Predicting Demographics and Affect in Social Networks

by

Svitlana Volkova

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

October, 2015

© Svitlana Volkova 2015

All rights reserved
Abstract

The recent explosion of social media services like Twitter, Facebook and Google+ has led to an interest in predicting hidden information from the large amounts of freely available public content. As compared to the earlier explosion of documents arising from the web, social media content is significantly more personalized – written in the first person, informal, and often revealing of latent attributes of users. The task of inferring latent user properties from social media data has become known as user modeling, personal analytics or user profiling task.

Previous approaches treated the task of user attribute prediction as static supervised classification, applied textual features extracted from user tweets and relied on an unrealistic amount of content per user (thousands of tweets). This dissertation relies mainly on Twitter data and focuses on several important but previously unexplored aspects of the task of user attribute prediction: (1) developing novel models and practical techniques that reflect the dynamic streaming nature of social media; (2) studying predictive power and latent relationships between user demographics, emotions and interests in social media; and (3) showing that extra-linguistic features such
as user demographics, personality and emotions can improve a variety of downstream applications, e.g., sentiment analysis and attribute-affect specific language modeling.

We start by developing models for low-resource batch prediction for Twitter users with no or limited content that rely on both text and network features extracted from user profiles – language of user neighbors, emotions, opinions and interests. We then treat the task of user profiling as iterative inference over communication streams. We extend our online streaming models to being able to learn and improve model quality iteratively over time. While allowing iterative learning, we further advance our models via iterative rationale (feature) annotation and weighting techniques. The major contributions of this thesis are the following:

- We present *models for low-resource static prediction* for Twitter users with no or limited content. Our batch models rely on language from user neighbors to overcome the *issue of topical sparsity* – when users’ posts are biased towards specific topics. Our results show that language of friends is the most predictive of user gender, language of followers is the most predictive of age, and language of retweeted and mentioned neighbors is predictive of user political preferences.

- We propose *models for iterative online inference* of latent user properties over real-time communication streams to directly model the *dynamic streaming nature of social media*. Our results demonstrate that streaming models that rely on joint user-neighbor streams are more efficient than static (batch) models.

- We develop *models for iterative learning and prediction* of user attributes over
ABSTRACT

time. We incorporate a “human in the loop” via crowdsourcing for iterative rationale (feature) annotation to overcome the issue of concept drift – when language and topics in social media change dramatically over time. Our results show that active learning approaches outperform iterative batch retraining, and learning from joint user-neighbor communication streams with iterative feature annotation significantly improves prediction results.

- We demonstrate how data sampling and annotation approaches influence attribute prediction accuracy. We experiment on multiple datasets annotated with self-reported attributes, following patterns and perceived crowdsourced annotations. Our results highlight significant differences in prediction performance depending on user activeness, on how Twitter users were sampled and on how their attribute annotations were collected.

- We study correlations between perceived user demographics and (i) emotions users express in social media, and (ii) the propensity of users to express different emotions than those expressed by their neighbors. Our results demonstrate that users with contrastive demographics express different emotions in their posts and tend to express more or less emotions compared to their neighbors.

- We release software that implements online streaming models and enables iterative predictions from Twitter communication streams.

- We present a prototype system for inferring psycho-demographics, emotions and opinions for any Twitter user by accessing user communications.
ABSTRACT

- We release seven datasets, which include between 350 and 5000 Twitter users and their tweets (mainly in English, except two additional datasets in Russian and Spanish), annotated with a variety of demographic attributes. We also release lists of annotator rationales for gender, age and political preference prediction in English and Spanish.

In this thesis we focus on practical and methodological aspects of the user attribute prediction task. We gradually improve our techniques starting with low-resource static classifiers, then experiment with online streaming models, and, finally, study models for iterative learning and prediction with crowdsourced annotations. In addition, we improve prediction performance by using better features (e.g., user emotions, interests, rationales). In conclusion, we demonstrate how user demographics can be useful for downstream prediction tasks such as gender-informed sentiment analysis and attribute-affect specific language modeling.

We also hope that this thesis will allow users to see how they may be perceived by their peers in social networks, so they can better understand what drives the image they project online.

Primary Reader: Benjamin Van Durme
Secondary Reader: David Yarowsky
Committee Member: Philip Resnik
Acknowledgments

I was fortunate to do my Ph.D. in the Center for Language and Speech Processing and the Human Language Technology Center of Excellence, which are the best academic and research communities in the world.

I am grateful to my adviser, Benjamin Van Durme, for his trust, support and guidance. He has pushed me to expand my boundaries and he was always there for me when I needed help. I owe my current knowledge and skills to him.

I also wish to thank my committee members, David Yarowsky and Philip Resnik. I am grateful to David for endless support, numerous ideas and so much wisdom to share. He taught me things beyond research, including networking and creative thinking. I am grateful to Philip for reading my whole dissertation front to back and giving a lot of feedback and detailed comments. I thank you both for the incredibly generous and beneficial outside guidance you provided, and I look forward to working together again in the future.

Bill Dolan and Yoram Bachrach have been my amazing mentors both during and outside of my time at Microsoft Research, where I had the pleasure of interning for
three summers. I thank Bill for his time and inspiration, and Yoram for being such a brilliant and fun collaborator. I hope to work with you both again in the future.

It has been a privilege to work with a group of exceptional researchers at Microsoft Research: Chris Quirk, Luke Zettlemoyer, Asli Celikyilmaz, Pallavi Choundry, Chris Brockett, Michael Gamon, Will Lewis, Patrick Pantel, Kristina Toutanova, Lucy Vanderwende, Katja Hofmann, Pushmeet Kohli and Chris Bishop.

I have learned much from other faculty members at CLSP: Mark Dredze, Jason Eisner, Sanjeev Khudanpur, Matt Post, Suchi Saria and others. I owe a special thank you to Chris Callison-Burch who admitted me to Hopkins and completely changed my life forever. I am also grateful to Theresa Willson and Glen Coppersmith, with whom I have been lucky to collaborate with on multiple projects.

I owe a lot to my Hopkins friends Courtney Napoles, Anni Irvine, Katie Henry, Juri Ganitkevitch and Ruth Scally. You brought fun into my Ph.D., and have contributed far more to my sanity than you realize.

I am thankful for great friendships with NLP students outside of JHU: Ekaterina Kochmar, Alisa Zhila, Ilija Chetviorkin, Mark Yatskar and Dirk Hovy. I owe thanks to my CLSP labmates: Xuchen Yao, Yuan Cao, Travis Wolfe, Peter Schulam, Frank Ferraro, Rachel Rudinger, Matt Gormley, Rebecca Knowles, Chandler May, Nanyun Peng (Violet), Jonathan Weese, Tim Vieira, Nicholas Andrews, Mike Carlin, Michael Paul and certainly others. I’ve also had the fortunate opportunity to collaborate with Daniel Preotiuc-Pietro, Vasileios Lampos and Nikolaos Aletras in recent months.
ACKNOWLEDGMENTS

I also wish to thank my friends from DoGEE at JHU: Jessica Lawson, Christine Kavanagh and Amar Wadhawan. I am grateful to Edward Bouwer for advising and graduating my husband with a Ph.D., and to Pat Bouwer for her amazing Thanksgiving dinners.

I owe a lot of my success to my undergraduate advisors and colleagues at the Perto Mohyla Black Sea University in Ukraine, Oleksandr Trunov and Yuriy Kondratenko, and I am grateful to my master’s advisors at Kansas State University, Doina Caragea and William Hsu.

I am grateful for the financial support I received from the Fulbright Program, The Institute of International Education and Google. These fellowships provided me with a huge amount of academic freedom. I wish to extend special thanks the Fulbright Graduate Scholarship coordinator in Ukraine, Inna Barysh for her help, support and patience.

Finally, I am grateful to my husband, Pavlo Bohutskyi, for supporting and loving me every day. I am especially thankful to my parents and my sister Anna, for always encouraging and supporting me. I cherish every single one of you.
Dedication

This thesis is dedicated to my parents Valentyna Volkova and Oleksandr Volkov, for always believing in me and always being there for me.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xxi</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Structure of this Document</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Datasets and Annotations</td>
<td>12</td>
</tr>
<tr>
<td>1.4 Related Publications</td>
<td>18</td>
</tr>
<tr>
<td>1.5 Summary of Contributions</td>
<td>20</td>
</tr>
<tr>
<td><strong>PART I: Methods</strong></td>
<td></td>
</tr>
<tr>
<td>2 Background: Personal Analytics in Social Media</td>
<td>25</td>
</tr>
</tbody>
</table>
CONTENTS

2.1 Supervised Models ............................................................... 26
  2.1.1 Features ................................................................. 26
  2.1.2 Classification .......................................................... 29
  2.1.3 Regression ............................................................... 30
2.2 Unsupervised Models .......................................................... 31
2.3 Graph-based Models .......................................................... 32
2.4 Streaming Approaches ....................................................... 36

3 Static Prediction ................................................................ 38
  3.1 Data .............................................................................. 40
    3.1.1 Twitter Language and Network Properties ....................... 40
    3.1.2 Twitter Social Network Graph ....................................... 41
    3.1.3 Twitter Social Relationship Types ................................. 43
      3.1.3.1 Graph-based Relationships ........................................ 43
      3.1.3.2 Content-based Relationships ..................................... 43
    3.1.4 Political Preference Graphs .......................................... 45
    3.1.5 Gender and Age Graphs .............................................. 46
  3.2 Batch Models ................................................................. 48
    3.2.1 Baseline User Model ................................................... 48
    3.2.2 Neighbor Model ........................................................ 50
    3.2.3 Joint User-Neighbor Model ......................................... 51
  3.3 Experiments ..................................................................... 51
5 Iterative Learning and Prediction \hspace{1cm} 88

5.1 Methodology \hspace{1cm} 91
  5.1.1 Data \hspace{1cm} 91
  5.1.2 Models \hspace{1cm} 92

5.2 Experimental Setup \hspace{1cm} 95
  5.2.1 Iterative Batch \hspace{1cm} 96
  5.2.2 Iterative Batch With Rationale Filtering \hspace{1cm} 97
  5.2.3 Active Without Oracle \hspace{1cm} 98
  5.2.4 Active With Oracle \hspace{1cm} 98
  5.2.5 Active With Rationale Filtering \hspace{1cm} 99
  5.2.6 Evaluation \hspace{1cm} 101

5.3 Experimental Results \hspace{1cm} 102
  5.3.1 Iterative Batch Results \hspace{1cm} 102
  5.3.2 Analyzing Threshold Influence \hspace{1cm} 105
  5.3.3 Studying Data Stream Type Influence \hspace{1cm} 105
  5.3.4 Comparing Performance for Batch vs Active Learning Setting \hspace{1cm} 106
  5.3.5 Evaluation in Constrained-Resource Conditions \hspace{1cm} 108
  5.3.6 Classification with Oracle Annotations \hspace{1cm} 110
  5.3.7 Applying Rationale Filtering \hspace{1cm} 110

5.4 Use-cases and Recommendations for Online Prediction \hspace{1cm} 114

5.5 Conclusions \hspace{1cm} 115
CONTENTS

6 Learning with Annotator Rationales  

6.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119

6.2 Methodology . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122

6.2.1 Models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122

6.2.1.1 Direct Model . . . . . . . . . . . . . . . . . . . . . . 123

6.2.1.2 Transitivity I Model . . . . . . . . . . . . . . . . . . . . . 123

6.2.1.3 Transitivity II Model . . . . . . . . . . . . . . . . . . . . . 123

6.2.2 Weighting Rationales . . . . . . . . . . . . . . . . . . . . . . . 124

6.3 Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125

6.3.1 Experimental Setup . . . . . . . . . . . . . . . . . . . . . . . . . 125

6.3.2 Experimental Results . . . . . . . . . . . . . . . . . . . . . . . 126

6.3.3 Comparison with Previous Work . . . . . . . . . . . . . . . . . 129

6.4 Additional Validation Experiments . . . . . . . . . . . . . . . . . . . 130

6.4.1 Removing Rationales from Tweets . . . . . . . . . . . . . . . . . 131

6.4.2 Matching Rationales . . . . . . . . . . . . . . . . . . . . . . . . 133

6.5 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 135

7 Background: Affect and Demographics in Social Networks  

7.1 Emotion Detection ... 139

7.2 Emotion Contagion and User Attributes ... 141
CONTENTS

7.3 Opinion Mining and User Properties ........................................... 143

8 Learning from Emotions and User-Neighbor Emotional Contrast .......... 146

8.1 Data ................................................................................................. 151

8.1.1 Perceived Demographic Data ....................................................... 151

8.1.1.1 Crowdsourced Participant Selection ....................................... 152

8.1.2 Emotion Data .............................................................................. 155

8.1.3 External Sentiment Data .............................................................. 157

8.2 Predicting User Emotions, Opinions and Demographics from Tweets .... 158

8.2.1 Perceived Attribute Prediction ...................................................... 159

8.2.2 Emotion and Opinion Detection .................................................. 160

8.2.3 Estimating Emotion & Opinion Distributions ................................. 161

8.2.4 Experimental Setup and Results .................................................. 162

8.2.4.1 Perceived Attribute Prediction Quality ..................................... 163

8.2.4.2 Emotion and Sentiment Prediction Quality ............................... 168

8.3 Correlating User Emotions and Perceived Demographics ................. 172

8.3.1 Estimating Emotion and Opinion Differences ............................... 172

8.3.2 Experimental Results ................................................................. 173

8.4 Inferring Perceived Demographics from Emotions and Opinions ....... 177

8.5 Correlating User-Environment Emotional Contrast and Demographics .... 179

8.5.1 Motivation ................................................................................... 179

8.5.2 Estimating User-Environment Emotional Contrast ....................... 181
CONTENTS

8.5.3 Experimental Results ........................................... 183
  8.5.3.1 User-Environment Emotional Contrast .................... 184
  8.5.3.2 User-Environment Opinion Contrast ...................... 185

8.5.4 Discussion ..................................................... 194
  8.5.4.1 Limitations and Caveats ................................... 196

8.6 Inferring Perceived Attributes from User-Neighbor Emotional Contrast 197

8.7 Conclusions ..................................................... 201

9 Learning from User Interests ........................................ 203
  9.1 Data ............................................................. 205
    9.1.1 Psycho-Demographic Attributes ............................. 205
    9.1.2 Interests .................................................. 206
  9.2 Methodology .................................................... 206
  9.3 Experimental Results ........................................... 210
  9.4 Conclusions .................................................... 213

10 Exploring Demographic Language Variations for Sentiment Analysis 216
  10.1 Motivation ..................................................... 217
  10.2 Data ........................................................... 219
  10.3 Subjective Language and Gender ................................ 222
    10.3.1 Bootstrapping Subjectivity Lexicons ....................... 223
    10.3.2 Lexical Evaluation ........................................ 225
CONTENTS

10.3.3 Hashtags ............................................. 230
10.3.4 Emoticons ............................................. 232
10.4 Experiments .............................................. 233
10.4.1 Models ................................................. 234
10.4.2 Results ................................................. 236
10.5 Conclusions .............................................. 240

11 Modeling Attribute-Affect Topic Variations in Social Media 244
11.1 Basic Topic Modeling ................................... 246
11.2 Data ..................................................... 247
11.3 Models .................................................. 248
11.3.1 Baseline Attribute-Affect LDA Model .............. 248
11.3.2 Attribute-Affect Mixture Model ................... 251
11.3.3 Comparing Attribute-Affect Models to Other Approaches ... 254
11.4 Experimental Results ................................... 256
11.4.1 Baseline Attribute-Affect LDA Model Results ...... 256
11.4.2 Preliminary Attribute-Affect Model Results ....... 257
11.5 Conclusions .............................................. 259

12 Conclusions and Future Directions 262
12.1 Summary of Contributions .............................. 268
12.2 Social Media Analytics Challenges Addressed ....... 271
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.3 Implications</td>
<td>272</td>
</tr>
<tr>
<td>12.4 Future Directions</td>
<td>273</td>
</tr>
<tr>
<td>A Mechanical Turk HIT Examples</td>
<td>275</td>
</tr>
<tr>
<td>B Twitter Graph Statistics</td>
<td>283</td>
</tr>
<tr>
<td>B.1 Candidate-Centric Graph Characteristics</td>
<td>285</td>
</tr>
<tr>
<td>C Querying Twitter API: Rate Limits</td>
<td>293</td>
</tr>
<tr>
<td>D A Note on Privacy and Ethics</td>
<td>295</td>
</tr>
<tr>
<td>E Visualizing Performance for The Most Predictive Neighborhoods</td>
<td>298</td>
</tr>
<tr>
<td>F Annotator Rationales for Gender, Age and Political Preference</td>
<td>305</td>
</tr>
<tr>
<td>G Emotion Hashtags</td>
<td>309</td>
</tr>
<tr>
<td>Bibliography</td>
<td>313</td>
</tr>
<tr>
<td>Vita</td>
<td>343</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Visualizing models for the task of latent user attribute prediction in social media .............................................. 8

2.1 Overview of approaches for gender prediction in social media ......................................................... 34
2.2 Overview of approaches for age prediction in social networks ......................................................... 35
2.3 Overview of techniques to infer ethnicity and origin on Twitter ...................................................... 36
2.4 Overview of methods for political preference classification on Twitter ............................................. 37

3.1 Dataset statistics: The number of users annotated with gender, age and political preference having at least one neighbor per user ......................................................... 48
3.2 Dataset statistics: Mean and standard deviation for the number of neighbors per user and the number of tweets per neighbor ......................................................... 48
3.3 Classification results for gender, age and political preference prediction using user, neighbor and user-neighbor models ......................................................... 57

5.1 Political preference classification accuracy for iterative batch models with and without rationale filtering ......................................................... 112
5.2 Political preference classification accuracy for active learning models with and without rationale filtering ......................................................... 113

6.1 Gender classification results for English using supervised and semi-supervised models with weighted annotator rationales ......................................................... 128
6.2 Comparing our rationale-weighted models with previously published approaches to gender classification ......................................................... 129
6.3 Gender classification performance when annotator rationales have been removed from tweets ......................................................... 132
6.4 Gender classification performance when matching annotator rationales ......................................................... 134

8.1 Crowdsourced annotations of perceived user properties ......................................................... 152
8.2 Crowdsourced multi-class annotations of perceived user properties ......................................................... 154
8.3 Statistics for a sampled social ego-network ......................................................... 154
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.4</td>
<td>Annotations for the external sentiment datasets</td>
<td>158</td>
</tr>
<tr>
<td>8.5</td>
<td>Cross-dataset comparison of accuracy for the gender attribute</td>
<td>164</td>
</tr>
<tr>
<td>8.6</td>
<td>Cross-dataset comparison of accuracy for the political preference attribute</td>
<td>164</td>
</tr>
<tr>
<td>8.7</td>
<td>Comparing model accuracy trained on true vs. perceived labels</td>
<td>167</td>
</tr>
<tr>
<td>8.8</td>
<td>Comparing our emotion classification results with existing systems</td>
<td>170</td>
</tr>
<tr>
<td>8.9</td>
<td>Sentiment differences among users with contrastive demographics</td>
<td>176</td>
</tr>
<tr>
<td>8.10</td>
<td>Mean JSD between incoming and outgoing sentiment distributions for users with different attributes</td>
<td>186</td>
</tr>
<tr>
<td>8.11</td>
<td>Mean JSD between incoming and outgoing emotion distributions for users with different attributes</td>
<td>187</td>
</tr>
<tr>
<td>9.1</td>
<td>Annotated attribute class distributions for the users with interests</td>
<td>205</td>
</tr>
<tr>
<td>9.2</td>
<td>Mean interest values in percentages for the most followed interest categories averaged within the groups of users with contrastive demographics</td>
<td>209</td>
</tr>
<tr>
<td>9.3</td>
<td>Attribute classification results using graph vs. content features</td>
<td>210</td>
</tr>
<tr>
<td>10.1</td>
<td>Gender and sentiment label distribution in the development and test sets for all languages</td>
<td>222</td>
</tr>
<tr>
<td>10.2</td>
<td>Initial and bootstrapped lexicon term counts with polarity across languages</td>
<td>225</td>
</tr>
<tr>
<td>10.3</td>
<td>Precision, recall and F-measure for subjectivity classification using the external, initial and bootstrapped lexicons for all languages</td>
<td>226</td>
</tr>
<tr>
<td>10.4</td>
<td>Sample of subjective terms to show lexical differences and polarity change across genders</td>
<td>228</td>
</tr>
<tr>
<td>10.5</td>
<td>Hashtag examples with opposite polarity across genders for English, Spanish, and Russian</td>
<td>230</td>
</tr>
<tr>
<td>10.6</td>
<td>Sentiment classification results obtained using gender-dependent and gender-independent joint and disjoint features</td>
<td>238</td>
</tr>
<tr>
<td>10.7</td>
<td>Precision and recall results for polarity classification: encoding gender as a binary feature vs. gender-dependent features</td>
<td>239</td>
</tr>
<tr>
<td>11.1</td>
<td>A set of example tweets annotated with perceived attributes and affect</td>
<td>247</td>
</tr>
<tr>
<td>11.2</td>
<td>Example topics with positive and negative sentiment for gender, age, children, stress and narcissism attributes</td>
<td>258</td>
</tr>
<tr>
<td>11.3</td>
<td>Negative topics for male and female users</td>
<td>260</td>
</tr>
<tr>
<td>11.4</td>
<td>Positive topics for male and female users</td>
<td>261</td>
</tr>
<tr>
<td>B.1</td>
<td>Clustering Twitter users by their tweeting frequency and the time since their accounts were created</td>
<td>285</td>
</tr>
<tr>
<td>B.2</td>
<td>User-follower and user-friend candidate-centric graph statistics</td>
<td>289</td>
</tr>
<tr>
<td>C.1</td>
<td>Twitter rate limits</td>
<td>294</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>User activeness on Twitter</td>
<td>5</td>
</tr>
<tr>
<td>1.2</td>
<td>Schematic overview of this thesis</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>Example tweets annotated with gender, political preference, income and education level</td>
<td>17</td>
</tr>
<tr>
<td>3.1</td>
<td>Twitter social graph constructed from followers, friends, replies, users with shared hashtags, mentioned and retweeted users</td>
<td>42</td>
</tr>
<tr>
<td>3.2</td>
<td>Classification results for age, gender and political preference prediction as a function of the number of tweets per user</td>
<td>55</td>
</tr>
<tr>
<td>3.3</td>
<td>Differences in classification decisions on a random partition of a candidate-centric graph while modeling the amount of content per user</td>
<td>59</td>
</tr>
<tr>
<td>3.4</td>
<td>Accuracy of model predictions for political preference as a function of the number of tweets per neighbor</td>
<td>62</td>
</tr>
<tr>
<td>3.5</td>
<td>Accuracy of model predictions for political preference as a function of the number of tweets per neighbor</td>
<td>63</td>
</tr>
<tr>
<td>3.6</td>
<td>Accuracy of political preference predictions as a function of number of neighbors using the best-performing neighborhoods</td>
<td>65</td>
</tr>
<tr>
<td>3.7</td>
<td>Summary of model results for political preference classification on multiple datasets</td>
<td>66</td>
</tr>
<tr>
<td>3.8</td>
<td>Best-performing neighborhoods recommended for gender, age and political preference classification</td>
<td>68</td>
</tr>
<tr>
<td>3.9</td>
<td>Top-ranked features learned from user communications for gender, age and political preference prediction</td>
<td>69</td>
</tr>
<tr>
<td>4.1</td>
<td>An example of stream-based prediction over a joint user-neighbor stream of communications</td>
<td>75</td>
</tr>
<tr>
<td>4.2</td>
<td>Political preference classification results using an iterative Bayesian update model over a user stream of communications</td>
<td>80</td>
</tr>
<tr>
<td>4.3</td>
<td>Time needed for the dynamic user and joint-user models to infer political preferences with different accuracy thresholds</td>
<td>82</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

4.4 Multi-dataset analysis of the prediction speed for political preference. 83
4.5 Multi-dataset comparison of batch vs. online models to infer user political preferences. 85

5.1 An example of a dynamic learning and prediction task 89
5.2 Active learning classification setup 94
5.3 Batch re-training setting with and without rationale filtering 96
5.4 Active learning setting with rationale filtering and oracle annotations 99
5.5 Comparing the performance of batch vs. active learning settings using different data streams and confidence thresholds 103
5.6 Comparing the performance of batch and active learning models with and without rationale filtering 104
5.7 Demonstrating the trade-off between model quality and classification accuracy for the batch setting 107
5.8 Demonstrating the trade-off between model quality and classification accuracy for the active learning setting 107
5.9 Measuring the influence of data stream type and learning setting on model quality over time 108
5.10 Daily classification performance in a batch setting 109
5.11 Classification performance for the active learning setting with oracle annotations 111
5.12 The number of training examples used for active vs. batch models over time 114

6.1 Gender rationale n-gram distribution for English and Spanish 120
6.2 Gender prediction accuracy using the direct model with weighted annotator rationales for English and Spanish 127
6.3 The distribution of tweets per user in MITRE gender dataset 132

8.1 An example prediction of perceived attributes, emotions and sentiments 147
8.2 An example of incoming and outgoing emotions and opinions for the users with predicted demographics 148
8.3 Our approach for predicting perceived psycho-demographic attributes, opinions and emotions 150
8.4 An example of psycho-demographic attribute predictions for a random user 160
8.5 An example of emotion and sentiment distributions for a random user 161
8.6 User demographic attribute classification performance 165
8.7 Word cloud illustration of the most predictive lexical markers (features) for income, gender, and relationship status attributes 166
8.8 Word cloud illustration of the most predictive lexical markers for gender learned from true vs. perceived annotations 168
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.9</td>
<td>Word cloud illustration of the most predictive lexical markers for six Ekman’s emotions.</td>
</tr>
<tr>
<td>8.10</td>
<td>Emotion differences among users with contrastive demographics.</td>
</tr>
<tr>
<td>8.11</td>
<td>Predicting hidden attributes from user emotions and opinions</td>
</tr>
<tr>
<td>8.12</td>
<td>User-environment emotional and opinion contrast given predicted perceived demographic traits.</td>
</tr>
<tr>
<td>8.13</td>
<td>User-environment emotional contrast between male and female users.</td>
</tr>
<tr>
<td>8.14</td>
<td>User-environment emotional contrast between older and younger users</td>
</tr>
<tr>
<td>8.15</td>
<td>User-environment emotional contrast between users with a college degree vs. high school education level.</td>
</tr>
<tr>
<td>8.16</td>
<td>User-environment emotional contrast between users with and without children.</td>
</tr>
<tr>
<td>8.17</td>
<td>User-environment emotional contrast between users with higher and lower annual income.</td>
</tr>
<tr>
<td>8.18</td>
<td>User-environment emotional contrast between single users and users in a relationship.</td>
</tr>
<tr>
<td>8.19</td>
<td>Opinion and emotion differences for optimism.</td>
</tr>
<tr>
<td>8.20</td>
<td>User-environment emotional contrast between users predicted to be satisfied and dissatisfied with life.</td>
</tr>
<tr>
<td>8.21</td>
<td>Opinion and emotion contagion differences between African American and Caucasian users.</td>
</tr>
<tr>
<td>8.22</td>
<td>Psycho-demographic attribute prediction from emotions and user-environment emotional contrast.</td>
</tr>
<tr>
<td>8.23</td>
<td>Correlations between user emotional-contrast features and demographic traits.</td>
</tr>
<tr>
<td>9.1</td>
<td>“Who to follow” popular profile distribution over 26 Twitter interest categories</td>
</tr>
<tr>
<td>9.2</td>
<td>Distribution of mean interests for male vs. female users.</td>
</tr>
<tr>
<td>9.3</td>
<td>Relationship between user interests and psycho-demographic attributes on Twitter</td>
</tr>
<tr>
<td>10.1</td>
<td>Gender-dependent vs. gender-independent subjectivity terms.</td>
</tr>
<tr>
<td>10.2</td>
<td>Distribution of gender-dependent and gender-independent sentiment terms.</td>
</tr>
<tr>
<td>10.3</td>
<td>Visualization of subjective language variations across genders on Twitter</td>
</tr>
<tr>
<td>10.4</td>
<td>Probability of gender and emoticons for English, Spanish and Russian.</td>
</tr>
<tr>
<td>10.5</td>
<td>Rule-based sentiment classification results for English.</td>
</tr>
<tr>
<td>10.6</td>
<td>Subjectivity and polarity classification results using supervised learning for English.</td>
</tr>
<tr>
<td>10.7</td>
<td>Relative accuracy improvements for subjectivity and polarity classification results across languages.</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.1</td>
<td>Baseline Attribute-Affect LDA Model</td>
<td>249</td>
</tr>
<tr>
<td>11.2</td>
<td>Attribute-Affect Mixture Model</td>
<td>252</td>
</tr>
<tr>
<td>A.1</td>
<td>Screen 1: MTurk HIT designed to crowdsource annotations for gender, age and political preferences of Twitter users.</td>
<td>276</td>
</tr>
<tr>
<td>A.2</td>
<td>Screen 2: MTurk HIT designed to crowdsource annotations for gender, age and political preferences of Twitter users.</td>
<td>277</td>
</tr>
<tr>
<td>A.3</td>
<td>Screen 1: MTurk HIT designed to crowdsource annotations for gender and age rationales.</td>
<td>278</td>
</tr>
<tr>
<td>A.4</td>
<td>Screen 2: MTurk HIT designed to crowdsource annotations for gender and age rationales.</td>
<td>279</td>
</tr>
<tr>
<td>A.5</td>
<td>Screen 1: MTurk HIT designed to crowdsource annotations for demographics and online personality of Twitter users.</td>
<td>280</td>
</tr>
<tr>
<td>A.6</td>
<td>Screen 2: MTurk HIT designed to crowdsource annotations for demographics and online personality of Twitter users.</td>
<td>281</td>
</tr>
<tr>
<td>A.7</td>
<td>Screen 3: MTurk HIT designed to crowdsource annotations for demographics and online personality of Twitter users.</td>
<td>282</td>
</tr>
<tr>
<td>B.1</td>
<td>Distribution of daily tweets for users sampled randomly over time from the 1% Twitter feed.</td>
<td>284</td>
</tr>
<tr>
<td>B.2</td>
<td>Candidate-centric graph example with Democratic and Republican candidates.</td>
<td>286</td>
</tr>
<tr>
<td>B.3</td>
<td>Distribution of followers and friends in the candidate-centric graph.</td>
<td>287</td>
</tr>
<tr>
<td>B.4</td>
<td>Distribution of retweets and user mentions in the candidate-centric graph.</td>
<td>290</td>
</tr>
<tr>
<td>B.5</td>
<td>Distribution of replies and hashtags in the candidate-centric graph.</td>
<td>291</td>
</tr>
<tr>
<td>B.6</td>
<td>True vs. sampled follower and friend neighborhoods in the candidate-centric graph</td>
<td>292</td>
</tr>
<tr>
<td>E.1</td>
<td>Complete age classification results obtained from our neighbor model</td>
<td>299</td>
</tr>
<tr>
<td>E.2</td>
<td>Complete age classification results obtained from our user-neighbor model</td>
<td>300</td>
</tr>
<tr>
<td>E.3</td>
<td>Complete gender classification results obtained from our neighbor model</td>
<td>301</td>
</tr>
<tr>
<td>E.4</td>
<td>Complete gender classification results obtained from our user-neighbor model</td>
<td>302</td>
</tr>
<tr>
<td>E.5</td>
<td>Complete political classification results obtained from our neighbor model</td>
<td>303</td>
</tr>
<tr>
<td>E.6</td>
<td>Complete political preference classification results obtained from our user-neighbor model</td>
<td>304</td>
</tr>
<tr>
<td>F.1</td>
<td>Gender rationale examples in English</td>
<td>306</td>
</tr>
<tr>
<td>F.2</td>
<td>Age rationale examples in English</td>
<td>307</td>
</tr>
<tr>
<td>F.3</td>
<td>Political preference rationale examples in English</td>
<td>308</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation

This thesis targets data emerging from social media networks, such as Twitter, Facebook and Google+. At the time of this writing, these services are already being used regularly by more than 1/7th of the world’s population. The explosion of such services has led to a growing application potential for personalization in human computer systems ranging from intelligent user interfaces or conversational agents and recommendation systems to large-scale healthcare analytics, real-time polling, online targeted advertising and marketing. Researchers have started mining the massive volumes of personalized and diverse data produced in public social media with the goal of learning about user demographics like gender, age, political preference, ethnicity (Burger et al., 2011; Zamal et al., 2012) and personality (Kosinski et al., 2013), as
well as users’ language (Kern, Eichstaedt, Schwartz, Park, et al., 2014), their interests or likes (Bachrach et al., 2012), the emotions and opinions they express (Bollen, Mao, & Zeng, 2011), their well-being (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al. 2013) and their interactions with the online environment. As a result, several social media predictive analytics services have been implemented. These services input a profile from a social network, e.g., Facebook or Twitter, and output predictions about the personality, emotions, sentiments, and demographic characteristics of the person behind the profile. Some example applications are listed below.

- Social Network Prediction: https://apps.facebook.com/snpredictionapp/
- Twitter Psycho-Demographic Profile and Affect Inference: http://twitterpredictor.cloudapp.net (pswd: twitpredMSR2014)
- You Are What You Like: http://youarewhatyoulike.com/
- Psycho-demographic trait predictions: http://applymagicsauce.com/
- IBM Personality: https://watson-pi-demo.mybluemix.net

The approaches proposed in this thesis for automatically classifying user properties from profiles in social media will enable researchers such as computational sociolinguists, psychologists, sociologists, social scientists and social media analysts to un-

CHAPTER 1. INTRODUCTION

derstand social media users and their behavior online. Moreover, it has been recently shown that socio- and psycho-linguistic information can help a variety of downstream natural language processing (NLP) tasks including sentiment analysis (Volkova et al., 2013a), language generation (Eisenstein et al., 2014) and topic classification (Hovy, 2015). Other potential applications include but are not limited to:

- **personalized conversational agents or intelligent user interfaces** that rely on user personality and demographics and are able to satisfy user intents by tracking user emotions and opinions in real-time;

- **online advertising** where the advertiser has the ability to personalize content, e.g., targeted ads, based on automatically predicted user features, so as to match the emotional tone the user expects;

- **personalized recommendation systems**, for example news or music recommendation based on user emotional tone, demographics and personality;

- **unique marketing strategies** e.g., detecting opinions and emotions users express about products or services within targeted populations, e.g., younger users;

- **large-scale real-time healthcare analytics** that could identify patterns such as depression or mental illness, track flu and disease outbreaks, etc.;

- **enhancing existing recruitment and human resource management strategies** by estimating emotional stability and personality of potential employees or measuring overall well-being, e.g., life satisfaction, happiness, stress;

- **large scale polling** to mine political opinions or predict voting outcomes.
CHAPTER 1. INTRODUCTION

However, predicting personal analytics from streaming social media communications is challenging for several reasons:

1. **Sampling and annotation biases:** The majority of work on social media analytics relies on samples from social network data. Therefore, biases introduced by sampling and annotation procedures have to be addressed and the representativeness and validity of the studies have to be verified (Tufekci, 2014).

2. **Dynamic streaming data:** The streaming nature of social media requires models that are able to handle dynamic updates of the environment. For example, the Twitter communication graph evolves rapidly (Van Durme, 2012b).

3. **Data (concept) drift:** Social media content changes dynamically. As new trending topics arise, statistical models that relied on language from the past may degrade over time (Dredze, Oates, & Piatko, 2010; Fromreide, Hovy, & Søgaard, 2014; Osborne, Lall, & Van Durme, 2014).

4. **Topical sparsity:** “Average” Twitter users talk about certain topics, e.g., politics, only sporadically (Cohen & Ruths, 2013) as compared to the political debate text (Thomas et al., 2006).

5. **User activeness and model generalization:** As shown in Figure 1.1, users post on social media with varying frequency. Some users may participate just once a week, while others post ten times a day. Models trained on data from active users may not generalize to the “average” users who have no or limited content (Burger et al., 2011; Sap et al., 2014).
CHAPTER 1. INTRODUCTION

Furthermore, language in social media is extremely sparse, noisy, heterogeneous and informal with many non-standard spellings (e.g., cuming “coming” or 2mr “tomorrow”) and abbreviations (e.g., u “you”). Tweets are short and sparse, unlike blogs (Goswami et al., 2009; Peersman et al., 2011), emails (Cheng et al., 2009), movie reviews (Otterbacher, 2010), scientific papers (Bergsma et al., 2012) or telephone speech (Garera & Yarowsky, 2009).

The objective of this thesis is to develop predictive models and tools for latent user demographic inference from social media streams. This thesis focuses on several main aspects of the problem: models – by developing iterative, scalable and accurate approaches; features – by detecting the most predictive signal; and applications – by showing how user demographics can be effectively used and improve other downstream tasks as shown in Figure 1.2.
CHAPTER 1. INTRODUCTION

In this thesis we experiment with static models and explore local user neighborhoods to infer user demographics from the language of neighbors of different types, e.g., friends, retweets, etc. We propose streaming and active learning models to handle iterative learning and prediction from social media streams over time, and we incorporate human knowledge via iterative rationale (feature) annotation. We experiment with a variety of features to predict user properties that reveal interesting and previously unexplored relationships between language, emotions, opinions, interests and psycho-demographic attributes in social networks. In addition, we demonstrate how our approaches can be effectively used for several applications such as gender-aware subjectivity and polarity classification, and personalized attribute and affect-aware language generation. We also show that our approaches for socio-linguistic content analysis in social media expand the applicability to other languages (Spanish and Russian), and that they can be used effectively in a constrained-resource predictive scenario, i.e., with no or limited content from a user. Finally, the techniques proposed in this thesis address the social media predictive analytics challenges discussed earlier.

1.2 Structure of this Document

Figure 1.2 gives a schematic overview of the thesis and shows how each chapter contributes to social media predictive analytics. In Table 1.1 we visualize the models proposed in each chapter.
CHAPTER 1. INTRODUCTION

Figure 1.2: Schematic overview of this thesis. It consists of two parts – Part I: Methods and Part II: Feature Analysis and Applications. In Chapter 2, we present the background for the user attribute prediction task, and in Chapter 7, we review the existing approaches for emotion detection and sentiment analysis in social networks. In Chapter 3, we define our constrained-resource static models for user attribute prediction that rely on content from user local neighborhoods of different types. In Chapter 4, we propose novel streaming models for predicting user properties from communication streams over time. In Chapter 5, we demonstrate how to iteratively update our models learned from joint user-neighbor streams using active learning and rationale (feature) filtering. In Chapter 6, we further investigate annotator rationales and develop models with rationale weighting to boost classification performance.

In part II, we show how models from Part I can be further improved through user emotions and user-neighbor emotional contrast (Chapter 8) and through user interests (Chapter 9). In Chapters 10 and 11, we analyze demographic language variations in social media for sentiment classification, and develop joint generative attribute-affect mixture models to further study lexical variations in social networks.
Table 1.1: Visualizing models for the task of latent user attribute prediction over multi-attributed, multi-relational and iteratively evolving social networks.

<table>
<thead>
<tr>
<th>User baseline</th>
<th>Multi-attributed</th>
<th>Multi-relational</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch. 3, Ch. 6</td>
<td>Content-based</td>
<td>Ch. 3, Ch. 8</td>
</tr>
<tr>
<td></td>
<td>discriminative</td>
<td>Context-based</td>
</tr>
<tr>
<td></td>
<td>single-attribute</td>
<td>discriminative</td>
</tr>
<tr>
<td></td>
<td>models</td>
<td>and generative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>models</td>
</tr>
<tr>
<td>a_k</td>
<td>v_i - user</td>
<td>N(v_i) - user</td>
</tr>
<tr>
<td></td>
<td>a_k - attribute</td>
<td>neighbors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ch. 4, Ch. 5</td>
<td>Content-based</td>
<td>Ch. 4, Ch. 5</td>
</tr>
<tr>
<td></td>
<td>single-attribute</td>
<td>Context-based</td>
</tr>
<tr>
<td></td>
<td>models with</td>
<td>discriminative</td>
</tr>
<tr>
<td></td>
<td>iterative</td>
<td>and generative</td>
</tr>
<tr>
<td></td>
<td>Bayesian updates</td>
<td>models</td>
</tr>
<tr>
<td>a_k</td>
<td>v_i - user</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a_k - attribute</td>
<td></td>
</tr>
<tr>
<td>a_k'</td>
<td>v_i' - user</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a_k' - attribute</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ch. 10, Ch. 11</td>
<td>Content-based</td>
<td>Ch. 10, Ch. 11</td>
</tr>
<tr>
<td></td>
<td>multi-attributed</td>
<td>Context-based</td>
</tr>
<tr>
<td></td>
<td>discriminative</td>
<td>discriminative</td>
</tr>
<tr>
<td></td>
<td>and generative</td>
<td>and generative</td>
</tr>
<tr>
<td></td>
<td>models</td>
<td>models</td>
</tr>
<tr>
<td>a_k</td>
<td>v_i - user</td>
<td></td>
</tr>
<tr>
<td>a_l</td>
<td>v_i' - user</td>
<td></td>
</tr>
<tr>
<td>a_k'</td>
<td>v_i' - user</td>
<td></td>
</tr>
<tr>
<td>a_l'</td>
<td>v_i' - user</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ch. 3 – Ch. 11</td>
<td>Content + context-based models</td>
<td>Ch. 3 – Ch. 11 Content + context-based models</td>
</tr>
<tr>
<td></td>
<td>for iterative learning and prediction over multi-relational multi-attributed social network streams</td>
<td>for iterative learning and prediction over multi-relational multi-attributed social network streams</td>
</tr>
</tbody>
</table>

Network properties: multi-attributed nodes, multi-relational edges, evolving structure.
Constraints: iterative data acquisition, sampling and annotation biases.
Challenges: concept drift, topical sparsity, user activeness and model generalization.
CHAPTER 1. INTRODUCTION

The thesis is structured as follows:

- Chapter 2 gives an overview of the existing approaches for latent user attribute prediction in social media.

- Chapter 3 presents results for the latent user attribute prediction task using static constrained-resource supervised models. It contains a range of experiments on multiple datasets for three personal attributes: gender, age and political preference. This chapter shows an analysis and empirical evaluation on the quality of the neighborhood content for predicting different attributes, feature types and the amount of content per user or neighbor (e.g., tweets necessary to obtain the highest results). This chapter also includes a direct comparison with existing approaches.

- Chapter 4 details a novel technique that relies on iterative Bayesian updates for streaming prediction over time from user or joint user-neighbor streams. It includes a range of experiments for political preference prediction on multiple datasets to analyze model generalization. This chapter summarizes prediction results in terms of prediction time and accuracy, and compares our streaming models with our batch models from Chapter 3.

- Chapter 5 extends our streaming models from Chapter 4 and presents a range of methods to iteratively learn and make predictions over time from user or user-neighbor communication streams. Our approaches rely on iterative re-
training, active learning and annotator rationale (feature) filtering techniques. This chapter empirically evaluates user demographic classification performance using iterative learning approaches, compares iterative re-training vs. active learning with and without rationale filtering, discusses the trade-off between model quality over time and prediction time, and outlines several use-cases and recommendations for targeted (demographic-specific) advertising.

- Chapter 6 details novel approaches for further improving static models for personal analytics in social media from Chapter 3. It includes experiments on rationale (feature) weighting techniques for user attribute classification taking as an example gender prediction task. We have released our lists of annotator rationales, which are highly predictive n-grams labeled by human annotators for inferring gender, age and political preference attributes.

- Chapter 7 gives an overview of the existing approaches for emotion and opinion detection in social networks and their correlation with user attributes.

- Chapter 8 applies our models from Part I to develop a framework for automatically inferring coarse-grained emotions, opinions and a range of previously unexplored psycho-demographic properties from public texts in social media. It also studies correlations between user demographics and (a) emotions expressed in user tweets and (b) emotional contrast between users and their neighbors of different types. This chapter demonstrates a set of experiments showing that
users with contrastive demographics both project different emotions in social media and react differently to the emotions expressed by their neighbors. Finally, it shows that one can predict user demographics solely from emotions expressed in user profiles and their local neighborhood.

- Chapter 9 presents experiments that qualitatively and quantitatively evaluate user interests (graph-based or network features) to predict a variety of latent user properties, and compare the results with the models that rely exclusively on language (graph-based features) as described in Chapters 3 and 8.

- Chapter 10 demonstrates how extra-linguistic features such as gender, which have been effectively predicted from texts published in social media in Chapters 3 – 9 can be successfully used to improve downstream NLP tasks such as opinion mining in multiple languages Russian, English and Spanish. We analyze demographic language variations including the usage of domain-specific items such as hashtags and emoticons by male and female users on Twitter. It shows how these gender differences in subjective language improve subjectivity and polarity classification in social media.

- Chapter 11 demonstrates the advantages of extra-linguistic features features in applications, taking the example of psycho-demographic attributes for personalized conversation agents. We analyze lexical variations in social media with novel attribute-affect mixture models for a variety of attribute-affect combina-
CHAPTER 1. INTRODUCTION

tions, e.g., female and fear, high income and anger, life satisfaction and positive
sentiment. In addition, the proposed models can be further extended with the
capability of modeling attribute-specific subjective language from either user or
joint user-neighbor streams over time to account for topical sparsity, concept
drift and the streaming nature of social media.

1.3 Datasets and Annotations

In this thesis we have constructed several datasets listed below.

- In Chapters 3 – 5 we run our experiments on Twitter user profiles:

  - Political Preference Datasets:

    Candidate-Centric: 1,031 users, 515 Democratic, 516 Republican (follow
    political candidates).

    Geo-Centric: 270 users, 135 Democratic, 135 Republican (self-reports in
    Delaware, Maryland and Virginia areas).

    Politically Active: 371 users, 185 Democratic, 186 Republican (politically
    active and follow political candidates).

  - Age Dataset: 387 users, 190 “18 - 23 years old”, 191 “25 - 30 years old”
    (self-reports, e.g., “Happy 25th birthday to me”).

  - Gender Dataset: 383 users, 191 female, 192 male (user name).

    For each user in the above datasets we collected the 200 most recent tweets,
CHAPTER 1. INTRODUCTION

and between 10 - 20 neighbors of six types including followers, friends, user
mentions, retweets, replies and shared hashtags.

• In Chapter 6 we run our experiments on Twitter user profiles:

  – **English Gender Dataset:** 357 users, 193 female, 164 male users in train,
    and a balanced set of 200 users in test (annotated via crowdsourcing); 1M
    tweets used in a semi-supervised setting (12.6k users);

  – **Spanish Gender Dataset:** 443 users, 192 female, 251 male, and a bal-
    anced set of 200 users as a held-out data (annotated via crowdsourcing);
    1M unlabeled tweets (7.5k users).

• In Chapters 8 – 9 we run our experiments on Twitter user profiles:

  – **Psycho-Demographic Profile Dataset:** 5,000 users annotated with 18
    psycho-demographic attributes;

  – **Emotion Dataset:** 70k tweets annotated with six Ekman’s emotions:
    sadness, joy, fear, surprise, anger, disgust (annotated via distant supervi-
    sion).

  – **Sentiment Dataset:** 20k tweets annotated with positive, negative and
    neutral sentiment.\(^2\)

\(^2\)We aggregate tweets from publicly available sentiment datasets outlined by Hassan Saif, Miriam
Fernandez and Alan\(l\) (2013).
CHAPTER 1. INTRODUCTION

– **Interest Dataset:** 4k users annotated with the distributions over 26 Twitter interest categories.

We use the above datasets to predict psycho-demographic profiles for 123k Twitter users and label their 25M tweets with emotions and sentiments.

• In Chapter 10 we run our experiments for three languages: English, Spanish and Russian. For each language we use:

  – **English, Spanish and Russian Gender Datasets:** a balanced sample of 1M tweets annotated with gender;

  – **English, Spanish and Russian Gender and Sentiment Datasets:** 4,000 tweets annotated with gender and sentiment.

Several approaches have been used to annotate the data with user attributes. Similar to other supervised classification tasks in a new domain, obtaining ground truth labels for latent user attributes is extremely costly and time consuming. Therefore, other approaches have been successfully used to obtain a good approximation of the ground truth as discussed below:

1. **Profile metadata:** Data such as relationship status, gender, age are often public on Facebook (Kosinski et al. 2013), reviews (Hovy 2015), YouTube (Filippova 2012) or Google+ (Fang et al. 2015) but extremely sparse on Twitter.

2. **Self reports:** Posts like “I am a republican ...” (used in our *geo-centric graph in Chapters 3 and 4*), “Happy #th/st/nd/rd birthday to me” (used in our *age
CHAPTER 1. INTRODUCTION

*graph in Chapter 3*, or other self-identifications mentioned in user biography fields “I have been diagnosed with ...” (Coppersmith, Dredze, & Harman, 2014), “I am a writer/actress/waitress/doctor...” (Beller et al., 2014). Note that self-reports are rare on Twitter. For instance, according to Burger et al. (2011), out of 1,000 randomly sampled Twitter users only 15% had an explicit gender cue. Interestingly, female users are more likely to provide specific indicators of their gender than male users.

3. **Online psychological tests:** Notably, personality tests voluntarily taken on Facebook – myPersonality project (http://mypersonality.org/wiki/doku.php) and other works that use the data (Kern, Eichstaedt, Schwartz, Park, et al., 2014; Sap et al., 2014; Schwartz, Eichstaedt, Dziurzynski, et al., 2013; Kern, Eichstaedt, Schwartz, Dziurzynski, et al., 2014; Youyou et al., 2015).

4. **Distant supervision:** Following Obama vs. Romney (used in our candidate-centric and politically active graphs in Chapters 3– 5 and other datasets annotated with political preferences (Zamal et al., 2012; Cohen & Ruths, 2013)), emotion hashtags (used in our emotion dataset in Chapter 8 and other datasets annotated with emotions (S. M. Mohammad & Kiritchenko, 2014)), user names (used in our gender graph in Chapter 3 and other datasets annotated with gender (Burger et al., 2011; Zamal et al., 2012; Bergsma et al., 2013)), following popular accounts on Twitter (Culotta et al., 2015).
CHAPTER 1. INTRODUCTION

5. **Crowdsourcing:** Subjective perceived attribute annotations (used in our *psychodemographic dataset in Chapter 9* and other datasets in multiple languages annotated with gender (Ciot et al., 2013), rationales (used to collect our *rationale lists and annotate a held out data with gender in English and Spanish in Chapter 6* and other similar datasets (Bergsma & Van Durme, 2013; W. Liu et al., 2012)), sentiment/polarity annotations (used in our *English, Spanish and Russian sentiment and gender datasets in Chapter 10*).

How difficult is it to annotate user profiles with personal attributes? Recent work on crowdsourced annotation experiments reports that 84% of annotators (on average 200 annotators per profile) agree on user gender (D.-P. Nguyen et al., 2014). In Figure 1.3 we present some example tweets authored by a female user, a Democratic user, a user with high income and a user with high education level. These tweets were available during the annotation process and were among those used to predict user gender and political preference attributes in Chapters 3–5, as well as education and income attributes in Chapters 8–9.

The annotation approaches listed above have certain limitations, and they are only approximations of the ground truth. Some of them are more reliable than others. For example, questionnaires completed voluntarily are superior to crowdsourced annotations. Distant supervision approaches do not take into account subjective use of the emotional hashtags nor do they disambiguate following strategies on Twitter. Self reports and other clues in user metadata are extremely sparse on Twitter.
CHAPTER 1. INTRODUCTION

(a) Gender: Female

(b) Political Preference: Democratic

(c) Income: Over $75k

(d) Education: Degree

Figure 1.3: Example tweets annotated with gender, political preference, income and education level.

In Appendix A we present example MTurk HITs developed to crowdsource gender, age and political preference annotations for the data used in Chapters 3, 6, 8, and 11.
CHAPTER 1. INTRODUCTION

1.4 Related Publications

The research described in this thesis is that of the author, some of which has contributed to articles previously published or in-progress:

- Chapters 2 and 3 presents models and the data we released along with “Language and Online Identity in Social Media: What You Write, Who You Know, and What Who You Know Writes” (to be published) and is joint work with Benjamin Van Durme, Glen Coppersmith and David Yarowsky.

- Chapters 3 and 4 present and extend “Inferring User Political Preferences from Streaming Communications,” which was published in the Proceedings of the Conference of the Association for Computational Linguistics (ACL) in 2014 and is joint work with Benjamin Van Durme and Glen Coppersmith.

- Chapter 5 presents “Online Bayesian Models for Personal Analytics in Social Media,” which was published in the Proceedings of the 29th Conference on Artificial Intelligence (AAAI) in 2015 and is a joint work with Ben Van Durme.

- Chapter 6 presents “Improving Gender Prediction of Social Media Users via Weighted Annotator Rationales,” which was published in the Proceedings of the NIPS Workshop on Personalization: Methods and Applications in 2014 and is a joint project with David Yarowsky.

- Chapter 8 presents the demo system that we built and released along with “In-
ferring Latent User Properties from Texts Published in Social Media,” which was published in the Proceedings of the 29th Conference on Artificial Intelligence (AAAI) in 2015 and is a joint work with Yoram Bachrach, Michael Armstrong and Vijay Sharma.

- Chapter 8 presents and extends “On Predicting Socio-Demographic Traits and Emotions in Social Networks and Implications to Online Self-Disclosure” (to appear) in Cyberpsychology, Behavior, and Social Networking journal in 2015 and is a joint work with Yoram Bachrach; “Inferring Perceived User Properties from Emotional Tone and User-Environment Emotional Contrast” (to be published), which is a collaboration with Yoram Bachrach and Benjamin Van Durme; and “Studying User Income through Language, Behaviour and Affect in Social Media,” which was published in PLOS One journal in 2015 and is a joint project with Daniel Preotiuc-Pietro, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras.

- Chapter 9 presents and extends “Mining User Interests to Predict Perceived Psycho-Demographic Traits on Twitter” (to be published), which is a joint project with Yoram Bachrach and Benjamin Van Durme; “Using Emotions to Predict User Interest Areas in Online Social Networks”, which is a collaboration with Yoad Lewenberg and Yoram Bachrach published in the Proceedings of IEEE Conference on Data Science and Advanced Analytics (DSAA).
CHAPTER 1. INTRODUCTION

• Chapter 10 presents and extends “Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media,” which was published in the Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP) in 2014 and is a joint project with Theresa Willson and David Yarowsky.

• Chapter 11 presents “Analyzing Lexical Variations in Social Media with Attribute-Affect Mixture Models” (to be published), a joint project with Benjamin Van Durme.

• This thesis presents and extends the tutorial on “Social Media Predictive Analytics” presented at the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) in June 2015.

1.5 Summary of Contributions

This thesis addresses several questions:

• How to infer latent user properties (gender, age, income); emotions (fear, surprise); interests (sports, news); and preferences (political favorites) from unstructured, informal, topically distributed and sparse communications in social networks.

• How to make predictions over dynamically evolving (streaming) social network data in a constrained-resource setting, when no or limited user tweets are avail-
CHAPTER 1. INTRODUCTION

able, when users post with different frequency and when topics being discussed are changing rapidly over time.

• What are the relationships between users’ languages, their psycho-demographic attributes, the emotions and sentiments they express and their interests in social media.

In this thesis we use machine learning (ML) and natural language processing (NLP) techniques to develop models for socio-linguistic content analysis over dynamic social network data. The major contributions of this thesis are as follows:

• **Methodologies and practical techniques:**

  – We propose a method that relies on a social network graph of user neighborhoods of different types, e.g., friends, and combines this networks with language in social media to infer latent user properties. Our method has been successfully applied to make predictions for users with no or limited content or users who talk about certain topics less frequently.

  – We develop several rationale annotation and weighting techniques to improve latent attribute prediction. The techniques can be efficiently ported via crowdsourcing to new languages, requiring access only to everyday native speakers of a target language.

  – We present a novel algorithm for iteratively inferring latent user properties from social media communication streams. Our approach handles the
CHAPTER 1. INTRODUCTION

dynamic nature of social media and takes into account user activeness.

– We develop a practical approach that learns models on the fly in addition to making predictions of latent user properties dynamically. It relies on iterative batch retraining, active learning and rationale annotation techniques via crowdsourcing. Our approach helps to address the issues of data drift and model generalization in social media.

– We study the relationships between user psycho-demographics, emotions, opinions and interests in social media. First, we show that users with contrastive demographics express different emotions and have different interests. Second, we demonstrate that users with contrastive demographics react differently to the emotions expressed by their neighbors.

• Software:

– We construct a prototype demonstration system, made available online, that highlights what users’ publicly posted social media content might reveal about them. This system accesses the most recent 200 tweets from a Twitter profile for a given user and outputs predictions for 18 psycho-demographic attributes and proportions of emotions and opinions for that person’s profile.

– We release a software package that implements our streaming models for user attribute prediction that rely on user, neighbor or joint user-neighbor
CHAPTER 1. INTRODUCTION

streams. This software can train, test and save pre-trained models; make prediction on the fly; and plot iterative prediction updates obtained from single or multiple communication streams across several user attributes.

• Datasets:

  – We release several datasets annotated with latent user properties including gender, age, political preferences and other 18 psycho-demographic attributes, six basic Ekman’s emotions and 26 Twitter interest categories.

  – We release the lists of annotator rationales (the most predictive n-grams) for gender, age and political preferences in English and Spanish.
PART I: Methods
Chapter 2

Background: Personal Analytics in Social Media

In this chapter we review prior work on predicting individual user attributes in batch static settings using supervised and unsupervised approaches including Bayesian graphical models and graph-based methods in sections 2.1, 2.2 and 2.3, respectively. We present the analysis of the limited number of streaming approaches for social media predictive analytics in 2.4. We perform a detailed comparison of the existing methods and aggregate performance numbers for individual attributes achieved using the baseline models trained on Twitter data.
CHAPTER 2. BACKGROUND: PERSONAL ANALYTICS IN SOCIAL MEDIA

2.1 Supervised Models

The vast majority of work on predicting latent user attributes in social media apply supervised models – Support Vector Machines (SVM) for binary classification e.g., gender and regression for continuous categorical attributes e.g., age with lexical bag-of-word features for predicting user gender (Garera & Yarowsky, 2009; Rao et al., 2010; Burger et al., 2011; Van Durme, 2012b; Ruths et al., 2014), age (Rao et al., 2010; D. Nguyen et al., 2011, 2013) or political orientation (Rao et al., 2010; Conover, Gonçalves, et al., 2011; Pennacchiotti & Popescu, 2011a; Zamal et al., 2012).

Bergsma et al. (2012), extending Rao et al. (2010) work on adding socio-linguistic features to improve user attribute prediction, show that incorporating stylistic and syntactic information to the bag-of-word features improves classification performance. Moreover, recent work by Bergsma and Van Durme (2013) presents an approach which allows bootstrapping the training data by learning lexical “common-sense attributes” (highly predictive n-grams) for male and female users on Twitter.

2.1.1 Features

The existing content-based models that rely on user communications apply bag-of-word (BOW) **lexical features** for latent attribute classification in social media:

- **normalized word n-gram frequencies** (Rao et al., 2010; Bergsma et al., 2012; Van Durme, 2012b; D. Nguyen et al., 2013);
• character-based n-grams (Burger et al., 2011; Bergsma et al., 2013; Pennacchiotti & Popescu, 2011b), user name or biography (Burger et al., 2011);

• class-based highly predictive n-grams (Bergsma & Van Durme, 2013; Sap et al., 2014), k-top discriminative words, digrams, trigrams, stems (Zamal et al., 2012; Cohen & Ruths, 2013; Bamman et al., 2014).

The rest of the feature types can be aggregated into three main groups: network structure features, communication behavior features, and psycho-linguistic, syntactic and stylistic features as shown below:

• network structure features follower-following ratio, follower or following frequency (Filippova, 2012; Preoiuc-Pietro et al., 2015); neighborhood size (Zamal et al., 2012), following strategies (Culotta et al., 2015);

• communication behavior features retweet and tweet frequency, reply rate (Singla & Richardson, 2008; Pennacchiotti & Popescu, 2011b; Golbeck et al., 2011; Conover, Gonçalves, et al., 2011); retweet rate (Zamal et al., 2012);

• psycho-linguistic smiles, excitement, puzzled punctuation, emoticons and socio-linguistic (Rao et al., 2010; Bamman et al., 2012; Kokkos & Tzouramanis, 2014; Hovy, 2015; Rangel & Rosso, 2015);

• syntactic and stylistic features e.g., Context-Free Grammar (CFG) rules, Tree-Substitution Grammar (TSG) fragments (Bergsma et al., 2012), punctuation (Hovy, 2015), stopwords and abbreviations (Cheng et al., 2011), lexicon e.g., LIWC features (Pennebaker et al., 2001; Fink et al., 2012; Sap et al.)
CHAPTER 2. BACKGROUND: PERSONAL ANALYTICS IN SOCIAL MEDIA


Other features include Facebook “likes” (Bachrach et al., 2012; Kosinski et al., 2013), user names (W. Liu & Ruths, 2013), links and images (Rosenthal & McKeown, 2011) as well as word embeddings, LDA topics, word clusters (Preotiuc-Pietro, Eichstaedt, et al., 2015; Hovy, 2015).

However, a variety of works on text classification show “discouraging results” that simple BOW features perform better than more sophisticated features e.g., syntax, stylistic features (Sebastiani, 2002) e.g., for sentiment classification (Pang et al., 2002; Volkova et al., 2013a) and age prediction (D. Nguyen et al., 2011, 2013). We also show that it is true when predicting multiple user attributes by comparing our models with the models from (Zamal et al., 2012) trained using more sophisticated features.

Some works have studied user demographic prediction in social media for languages other than English e.g., gender (Ciot et al., 2013; Hovy, 2015) and age (Peersman et al., 2011; Hovy, 2015), as well as the shared task on author profiling in English and Spanish (Rangel, Rosso, Potthast, Stein, & Daelemans, 2015).1

Besides the above discussed well-studied attributes e.g., gender, age and political views, some works have started building predictive analytics for other more subjective and, therefore, harder to guess, latent user properties including occupation or job (Preotiuc-Pietro, Lampos, & Aletras, 2015; Sloan et al., 2015; Fang et al.

---

1PAN 15 author profiling shared task: [http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-profiling.html](http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-profiling.html)
CHAPTER 2. BACKGROUND: PERSONAL ANALYTICS IN SOCIAL MEDIA

..., education (J. Li, Ritter, & Hovy, 2014), relationship (Fang et al., 2015) and income (Preoïnc-Pietro et al., 2015).

2.1.2 Classification

Given an input vector $f(v_i) \in \mathbb{R}^d$, where $f(v_i)$ represent features extracted from user $v_i$’s communications, the goal is to find a prediction of user attributes $\hat{y} \in \{−1, 1\}$, for example, binary attributes: gender (Male or Female), age (Above 30 or Below 30), political affiliation (Democratic or Republican) as shown below:

$$\hat{y} = \theta^T f(v_i) + b. \quad (2.1)$$

Thus, the baseline linear classifier for user attribute classification e.g., gender is defined by (Van Durme, 2012b) and given below:

$$\hat{y} = \begin{cases} Male & \text{if } \theta \cdot f(v_i) \geq 0, \\ Female & \text{otherwise.} \end{cases} \quad (2.2)$$

The log-linear model for gender classification in Twitter applied by (Van Durme, 2012b):

$$\hat{y} = \begin{cases} Male & \text{if } [1 + \exp(-\theta \cdot f(v_i))]^{-1} \geq 0.5, \\ Female & \text{otherwise.} \end{cases} \quad (2.3)$$

D. Nguyen et al. (2011, 2013) use one versus all method to handle multi-class
classification of user’s age compared to the classification models with two classes \( \hat{y} \in \{-1, 1\} \) described above. The model estimates a conditional distribution:

\[
p(\hat{y} \mid x, \theta) = \frac{1}{1 + \exp[-(\theta^\top f(v_i) + b)]}.
\]

(2.4)

### 2.1.3 Regression

Similarly, given an input vector \( f(v_i) \in \mathbb{R}^d \), where \( f(v_i) \) represent features extracted from \( v_i \)’s self-authored communications, the goal is to predict \( \hat{y} \in \mathbb{R} \) for the exact age of a person \( y \in \mathbb{R} \) (age is a continuous variable here) using a linear regression model:

\[
\hat{y} = \beta_0 + f(c)^\top \beta,
\]

(2.5)

where \( \beta_0 \) and \( \beta \) are the parameters to estimate. For that D. Nguyen et al. (2011, 2013) use ridge with \( L_2 \) regularization implementation from LibLinear package (Rong et al., 2008) and scikit-learn library (Pedregosa et al., 2011).

Unlike prior work, this thesis uses log-linear aka logistic regression or max-entropy models rather than SVM for classification as defined in the Eq. 2.3. We use log-linear models over reasonable alternatives such as perceptron or SVM, following the practice of a range of previous work in related areas (Smith, 2004; Y. Liu et al., 2005; Poon et al., 2009) including text classification in social media (Y. Yang & Eisenstein, 2013).
2.2 Unsupervised Models

Other works for personal analytics exploit unsupervised approaches. Bergsma et al. (2013) show that large-scale clustering of user names in Twitter improves gender, ethnicity and location classification performance. O’Connor et al. (2010), following the work by Eisenstein et al. (2010), propose a Bayesian generative model to discover demographic language variations in Twitter. Rao et al. (2011) suggest a hierarchical Bayesian model which takes advantage of user name morphology for predicting user gender and ethnicity. Golbeck et al. (2010) incorporate Twitter data in a spatial model of political ideology.

More works use social media for mining political opinions (O’Connor et al., 2010; Maynard & Funk, 2012) or understanding socio-political trends and predicting voting outcomes (Tumasjan et al., 2010; Gayo-Avello, 2012; Lampos et al., 2013) is becoming a common practice. For instance, Lampos et al. (2013) propose a bilinear user-centric model for predicting voting intentions in the UK and Australia from social media data. Other works explore political blogs to predict what content will get the most comments (Yano et al., 2013) or analyze communications from Capitol Hill (http://www.tweetcongress.org) to predict campaign contributors based on this content (Yano & Smith, 2013).
CHAPTER 2. BACKGROUND: PERSONAL ANALYTICS IN SOCIAL MEDIA

2.3 Graph-based Models

Other methods characterize Twitter users by applying limited amounts of network structure information in addition to lexical features (Mislove et al., 2010; Conover, Gonçalves, et al., 2011; Pennacchiotti & Popescu, 2011a; S.-H. Yang et al., 2011; Golbeck & Hansen, 2011; Zamal et al., 2012).

Conover, Gonçalves, et al. (2011) rely on identifying strong partisan clusters of Democratic and Republican users on Twitter based on retweet and user mention degree of connectivity, and then combine this clustering information with the follower and friend neighborhood size features. Pennacchiotti and Popescu (2011a, 2011b) focus on user behavior, network structure and linguistic features. Similar to our work, they assume that users from a particular class tend to reply and retweet messages of the users from the same class. We extend this assumption and study other relationship types e.g., friends, user mentions etc.

Recent work by Ming Fai Wong et al. (2013) investigate tweeting and retweeting behavior for political learning in social media during 2012 US Presidential election.

The most similar work to ours is by Zamal et al. (2012), where the authors apply features from the tweets authored by a user’s friend to infer attributes of that user. In this thesis, we study different types of local neighborhoods (in addition to a friend network explored by Zamal et al.) and make a comparison with their results, and we are the first to propose a model for calculating personal analytics where the low-resource aspect is explicitly taken into account.
CHAPTER 2. BACKGROUND: PERSONAL ANALYTICS IN SOCIAL MEDIA

In Tables 2.1–2.3 we report classification performance for the existing discriminative fully supervised content-based models for age, gender, political preference, ethnicity, origin and race prediction. For each approach we present the number of users, the average number of tweets per user, the learning algorithm e.g., it is often SVM, feature types and the corresponding accuracy numbers.

These results allow us to estimate the bounds of prediction performance for latent user demographics and to summarize the limitations of the existing approaches to personal analytics in social media:

- the majority of them rely on unrealistic amount of content per user e.g., more than 1,000 tweets per user and on complicated features e.g., socio-linguistic, LWIC, which is less feasible in constrained-resource prediction scenario (as we have demonstrated in Figure 1.1 and Appendix B, an average Twitter user is not associated with 1,000 tweets);

- only some of them (Pennacchiotti & Popescu, 2011a; Conover, Gonçalves, et al., 2011; Zamal et al., 2012) started to explore the network structure or content from the network in order to incorporate this signal as features into the model;

- only models proposed by Cohen and Ruths (2013); Volkova et al. (2014) started to explore data sampling and annotation biases, and none of the existing models investigated the imbalanced data issues and small priors.
Table 2.1: Overview of approaches for gender prediction in social media (YouTube\(^1\), Facebook\(^2\), Google+\(^3\)). Numbers in bold highlight the accuracy of the best-performing features for a given approach.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Users</th>
<th>Av. tweets per user</th>
<th>Model</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender (Binary Classification: Male, Female)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rao et al. (2010)</strong></td>
<td>1,000</td>
<td>405</td>
<td>SVM</td>
<td>n-grams, socio-linguistic, combined</td>
<td>0.687</td>
</tr>
<tr>
<td><strong>Burger et al. (2011)</strong></td>
<td>183,875</td>
<td>22</td>
<td>Balanced Winnow 2</td>
<td>screen/full name, bio/n-grams</td>
<td>0.918</td>
</tr>
<tr>
<td><strong>Liu et al. (2012)</strong></td>
<td>400</td>
<td>1,000</td>
<td>SVM</td>
<td>n-grams, stems, behaviour, neigh</td>
<td>0.868</td>
</tr>
<tr>
<td><strong>Zamal et al. (2012)</strong></td>
<td>384</td>
<td>1,000</td>
<td>SVM</td>
<td>User-Only Nbr-Most Avg-Least Joint-Closest</td>
<td></td>
</tr>
<tr>
<td><strong>Filippova (2012)</strong>(^1)</td>
<td>4.9M train 1.2M test</td>
<td>20 - 30</td>
<td>MaxEnt</td>
<td>char/token/sent, graph, propagate</td>
<td>0.880</td>
</tr>
<tr>
<td><strong>Kosinski et al. (2013)</strong>(^2)</td>
<td>57,505</td>
<td>–</td>
<td>Linear regression</td>
<td>likes (SVD)</td>
<td>0.93 (AUC)</td>
</tr>
<tr>
<td><strong>Bergsma et al. (2013)</strong></td>
<td>33,805</td>
<td>only first, last name</td>
<td>SVM</td>
<td>n-grams clusters</td>
<td>0.895</td>
</tr>
<tr>
<td><strong>Liu and Ruths (2013)</strong></td>
<td>8,000</td>
<td>1,000</td>
<td>SVM</td>
<td>n-grams, stems, behavior, names</td>
<td>0.871</td>
</tr>
<tr>
<td><strong>Schwartz et al. (2013)</strong>(^2)</td>
<td>74,941</td>
<td>209 posts</td>
<td>SVM</td>
<td>word phrases, topics, lexicon</td>
<td>0.919</td>
</tr>
<tr>
<td><strong>Alowidhi et al. (2013)</strong></td>
<td>180,000</td>
<td>–</td>
<td>NB-Tree</td>
<td>profile color profile name</td>
<td>0.740</td>
</tr>
<tr>
<td><strong>Sap et al. (2014)</strong></td>
<td>1,000 (test)</td>
<td>100+</td>
<td>SVM</td>
<td>word unigrams</td>
<td>0.900</td>
</tr>
<tr>
<td><strong>Bamman et al. (2014)</strong></td>
<td>14,464</td>
<td>636 (9.2M)</td>
<td>Logit regression</td>
<td>10K freq. words, emotions, punct</td>
<td>0.880</td>
</tr>
<tr>
<td><strong>Fang et al. (2015)</strong>(^3)</td>
<td>2,548</td>
<td>332 (846K)</td>
<td>LSVM</td>
<td>word n-grams, socio-linguistic above + visual</td>
<td>0.723</td>
</tr>
<tr>
<td><strong>Culotta et al. (2015)</strong></td>
<td>1,000</td>
<td>200 friends</td>
<td>Logistic regression</td>
<td>friends</td>
<td>0.75 (F1)</td>
</tr>
<tr>
<td><strong>Hovy (2015)</strong></td>
<td>–</td>
<td>46.83K sentences</td>
<td>Logistic regression</td>
<td>word embeddings</td>
<td>0.609 (F1)</td>
</tr>
</tbody>
</table>
CHAPTER 2. BACKGROUND: PERSONAL ANALYTICS IN SOCIAL MEDIA

Table 2.2: Overview of approaches for age prediction in social networks (Blogs\(^1\) \(\text{www} \).blogger.com, YouTube\(^2\), Facebook\(^2\), Google+\(^3\), LiveJournal\(^1\)). Numbers in bold highlight the accuracy of the best-performing features for a given approach.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Users</th>
<th>Av. tweets per user</th>
<th>Model</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (Binary or Multi-Class Prediction)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schler et al. (2006)</td>
<td>37,478</td>
<td>[13-17, 23-27, 33-42]</td>
<td>Winnow</td>
<td>n-grams, POS, function words</td>
<td>0.762</td>
</tr>
<tr>
<td>Argamon et al. (2007) (^9)</td>
<td>19,320</td>
<td>[13-17, 23-27, 33-47]</td>
<td>Logistic regression</td>
<td>1K high inf. gain words</td>
<td>0.774</td>
</tr>
<tr>
<td>Rao et al. (2010)</td>
<td>2,000</td>
<td>1,183</td>
<td>SVM</td>
<td>n-grams, socio-linguistic, combined</td>
<td>0.731</td>
</tr>
<tr>
<td>D. Nguyen et al. (2011)</td>
<td>2,500</td>
<td>200</td>
<td>Multi-class</td>
<td>unigrams, LWIC</td>
<td>0.767 (F1)</td>
</tr>
<tr>
<td>Rosenthal and McKown (2011) (^4)</td>
<td>24,500</td>
<td>[18-22, 28-32, 38-42]</td>
<td>Logistic regression</td>
<td>n-grams, style behav, interests,</td>
<td>0.57</td>
</tr>
<tr>
<td>Zamal et al. (2012)</td>
<td>386</td>
<td>1K user 20K neigh [18-23, 25-30]</td>
<td>SVM</td>
<td>User-Only Nbr-All Avg-Least Joint-least</td>
<td>0.751</td>
</tr>
<tr>
<td>Filippova (2012) (^1)</td>
<td>4.9M train 1.2M test</td>
<td>20-30 posts [13-19, 20-29, 30+]</td>
<td>MaxEnt</td>
<td>char, token, sent, graph, propagate</td>
<td>0.700</td>
</tr>
<tr>
<td>Hovy (2015)</td>
<td>–</td>
<td>46.83K sent. [&lt; 35, &gt; 45]</td>
<td>Logistic regression</td>
<td>word embeddings</td>
<td>0.626 (F1)</td>
</tr>
<tr>
<td><strong>Age (Linear Regression: continuous)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. Nguyen et al. (2011)</td>
<td>2,500</td>
<td>200</td>
<td>Linear regression</td>
<td>unigrams, LWIC</td>
<td>3.88 (MAE)</td>
</tr>
<tr>
<td>Schwartz et al. (2013) (^2)</td>
<td>74,941</td>
<td>209 posts</td>
<td>Ridge regression</td>
<td>word phrases, topics, lexicon</td>
<td>0.84 (Pears)</td>
</tr>
<tr>
<td>Kosinski et al. (2013) (^2)</td>
<td>57,505</td>
<td>–</td>
<td>Linear regression</td>
<td>likes (SVD)</td>
<td>0.75 (Pears)</td>
</tr>
<tr>
<td>Sap et al. (2014) (^2)</td>
<td>1,000 (test)</td>
<td>100+</td>
<td>Ridge regression, L2 regul.</td>
<td>word unigrams</td>
<td>3.4 (MAE)</td>
</tr>
<tr>
<td>Fang et al. (2015) (^3)</td>
<td>2,548</td>
<td>332 (846K)</td>
<td>LSVM</td>
<td>word n-grams, socio-linguistic above + visual</td>
<td>0.598</td>
</tr>
</tbody>
</table>

35
### Table 2.3: Overview of techniques to infer ethnicity and origin on Twitter. Numbers in bold highlight the accuracy of the best-performing features for a given approach.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Users</th>
<th>Av. tweets per user</th>
<th>Model</th>
<th>Features</th>
<th>Accuracy (F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnicity (Binary: African-American or not)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pennacchiotti and Popescu (2011a)</td>
<td>6,000</td>
<td>-</td>
<td>GBDT</td>
<td>n-grams, behavior, profile, network info</td>
<td>0.758</td>
</tr>
<tr>
<td>Ethnicity (Multi-class: African-American, Hispanic, Caucasian)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Culotta et al. (2015)</td>
<td>615</td>
<td>200 friends</td>
<td>Logistic regression</td>
<td>friends</td>
<td>0.61</td>
</tr>
<tr>
<td>Ethnicity (13 classes, one vs. all strategy)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergsma et al. (2013)</td>
<td>7,457</td>
<td>only first, last name</td>
<td>SVM</td>
<td>characters token, n-grams clusters</td>
<td>0.775 0.784 0.813</td>
</tr>
<tr>
<td>Race (White vs. African-American)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergsma et al. (2013)</td>
<td>7,977</td>
<td>only first, last name, location</td>
<td>SVM</td>
<td>character n-grams clusters</td>
<td>0.816 0.824 0.846</td>
</tr>
<tr>
<td>Origin (Southern, Northern India)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rao et al. (2010)</td>
<td>1,000</td>
<td>497</td>
<td>SVM</td>
<td>n-grams, socio-linguistic, combined</td>
<td>0.729 0.771 0.738</td>
</tr>
<tr>
<td>Country (53 classes, one vs. all strategy)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergsma et al. (2013)</td>
<td>781,920</td>
<td>only first, last name, location</td>
<td>SVM</td>
<td>characters clusters</td>
<td>0.845 0.848 0.867</td>
</tr>
</tbody>
</table>

### 2.4 Streaming Approaches

Van Durme (2012b) proposed streaming models to predict user gender in Twitter. Other works suggested to process text streams for other NLP tasks e.g., real-time opinion mining and sentiment analysis in social media (Pang & Lee, 2008), named entity disambiguation (Sarmento et al., 2009), statistical machine translation (Levenberg et al., 2011), first story detection (Petrović et al., 2010), and unsu-
Table 2.4: Overview of methods for political preference classification on Twitter. Numbers in bold highlight highlight the accuracy of the best-performing features for a given approach.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Users</th>
<th>Av. tweets per user</th>
<th>Model</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rao et al. (2010)</td>
<td>400</td>
<td>4,965</td>
<td>SVM</td>
<td>n-grams, socio-linguistic, combined</td>
<td>0.828</td>
</tr>
<tr>
<td>Pennacchioti and Popescu (2011a, 2011b)</td>
<td>10.3K</td>
<td>-</td>
<td>GBDT</td>
<td>n-grams, tweeting behavior, profile, network info</td>
<td>0.889</td>
</tr>
<tr>
<td>Conover, Gonçalves, et al. (2011)</td>
<td>1,000</td>
<td>1,000</td>
<td>SVM</td>
<td>full-text hashtags clusters</td>
<td>0.949</td>
</tr>
<tr>
<td>Zamal et al. (2012)</td>
<td>400</td>
<td>1,000 9,625 10,625 10,625</td>
<td>SVM</td>
<td>User-Only Nbr-All Avg-All Joint-All</td>
<td>0.932</td>
</tr>
<tr>
<td>Cohen and Ruths (2013)</td>
<td>397 1,837 262 196</td>
<td>1,000</td>
<td>SVM LLDA</td>
<td>k-top words, stems etc. from (Zamal et al., 2012)</td>
<td>0.910</td>
</tr>
</tbody>
</table>

Supervised dependency parsing (Goyal & Daume, 2011). The Massive Online Analysis (MOA) toolkit developed by Bifet et al. (2010, 2011) is an alternative to the Jerboa package developed by Van Durme (2012a) for streaming analytics.
Chapter 3

Static Prediction

As we showed in Tables 2.1 - 2.3 of the literature review, the existing approaches to latent user attribute prediction have relied on supervised classification models learned from thousands of tweets per user. However, most Twitter users are less prolific than those examined in these works, as has been shown in the user activeness diagram in Figure 1.1, and thus do not produce the thousands of tweets required to obtain their levels of accuracy. Furthermore, we estimated from the 1% Twitter feed that the median number of tweets produced by a random Twitter user per day is 10. In addition, recent changes to Twitter API querying rates further restrict the speed of access to this resource, effectively reducing the amount of data that can be collected for a certain user in a given time period.

In this Chapter\textsuperscript{1} we propose to go beyond the existing batch models that rely on

\textsuperscript{1}This chapter extends “Inferring User Political Preferences from Streaming Communications”, which was published in the Proceedings of the Conference of the Association for Computational Linguistics (ACL) in 2014, and is joint work with Benjamin Van Durme and Glen Coppersmith.
CHAPTER 3. STATIC PREDICTION

thousands of tweets per user and take advantage of content produced within local user neighborhoods. We start by describing the data in Section 3.1.3 – Twitter social network graph constructed using six types of relationships including graph-based e.g., friends and followers, and content-based e.g., retweets, replies, user mentions and shared hashtags. In Section 3.1.4 we present more details on our data collection and annotation methodology for 5 datasets that include users annotated with political affiliation, gender and age, their immediate neighbors of different types and communications e.g., 200 tweets per user/neighbor. Throughout the thesis we use these datasets for the experiments. Relying on multiple datasets allows us to analyze biases present in the data arising from the underlying collection and annotation strategies that significantly influence performance. In Sections 3.2 and 3.3 we describe our static models learned from the content of different neighbors, outline our experimental setup and discuss baselines. We present our results in Section 3.4 and summarize them as recommendations in terms of the most predictive (a) neighbor types and (b) feature combinations for predicting user gender, age and political preferences in a classic batch setting given varied amount of content available for predictions. To the best of our knowledge, this is the first work that makes explicit the trade-off between accuracy and cost (manifest as cost to the Twitter API), and optimizes to a different trade-off than baseline approaches, seeking maximal performance when limited user content is available.
CHAPTER 3. STATIC PREDICTION

3.1 Data

3.1.1 Twitter Language and Network Properties

Twitter Language Properties The heterogeneity and streaming nature of social media makes prediction of latent user attributes a challenging task compared to e.g., Switchboard telephone speech data (Godfrey et al., 1992a). Moreover, in contrast to political debate text (Thomas et al., 2006; Yano & Smith, 2013; Yano et al., 2013) or news data (Zhou et al., 2011; O’Connor et al., 2013), average users in social media discuss political topics very sporadically compared to politically active users or political figures. This makes prediction of political preferences more difficult, as recently discussed (Cohen & Ruths, 2013; Volkova et al., 2014).

To address these issues, our approach takes advantage of content derived from various social circles of a user, as realized by the social network. Consider the case where a given user does not often tweet about their political views, but follows people whose content is more politically oriented. It is reasonable to assume that the content of the users the user follows may provide insight into the given user’s political beliefs.

Twitter Network Properties Twitter is a multi-relational network, various signals can be used to create a social network e.g., retweets, friends, each encoding different social relationships between users. For instance, retweet or user mention edges imply direct involvement in communication and may reveal shared interests or preferences between the users, whereas follower edges may have different implications.
CHAPTER 3. STATIC PREDICTION

Moreover, social networks like Twitter have homophilic properties which means that similar nodes are likely to share a link. Homophily of social media has been briefly studied for latent user attribute classification [Thelwall 2009, Zamal et al. 2012], but have been extensively investigated for other networks [Bilgic & Getoor 2010, Namata et al. 2011]. Given the homophilic nature of these networks, it is reasonable to assume that immediate neighbors of a user (the user-local neighborhood, social circle or environment [Filippova 2012]), can provide information about user preferences, interests, or other information relevant to predicting their latent attributes. In this work we analyze these principal characteristics of a Twitter network to evaluate their utility for a gender, age and political preference prediction in social media.

3.1.2 Twitter Social Network Graph

Definition 1. Let us define an attributed undirected graph $G = (V, E)$, where $V$ is a set of vertices and $E$ is a set of edges. Each vertex $v_i$ represents someone in a communication graph i.e., communicant: here a Twitter user. Each vertex is associated with a feature vector $f(v_i)$ which encodes communications e.g., tweets $T$ available for a given user $v_i$ and a corresponding attribute $a(v_i)$ – in our case it is binary for age, gender or political preference e.g., for gender $a(v_i) \in \{\text{Male}, \text{Female}\}$. Each edge $e_{ij} \in E$ represents a connection between $v_i$ and $v_j$, $e_{ij} = (v_i, v_j)$ and defines different communication relations between Twitter users e.g., graph-based – follower (l) and friend (b), and content-based – user mention (s), hashtag (h), reply (y) and retweet.
CHAPTER 3. STATIC PREDICTION

Figure 3.1: Twitter social graph constructed using content-based relationships (aka interaction network): user mention $\phi_s(E)$, retweet $\phi_w(E)$, hashtag $\phi_h(E)$ and reply $\phi_y(E)$, and graph-based relationships (aka social network): follower $\phi_l(E)$ and friend $\phi_b(E)$ between the users; blue vertices – Democratic users, red – Republican users; $f(v_i)$ and $f[N_h(v_i)]$ are user and neighbor feature vectors extracted from user and neighbor tweets, respectively.

Thus, $E \in V^{(2)} \times \{l, b, s, w, y, h\}$. We denote a set of edges of a given type as $\phi_r(E)$ for $r \in \{l, b, s, w, y, h\}$. We denote a set of vertices adjacent to $v_i$ by relationship type $r$ as $N_r(v_i)$ which is equivalent to $\{v_j \mid e_{ij} \in \phi_r(E)\}$. We refer to $N_r(v_i)$ as $v_i$’s local $r$-neighborhood.

In most cases, we only work with a sample of a neighborhood, denoted $N'_r(v_i)$ where $|N'_r(v_i)| = \text{deg}(v_i) = d$ is the size of the sampled $r$-neighborhood for $v_i$. For the experiments in this work $d$ ranges from 10 to 25. We elaborate on how $r$-neighborhoods are constructed and present an example Twitter social network graph in Figure 3.1.
CHAPTER 3. STATIC PREDICTION

3.1.3 Twitter Social Relationship Types

Here we describe how graph-based and content-based relationships were inferred for each Twitter user (Volkova, 2014). Content-based edges are derived from self-authored content (user mentions, hashtags, replies and retweets) while graph-based edges are derived from the user’s metadata (followers and friends).

3.1.3.1 Graph-based Relationships

A follower of user $v_i$ is any $v_j$ who subscribes to the tweets of $v_i$. We explore user-follower relationships by downloading lists of followers for each user $v_i$. We then uniformly at random sample 20 followers for each user and download a sample of 200 of their tweets as of the time of the sampling to construct user-follower neighborhood $G_l = [N_l(v_i), \phi_l(E)]$.

A friend is a bidirectional following relationship – $v_i$ follows $v_j$ and $v_j$ also follows $v_i$. These friends comprise the user-friend neighborhood $G_b = [N_b(v_i), \phi_b(E)]$. We sample 20 friends for each user and download a sample of 200 of their tweets.

3.1.3.2 Content-based Relationships

Twitter users interact with one another primarily in four ways – retweets, user mentions, replies and shared hashtags. User mentions allow a user $v_i$ to address another user $v_j$ directly in a public feed or just make a reference to another user in the third person, for example, “I miss @Reem_Fayek so much”. We extract all
CHAPTER 3. STATIC PREDICTION

user mentions from all self-authored content available, each such mention creates a user-mention relationship between the author and the mentioned user. Some users are ubiquitous (e.g., @youtube, @FoxNews) and, thus, were treated similarly to stopwords and removed. Specifically, we sorted all user mentions by frequency in each of the labeled collections and eliminated top most frequent (> 100) and the least frequent (< 2) user mentions. We sampled user-mentioned users based on the frequency of mention. The extracted set of user-mentioned users comprises $G_s = [N_s(v_i), \phi_s(E)]$.

Retweets allow users to rebroadcast content generated by other users to increase the content’s spread and influence, for example “RT @theellenshow: If only Bradley’s arm was longer. Best photo ever. #oscars”. More formally, a retweet allows a user $v_i$ to re-post a tweet from $v_j$ within $v_i$’s follower circle. We explore user-retweet relationships by extracting all retweets from user content. The users who were retweeted comprise $G_w = [N_w(v_i), \phi_w(E)]$.

A reply is a tweet sent in direct response to another tweet. To construct user-reply social circle $G_y = [N_y(v_i), \phi_y(E)]$ we consider tweets with filled in reply to field and extract reply user information.

Finally, to extract user-hashtag social circles $G_h = [N_h(v_i), \phi_h(E)]$, we get a sample of hashtags e.g., #tcot, #gop or #Obama and users sharing a specific hashtag as follows: for each user we eliminate hashtags that occur only once, then randomly sample a set of 5 hashtags; next, we download 100 most recent tweets per hashtag, extract user information associated with each tweet.
CHAPTER 3. STATIC PREDICTION

For all neighborhood types, there is more data to be downloaded but it is not feasible due to the rate limits on Twitter API. Thus, we randomly sample $d$ users in the neighborhood of each user to be classified. For these randomly selected neighbors, we download 200 of their tweets and use this as the respective $N'_r(v_i)$ and $\phi_r(E)$.

3.1.4 Political Preference Graphs

3.1.4.0.1 Candidate-Centric Graph

We construct candidate-centric graph $G_{cand}$ by looking into following relationships between the users and Democratic or Republican candidates during the 2012 US Presidential election. In the Fall of 2012, leading up to the elections, we randomly sampled 516 Democratic and 515 Republican users. We labeled users as Democratic if they exclusively follow both Democratic candidates – BarackObama and JoeBiden but do not follow both Republican candidates – MittRomney and RepPaulRyan and vice versa. We collectively refer to $D$ and $R$ as our “users of interest” for which we aim to predict political preference. For each such user we collect recent tweets and randomly sample their immediate $d = 10$ neighbors from follower, friend, user mention, reply, retweet and hashtag social circles.

3.1.4.0.2 Geo-Centric Graph

We construct a geo-centric graph $G_{geo}$ by collecting 135 Democratic and 135 Republican users from the Maryland, Virginia and Delaware region of the US (using the
CHAPTER 3. STATIC PREDICTION

location field in the metadata) with self-reported political preference in their biographies e.g., I am a Republican. Similar to the candidate-centric graph, for each user we collect recent tweets and randomly sample user social circles in the Fall of 2012. We collect this data to get a sample of politically less active users compared to the users from candidate-centric graph.

3.1.4.0.3 Active Graph

We also consider a $G_{\text{active}}$ graph constructed from a dataset previously used for political affiliation classification [Zamal et al., 2012]. This dataset consists of 200 Republican and 200 Democratic users associated with 925 tweets on average per user. Each user has on average 6155 friends with 642 tweets per friend. Sharing restrictions and rate limits\(^2\) on Twitter data collection only allowed us to recreate a semblance of their data – 371 out of 400 users labeled with political orientation as shown in Table 3.1. Political labels were extracted from http://www.wefollow.com following the approach by (Pennacchiotti & Popescu, 2011b).

3.1.5 Gender and Age Graphs

We replicate and significantly extend previously used dataset for user attribute classification including gender – 192 Male, 191 Female users and age – 190 “18 - 23”,

\(^2\)This inability to perfectly replicate prior work based on Twitter is a recognized problem throughout the community of computational social science, arising from the data policies of Twitter itself, it is not specific to this work.
CHAPTER 3. STATIC PREDICTION

191 “25 - 30” users \cite{Zamal2012}. This data, collected in 2012, contains users labeled with attributes, user tweetIDs (up to 1K per user), user-to-friend mappings and friend tweetIDs. Sharing restrictions on the data prevented us from doing a direct comparison, but we were able to recreate almost all of the dataset, excluding a subset of user profiles that were deleted or became private since the original collection was formed. Specifically, we recovered 383 out of 384 users labeled with gender, 381 out of 386 users labeled with age as reported in the second column of Table 3.1. We also report the number of users with at least one neighbor per user of each type (follower \( l \), friend \( b \), mention \( s \) and retweet \( w \)). In Table 3.2, we present the means and standard deviations for the number of neighbors of each type per user.

We briefly review the method for obtaining labels for the data, and refer the reader to \cite{Zamal2012} for further details. Gender labels were extracted using 100 most common baby boy and girl names on record with US social security department in 2011, following the technique suggested by \cite{Mislove2011}. Age labels were extracted from user accounts with announced birthdays e.g., “Happy # th/st/nd/rd birthday to me”. These annotation strategies include annotation biases as discussed in section 1.3.

\footnote{The original dataset with userIDs and friendIDs is available at \url{http://icwsm.cs.mcgill.ca/}.}

\footnote{Twitter policy restricts to sharing only tweet IDs rather than complete tweets. Therefore, we downloaded user and friend tweets using their tweet IDs by querying Twitter API as described in Appendix C. As of August 2015, a registered developer account on Twitter allows to make 450 Twitter API calls every 15 min to get 450 tweets using tweetIDs.}
CHAPTER 3. STATIC PREDICTION

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Users</th>
<th>$v$</th>
<th>Follower $l$</th>
<th>Friend $b$</th>
<th>Retweet $w$</th>
<th>Mention $s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>384</td>
<td>383</td>
<td>302</td>
<td>378</td>
<td>243</td>
<td>325</td>
</tr>
<tr>
<td>Age</td>
<td>386</td>
<td>381</td>
<td>310</td>
<td>359</td>
<td>217</td>
<td>324</td>
</tr>
<tr>
<td>Political</td>
<td>400</td>
<td>371</td>
<td>360</td>
<td>378</td>
<td>332</td>
<td>353</td>
</tr>
</tbody>
</table>

Table 3.1: The number of recovered users $v$ from the original dataset, and a subset of these users with at least one neighbor per user – $l$ followers, $b$ friends, $w$ retweets, and $s$ user mentions.

<table>
<thead>
<tr>
<th>Type</th>
<th>Follower $l$</th>
<th>Friend $b$</th>
<th>Retweet $w$</th>
<th>Mention $s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N^G_v$</td>
<td>19 ± 3</td>
<td>19 ± 3</td>
<td>16 ± 4</td>
<td>8 ± 7</td>
</tr>
<tr>
<td>$T^G_v$</td>
<td>157 ± 79</td>
<td>193 ± 92</td>
<td>215 ± 72</td>
<td>217 ± 63</td>
</tr>
<tr>
<td>$N^A_v$</td>
<td>19 ± 7</td>
<td>18 ± 4</td>
<td>9 ± 7</td>
<td>20 ± 3</td>
</tr>
<tr>
<td>$T^A_v$</td>
<td>170 ± 100</td>
<td>174 ± 74</td>
<td>212 ± 66</td>
<td>212 ± 69</td>
</tr>
<tr>
<td>$N^P_v$</td>
<td>19 ± 1</td>
<td>40 ± 20</td>
<td>19 ± 7</td>
<td>22 ± 9</td>
</tr>
<tr>
<td>$T^P_v$</td>
<td>181 ± 110</td>
<td>180 ± 66</td>
<td>221 ± 67</td>
<td>220 ± 69</td>
</tr>
</tbody>
</table>

Table 3.2: Mean ± standard deviation for the number of neighbors per user ($N_v$) and the number of tweets per neighbor ($T_v$) for gender ($G$), age ($A$), and political affiliation ($P$) for $G_{active}$.

### 3.2 Batch Models

#### 3.2.1 Baseline User Model

As input we are given a set of vertices representing users of interest $v_i \in V$ along with feature vectors $f(v_i)$ derived from content authored by the user of interest. Each user is associated with a non-zero number of publicly posted tweets $T$. Our goal is to assign to a category each user $v_i$ based on $f(v_i)$. Here we focus on a binary assignment into the categories e.g., Democratic $D$ or Republican $R$. As we briefly outlined in
Chapter 3. Static Prediction

Chapter 2 in the Eq. 2.3 the log-linear model for such binary classification is:

\[
\Phi_{v_i} = \begin{cases} 
D & (1 + \exp[-\theta \cdot f(v_i)])^{-1} \geq 0.5, \\
R & \text{otherwise}.
\end{cases}
\tag{3.1}
\]

where features are normalized word n-gram counts extracted from \(v_i\)’s tweets \(f(v_i)\).

The proposed baseline model follows the same trends as the existing approaches for user attribute classification in social media as described in Tables 2.1 – 2.3 in Chapter 2.

Let’s see how to represent a tweet in terms of user features \(f(v_i)\). For instance, a female tweet from Figure 1.3 “Delighted I kept my Xmas vouchers – Happy Friday to me :-) #shopping” is represented as a binary feature vector below:

<table>
<thead>
<tr>
<th>:-)</th>
<th>#shopping</th>
<th>…</th>
<th>delighted</th>
<th>…</th>
<th>…</th>
<th>happy</th>
<th>…</th>
<th>kept</th>
<th>…</th>
<th>me</th>
<th>…</th>
<th>my</th>
<th>…</th>
<th>to</th>
<th>…</th>
<th>vouchers</th>
<th>…</th>
<th>xmas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

To make prediction for a given user we aggregate the most recent \(k\) tweets per user, thus, the feature vector \(f(v_i)\) becomes more dense; when the count-based (frequency) features are used instead of binary features, 0s and 1s in the feature vector \(f(v_i)\) are converted to word frequencies.

Next we propose to extend the baseline model by taking advantage of language in user social circles (neighborhoods) as described below.
CHAPTER 3. STATIC PREDICTION

3.2.2 Neighbor Model

As input we are given user-local neighborhood $N_r(v_i)$, where $r$ is a neighborhood type. Besides the neighborhood’s type $r$, each neighbor is characterized by:

- the number of tweets per neighbor $t \in T$, $t = \{5, 10, 15, 25, 50, 100, 200\}$;
- the order of the neighborhood (aka social circle) – the number of neighbors per user of interest $|N_r| = \text{deg}(v_i)$, $d = \{1, 2, 5, 10\}$.

Our goal is to classify users of interest using evidence e.g., tweets aggregated over neighbors $\sum_d T^{(d)}$ within the local neighborhood converted into neighbor features $f[N_r(v_i)] \equiv f(N_r)$ as e.g., Democratic or Republican. The corresponding log-linear model is defined as:

$$\Phi_{N_r} = \begin{cases} 
D & (1 + \exp[-\theta \cdot f(N_r)])^{-1} \geq 0.5, \\
R & \text{otherwise.}
\end{cases} \quad (3.2)$$

To check whether our static models can be effectively used in a constrained-resource setting we compare the performance of the user model defined in Eq.3.1 and the neighborhood model defined in Eq.3.2. As we present in Appendix C, we see the scarce available resources as the number of requests allowed per day to the Twitter API. Here we abstract this to a model assumption where we receive one tweet $t_k$ at a time and aim to maximize performance with as few tweets per user as possible.\(^5\)

\(^5\)As we discuss in Appendix B, many Twitter users simply don’t tweet very often. For instance, 85.3% of all Twitter users post less than one update per day as reported at [http://www.sysomos.com/insidetwitter/](http://www.sysomos.com/insidetwitter/)
CHAPTER 3. STATIC PREDICTION

- for the baseline user model:

  \[
  \arg\min_{t_k} \sum_k t_k^{(v_i)},
  \tag{3.3}
  \]

- for the neighborhood model:

  \[
  \arg\min_{t_k} \sum_d \sum_k t_k^{[N_r(v_i)]}, d = \text{deg}|N_r(v_i)|.
  \tag{3.4}
  \]

3.2.3 Joint User-Neighbor Model

For the user-neighbor model (User+Nbr), the joint feature vector \( f(N_r, v_i) \) is derived from both user and neighbor communications. One might try treating user and neighbor n-grams using disjoint prefixed features e.g., \( u_{_\text{nbox}} \) and \( n_{_\text{xbox}} \). However, we found that non-prefix shared features between users and their neighbors lead to higher performance. Thus, we extract shared user and neighbor features. The corresponding User+Nbr model is defined similarly to the Eq. 3.1 and 3.2.

3.3 Experiments

We design a set of experiments to analyze static models for inferring user latent attributes in social media. We evaluate our models for classifying user gender (Male
CHAPTER 3. STATIC PREDICTION

or Female), age (18-23 or 25-30 years old) and political preferences (Democratic or Republican) on Twitter and compare them with the models from (Zamal et al., 2012). For all experiments we use the LibLinear package (Rong et al., 2008) integrated in the Jerboa toolkit (Van Durme, 2012a).

To predict an attribute for a given user, we apply standard supervised classification setting with 10-fold cross-validation to a Twitter social graph $G$ as shown in Figure 3.1. For each fold we train a classifier on a random partition of a training graph $G^{\text{train}}$ and predict attributes for the users (vertices) in a test graph $G^{\text{test}}$ in batch setting. We evaluate our models as described below.

User Model We train the baseline user model using lexical features extracted from self-authored content including binary unigram (bin) and normalized count-based unigram (uni), bigram (bi) and trigram (tri) features. Moreover, to estimate how many self-authored communications is sufficient for reasonable accuracy classifying gender, age and political preferences of a user, we vary the number of tweets used per user $t = [5,..,1000]$. Since our User model relies exclusively on self-authored communications it can only be applied when at least some self-authored tweets are available for a test user.

Neighbor Model When no self-authored data is available for a given test user, we can learn a classifier using the discourse from user-local neighborhood of different types represented by follower, friend, user mention or retweet social circles. We learn our neighbor model using both binary and count-based n-gram features extracted
CHAPTER 3. STATIC PREDICTION

from tweets in user social circles of different types. We measure the effect of the following parameters on classification performance: the number of tweets sampled per neighbor $t = [5, ..., 200]$ and the number of neighbors per user $d = [2, ..., 20]$.

**User-Neighbor Model** We learn another model using features from user self-authored communications combined with the discourse from user social circles. Our User+Nbr model assumes that both kinds of communications available for a given test user. Similar to our neighbor model we vary the size of the neighborhood $d$ and the amount of tweets $t$ for each user/neighbor. To estimate accuracy and variance of classification performance, we conduct 100 replicated experiments for each parameter setting – each value of $t$ and, if applicable, the Cartesian product with each value of $d$. For each replicate we randomly sample (without replacement) $t$ tweets from all available tweets per user/neighbor.

### 3.3.1 Comparisons and Baselines

As a first baseline we apply the majority class model that learns the prior distribution over binary classes for three attributes: 0.501 for gender (majority class: Male), 0.506 for age (majority: 18-23 years old), and 0.520 for political preference for $G_{active}$ (majority: Democratic). As a second baseline we consider the basic user model and compare both neighbor and user-neighbor models against it.

Finally, we compare our User, Nbr and User+Nbr models for gender, age and political preference classification with the models learned from the same data (Zamal et al.)
Their models include user UserOnlyZLR, neighbor Nbr-ClosestZLR and user-neighbor Joint-ClosestZLR trained on k-top differentiating words, bigrams, trigrams, stems, co-stems, hashtags, frequency statistics, retweeting tendency, neighborhood size using SVM classifier. They restricted their analysis to friendship relationships and with a much larger pool of resources than we do: their equivalent of Nbr-AllZLR model is trained on approximately 6,000 friends per user with 642 tweets per friend.

3.4 Results

We report accuracy results for gender (Male or Female), age (18-23 or 25-30 years old) and political preference (Democratic or Republican) classification on Twitter and compare them with the models from (Zamal et al. 2012) in Table 3.3. We present more detailed experimental results for the User model in Figure 3.2. We outline concrete recommendations for the Nbr and User+Nbr models regarding (a) the types of neighbors, and (b) features given the amount of content available per user e.g., $t = [5, .., 200]$ tweets from $d = [2, .., 20]$ neighbors in Figure 3.8.

3.4.1 Age Classification

Figure 3.2a shows that our baseline User model trained on count-based word bigram and trigram features yields lower results compared to unigram features regardless of the number of tweets per user $t = [5, \ldots, 1000]$. Moreover, when the
CHAPTER 3. STATIC PREDICTION

Figure 3.2: Classification results for age (a), gender (b) and political preference prediction on $G_{active}$ (c) as a function of the number of tweets per user.

number of tweets is less than 150, User\textsuperscript{uni} significantly outperforms User\textsuperscript{bin}. When the number of tweets is higher, User\textsuperscript{bin} is better than User\textsuperscript{uni} and yields the same results as UserOnly\textsuperscript{ZLR} models.
CHAPTER 3. STATIC PREDICTION

The User model with binary features yields the highest accuracy (0.774) compared to the best performing neighbor (0.688) and user-neighbor (0.709) models. Our Nbr model demonstrates reasonable performance compared to the User model with follower neighbors yielding the highest $\text{Nbr}_{\text{follower}}=0.688$ and $\text{Nbr}_{\text{follower}}=0.720$ and user mention neighbors the lowest $\text{Nbr}_{\text{mention}}=0.541$ and $\text{Nbr}_{\text{mention}}=0.598$ accuracy for binary and count-based features, respectively. These results are very encouraging for latent attribute prediction because our Nbr model is exclusively learned from tweets in the social circles of a user, and thus, minimally relies on user’s self-authored content.

Interestingly, when combining user tweets with the tweets from user neighborhoods we observe that retweets lead to the best performance compared to the other types of neighbors – User+Nbr$_{\text{retweet}}=0.709$ for binary and User+Nbr$_{\text{retweet}}=0.738$ for count-based features. Moreover, we find that our User and Nbr models yield statistically significantly higher results compared to UserOnly$_{\text{ZLR}}$ and Nbr-Closest$_{\text{ZLR}}$ models trained using more advanced features. To test statistical significance we use a one-tailed single-sample $t$-test. The results in Table 3.3 marked with ↑ are statistically significantly higher (p-value $\leq 0.05$) than the corresponding accuracies in (Zamal et al., 2012).

3.4.2 Gender Classification

Figure 3.2b demonstrates accuracy results for gender classification using the baseline User model. Similarly to age classification, we found that word unigram features
CHAPTER 3. STATIC PREDICTION

Table 3.3: Classification results for gender, age and political preference on $G_{active}$. Predictions were made with user, neighbor and user-neighbor models trained using binary vs. count-based features and compared to the corresponding UserOnly$_{ZLR}$, Nbr-Closest$_{ZLR}$ and Joint-Closest$_{ZLR}$ models from Zamal et al. (2012).

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Age (Majority class)</th>
<th>Gender (User)</th>
<th>Political $G_{active}$ (UserOnly$_{ZLR}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.501</td>
<td>0.506</td>
<td>0.520</td>
</tr>
<tr>
<td>User</td>
<td>0.774 ↑</td>
<td>0.705</td>
<td>0.820 ↑</td>
</tr>
<tr>
<td>UserOnly$_{ZLR}$</td>
<td>0.751</td>
<td>0.795</td>
<td>0.890</td>
</tr>
<tr>
<td>Nbr$_{follower}$</td>
<td>0.688</td>
<td>0.720 ↑</td>
<td>0.534</td>
</tr>
<tr>
<td>Nbr$_{friend}$</td>
<td>0.655</td>
<td>0.679</td>
<td>0.609 ↑</td>
</tr>
<tr>
<td>Nbr$_{retweet}$</td>
<td>0.556</td>
<td>0.633</td>
<td>0.547</td>
</tr>
<tr>
<td>Nbr$_{mention}$</td>
<td>0.541</td>
<td>0.598</td>
<td>0.594</td>
</tr>
<tr>
<td>Nbr-Closest$_{ZLR}$</td>
<td>0.716</td>
<td>0.598</td>
<td>0.895</td>
</tr>
<tr>
<td>User+Nbr$_{follower}$</td>
<td>0.695</td>
<td>0.721</td>
<td>0.695</td>
</tr>
<tr>
<td>User+Nbr$_{friend}$</td>
<td>0.707</td>
<td>0.678</td>
<td>0.620</td>
</tr>
<tr>
<td>User+Nbr$_{retweet}$</td>
<td>0.709</td>
<td>0.738</td>
<td>0.667</td>
</tr>
<tr>
<td>User+Nbr$_{mention}$</td>
<td>0.630</td>
<td>0.673</td>
<td>0.672</td>
</tr>
<tr>
<td>Joint-Closest$_{ZLR}$</td>
<td>0.772</td>
<td>0.802</td>
<td>0.915</td>
</tr>
</tbody>
</table>

yield higher accuracy compared to bigrams and trigrams. Binary unigram features User$^{bin}$ are more effective than unigram count-based features User$^{uni}$ when the number of tweets is more than 400, and feature vectors representing users are less sparse. Moreover, User$^{bin}$=0.820 accuracy is statistically significantly better than the corresponding results for UserOnly$_{ZLR}$=0.795.

Communications from a user’s friends are the most predictive of gender compared to all other types of neighborhoods – Nbr$_{b}$=0.609 for binary and Nbr$_{c}$=0.628 for count-based features. Moreover, our neighbor models yield statistically significantly higher results compared to the corresponding neighbor model Nbr-Closest$_{ZLR}$=0.598. However, for our joint User+Nbr model all neighborhood types including friends, fol-
CHAPTER 3. STATIC PREDICTION

lowers, retweets and user mentions achieve lower results compared to Joint-Closest\textsubscript{ZLR} = 0.802. It may suggest that similar to age attribute gender attribute has low assortativity (the degree to which users are surrounded by those sharing the same attribute value) when a joint User+Nbr model is learned from both user and neighbor communications. Finally, similarly to age classification, count-based features lead to higher performance compared to binary features for Nbr and User+Nbr models for gender prediction, but the opposite is found for the User model.

3.4.3 Political Preference Classification

Figure 3.2c shows accuracy results for political preference classification for \( G_{active} \) using the baseline user model. In contrast to gender and age attribute inference:

- binary unigram features User\text{bin} outperform normalized count-based features User\text{uni} regardless the number of tweets available per user (this trend continues for the Nbr and User+Nbr models as shown in Table 3.3);

- count-based User\text{bi} model outperforms the unigram when the number of tweets per user is more than 350.

Tweets from friend and retweet neighborhoods Nbr\textsubscript{w}=0.913 are the most predictive of user political preferences when no self-authored tweets exist. These results are significantly better that the corresponding results for Nbr-Closest\textsubscript{ZLR}=0.895.

Content from user mention neighborhood yield the best accuracy when user and neighbor content are combined User+Nbr\textsubscript{w}=0.922. It is 3.6% higher than the User
CHAPTER 3. STATIC PREDICTION

baseline model, 0.9% higher than the the best neighbor model and 0.7% higher than Joint-Closest_{ZLR}=0.915 model. These results further confirm that political preference attribute has higher assortativity level compared to gender and age attributes.

Figure 3.3: Difference $\Delta d$ in $\Phi_{vi}$ classification decisions on a random partition of a candidate-centric graph while modeling the amount of content per user for 5 (blue circle) vs. 100 (green triangle) tweets. (A) $\Delta d$ with correctly classified (filled) and misclassified (not filled) markers; (B) averaged accuracies for batched classification decisions; (C) $\Delta d$ between two independent runs (top), and the average of two runs and true class (bottom); (D) fraction of correctly classified (dark) vs. misclassified (light) users.
CHAPTER 3. STATIC PREDICTION

3.4.4 User Content Influence

We investigate differences in classification decision for our user model $\Phi_{v_i}$ by making a prediction on a random set of 5 (constrained-resource setting) vs. 100 tweets per user. For that purpose, we take a random partition of the candidate-centric graph and perform two independent classification experiments (run 1 and run 2) using $t^{(v_i)} = 5$ and $t^{(v_i)} = 100$ tweets.

Figure 3.3A demonstrates that more tweets during prediction lead to higher accuracy by showing that more users with 100 tweets are correctly classified. Moreover, a lot of users with 100 tweets are close to 0.5 decision probability which suggests that the classifier is just uncertain rather than being completely off e.g., users with 5 tweets are close to 0 or 1 probability. To analyze this phenomena further we produce Figure 3.3B where we group users into 10 batches which corresponds to a decision probability from 0 to 1 e.g., 0-0.1, 0.1-0.2 and calculate an average accuracy level in each batch. The corresponding accuracy results shown as four fitted lines for four independent experiments trained using 5 vs. 100 tweets overlayed with the actual decision points: □ – 100, ◆ – 5 tweets.

Figure 3.3C shows that $\Delta d$ between two independent runs for users with 5 vs. 100 tweets and $\Delta d$ between the average of two runs and a true class is higher for 5 tweets compared to 100 tweets, which further confirms more content is preferable to make a prediction. Finally, we present the fraction of correctly classified vs. misclassified users in Figure 3.3D. We find that for our low-resource setting with 5 tweets and
CHAPTER 3. STATIC PREDICTION

100 tweets the number of correctly classified points is higher than the number of misclassified points at every decision probability cutoff.

3.4.4.1 Neighbor Content Influence

Here we study the influence of the neighborhood type $r$ and size in terms of the number of neighbors $n$ and tweets $t$ per neighbor using our static neighborhood model for political preference prediction.

In Figure 3.4 we present accuracy results for $G_{cand}$ and $G_{geo}$ graphs. Following Eq.3.3 and 3.4 we spent an equal amount of resources to obtain 100 user tweets and 10 tweets from 10 neighbors. We annotate these “points of equal number of communication” with a line on top marked with a corresponding number of user tweets. We show that three of six social circles – friend, retweet and user-mention yield better accuracy compared to the user model for all graphs when $t \geq 250$. Thus, for effectively classifying a given user $v_i$ it is better to take 200 tweets each from 10 neighbors rather than 2,000 tweets from the user.

The best accuracy for $G_{cand}$ is 0.75 for friend, follower, retweet and user-mention neighborhoods which is 0.03 higher than the user baseline for $G_{cand}$; for $G_{geo}$ is 0.67 for user-mention and 0.64 for retweet circles compared to 0.57 for the user model. However, these accuracies are much lower compared to the results from $G_{active}$ shown in Table 3.3. These extreme divergences in performance demonstrate how dataset annotation and sampling biases influence the classification performance and justify
CHAPTER 3. STATIC PREDICTION

Figure 3.4: Accuracy of model predictions for political preference on $G_{cand}$ and $G_{geo}$ graphs as a function of the number of tweets per neighbor for smaller ($n = 2$; a and c) and larger ($n = 10$; b and d) neighborhood size.

the necessity to report prediction results on multiple datasets. Finally, we show that increasing the number of tweets per neighbor from 5 to 200 leads to a significant gain in performance for all neighborhood types.
CHAPTER 3. STATISTICAL PREDICTION

Figure 3.5: Accuracy of model predictions for political preference on $G_{\text{cand}}$ and $G_{\text{geo}}$ graphs as a function of the number of tweets per neighbor for smaller ($t = 5$; a and c) and larger ($t = 200$; b and d).
CHAPTER 3. STATIC PREDICTION

3.4.4.2 Neighborhood Size Influence

In Figure 3.5 we present accuracy results to show neighborhood size influence on classification performance for $G_{geo}$ and $G_{cand}$ graphs. Our results demonstrate that even small changes to the neighborhood size $n$ lead to better performance which does not support the claims by (Zamal et al., 2012). We demonstrate that increasing the size of the neighborhood leads to better performance across six neighborhood types. Friend, user mention and retweet neighborhoods yield the highest accuracy for all graphs. We observe that when the number of neighbors is $n = 1$, the difference in accuracy across all neighborhood types is less significant but for $n \geq 2$ it becomes more significant.

3.4.4.3 Comparing User and Neighbor Models

In Figure 3.6 we report accuracy results for the best performing neighborhoods for political preference prediction and their cost following the Eq.3.4 for $G_{cand}$ and $G_{geo}$ graphs. Each line represents the cost of querying Twitter API for “20-25”, “50”, “100-125” or “200-250” tweets. For instance, “20-25” tweet line demonstrate the difference in performance for 50 tweets from 1 neighbor, 25 tweets from 2 neighbors, 10 tweets from 5 neighbors, and 5 tweets from 10 neighbors. We further confirm that:

- increasing the number of tweets per neighbor and the number of neighbors per user improves performance;
- querying for more neighbors per user is more beneficial than querying for addi-
CHAPTER 3. STATIC PREDICTION

...tional content from the existing neighbors, e.g., 5 tweets from 10 neighbors is better than 25 tweets from 2 neighbors or 50 tweets from 1 neighbor.

Furthermore, the empirical results from Figures 3.4 and 3.5 demonstrate that our neighborhood model is more cost-effective than the user model, e.g., to achieve 0.72

Figure 3.6: Accuracy of political preference predictions as a function of number of neighbors using the best-performing neighborhoods on $G_{\text{can}}$ (top) and $G_{\text{geo}}$ (bottom). At left, accuracy of predictions made from friends. At right, accuracy of predictions made from retweets (top, $G_{\text{can}}$) or user mentions (bottom, $G_{\text{geo}}$). Lines represent the corresponding cost in number of tweets. For example, the “20–25” tweet lines represent accuracy obtained with 50 tweets from one neighbor, 25 tweets from 2 neighbors, 10 tweets from 5 neighbors, or 5 tweets from 10 neighbors.
accuracy for $G_{\text{cand}}$ we need to issue at least 250 calls to get user tweets compared to only 150 calls for tweets from the local neighborhood. For $G_{\text{geo}}$, the neighborhood model is as effective as the user model. To achieve 0.57 accuracy we need to query either 50 self-authored tweets or 50 tweets from the local neighborhood.

### 3.4.5 Model Generalization

In section 3.4.3 we applied our models for political preference classification to three different datasets: candidate-centric, geo-centric and active. In Figure 3.7 we summarize classification results obtained using user, neighbor and user-neighbor models to demonstrate variability in prediction quality.

As has been shown by Volkova et al. and Cohen and Ruths user attribute prediction task is not easy! – prediction quality ranges from 0.57 to 0.69 for the baseline.
user model, and 0.72 - 0.87 for the neighbor model. Such variability can be explained by data sampling and annotation biases. The users in our Active graph are politically active e.g., talk about politics a lot; users in our candidate-centric graph are moderately active and follow political figures e.g., Obama, Romney; whereas our Geocentric graph includes users who report their political preferences in the biography field, but express their political opinions significantly less. That’s why prediction quality for these users are significantly lower. To fix for that, the first step is to test models on different datasets as we demonstrate here. The next step is to avoid biases while sampling Twitter users as we do in Chapter 8. And finally, more importantly, creating benchmark datasets for testing competing models is necessary.

3.4.6 Recommendations for Constrained-Resource Batch Prediction

As a summary of our findings, we provide recommendations for (1) neighbor and (2) feature types e.g., “friend.count” for predicting different attributes depending on the amount of content available in Figure 3.8. For example, for political preference classification using User+Nbr model it is better to use “friend.counts” – count-based features from $k > 17$ friends when the number of tweets is $t \leq 10$; when $10 < t \leq 135$ – binary features from retweet neighbors; when $t > 135$ – binary features from user mention neighbors.
CHAPTER 3. STATIC PREDICTION

Figure 3.8: Best performing neighborhoods recommended for gender, age and political preference classification using Nbr (left) vs. User+Nbr models (right).
CHAPTER 3. STATIC PREDICTION

(a) 18 - 23 y.o
(b) 25 - 30 y.o.
(c) Male
(d) Female
(e) Democratic
(f) Republican

Figure 3.9: Top-ranked features learned from user communications for gender, age and political preference prediction.

Similar to other work on predictive analytics in social media \cite{Van-Durme:2012:PhD,Schwartz:2013,Sap:2014} and other domains \cite{Boulis:2005,Garera:2009}, we visualize highly predictive unigrams for each class using word clouds – top ranked features
CHAPTER 3. STATIC PREDICTION

learned from user tweets for gender, age and political preference in Figure 3.9. In addition, we present detailed results on how accuracy depends on (a) the feature type (binary vs. counts), (b) the neighbor type (friend, follower retweet, mention) and (c) the content type (only neighbor vs. joint user-neighbor) for each attribute including gender, age and political preferences in Appendix E, Figures E.3 – E.6.

3.5 Conclusions

In Chapter 3 we experimented with static models for user attribute prediction in social media for gender, age and political preference attributes. We first implemented and evaluated the predictive power of several baseline user models learned from user tweets. We then proposed to use neighbor and joint user-neighbor models and empirically evaluated (a) what neighbors are the most predictive of each attribute and (b) how the content should be acquired within the local neighborhood (exploration vs. exploration strategy) in the constrained-resource prediction scenario. Our results demonstrate that neighbor models are as effective or even more effective for some attributes as the baseline user models. In other words, neighbors “give us away” in social networks, also known as homophily. Moreover, we observed that the predictive power of the neighborhood depends on the attribute type – friends are the most predictive of user gender, followers of age and retweets and mentions of political preferences. Moreover, in the constrained-resource prediction scenario when a user has no
CHAPTER 3. STATIC PREDICTION

or limited content, it is better query for the additional content from more neighbors per user (exploring the neighbors) rather than querying for more content from the existing neighbors (exploiting the neighbors).

We experimented with a variety of features and outlined concrete recommendations regarding (a) the feature types, (b) the neighbor types and (c) the amount of content – the number of tweets per user one should get to obtain certain levels of accuracy for each given attribute. We found that binary features are better to be used while predicting gender and age and count-based features for political preferences.

We also addressed the problem of model generalization caused by sampling and annotation biases by experimenting with multiple datasets. We showed that the prediction quality not only depends on the amount of tweets available during prediction, but also on the way the dataset was sampled. For instance, we showed that the accuracy for political preference classification ranges between 0.57 - 0.92 when applying the same models to different datasets with more or less politically active users.

Our key findings on the importance of user neighbors for user attribute classification are presented below.

- **Language of the most predictive neighbors varies among attributes.**

  Table 3.3 and Figures 3.4 – 3.8 show that followers yield the best accuracy for age Nbr_f = 0.720, friends for gender Nbr_b = 0.628, and retweets or usermentions for political preference prediction Nbr_w = 0.913 for $G_{active}$, Nbr_w = 0.75 for $G_{cand}$, Nbr_m = 0.67 for $G_{geo}$.
• **Combining user and neighbor content improves performance only for some attributes.** Table 3.3 results show that when user data is available and can be combined with neighbor data, retweets yield the best accuracy for age User+Nbr\(_{f}\)=0.738, user mentions for both gender Nbr\(_{f}\)=0.734 and political preference Nbr\(_{m}\)=0.922 prediction. Despite our joint User+Nbr model always outperforms Nbr model for all feature types, neighbors and attributes, we found that for gender and age the combination of user and neighbor features does not improve performance compared to the baseline user model. Only for political preference prediction User+Nbr model yields 3.6% gain over user baseline.

• **Attribute assortativity is important when relying on neighborhood content.** Assortativity measures the degree to which a user is exclusively surrounded by users like him or different users. For example, political affiliation attribute has been reported to be highly assortative (McPherson et al., 2001; Conover, Ratkiewicz, et al., 2011). Our results also demonstrate the benefit from including neighborhood content into our classification framework. However, gender and age attributes have low degree of assortativity (Thelwall, 2009; Zamal et al., 2012). Thus, relying on features from neighbors leads to much lower or no accuracy improvements.

• **Spotting lexical variations for user attribute prediction is more challenging for some attributes than others.** An interesting observation is
CHAPTER 3. STATIC PREDICTION

that for gender and age attributes count-based features are more predictive
than binary features extracted from either user or neighbor communications. It
means that frequencies of certain words e.g., shopping, kids vs. sports, money
for gender or lmao, school, partying vs. homie, love, family for age mentioned
by a user or a neighbor matter more than just the presence of these words in
the discourse. However, for political preference prediction, the presence of a
particular unigram e.g., #teaparty, tcot vs. LGBT, #solar is more effective for
distinguishing Democratic from Republican users.

In Chapter 4 we go beyond batch models and formulate user attribute prediction
as a streaming task. Such formulation allows us to make predictions from a stream of
communications and, therefore, takes into account the dynamic (streaming) nature of
social media. Unlike the existing approaches it does not assume access to thousands
of messages per user and makes iterative predictions from user or user-neighbor tweet
streams in real time as soon as the new tweet arrives.

\footnote{Men also talk about kids but less frequently than women; and women talk about sports but less
frequently than men.}
Chapter 4

Streaming Online Prediction

In this Chapter\textsuperscript{1} we go beyond batch models that are cognizant of a constrained-resource prediction setting and maximize the efficiency of content in calculating personal analytics, by formulating user attribute inference as a streaming prediction task. We propose streaming models that dynamically update beliefs about user attributes based on the stream of communications, as was previously motivated by Van Durme (2012b). Such models better capture the real-time nature of evidence being used for attribute prediction, and take advantage of dynamic lexical content available from joint user-neighbor communication streams. We define our streaming models in section 4.1, outline experimental setup in section 4.2, and present results in section 4.3. We also contrast the performance of streaming models with the batch models from Chapter 3 and outline our key findings in section 4.5.

\textsuperscript{1}This chapter presents “Inferring User Political Preferences from Streaming Communications”, which was published in the Proceedings of the Conference of the Association for Computational Linguistics (ACL) in 2014, and is joint work with Benjamin Van Durme and Glen Coppersmith.
CHAPTER 4. STREAMING ONLINE PREDICTION

4.1 Bayesian Iterative Update Models

To simulate the real-world streaming prediction scenario we rely on straightforward Bayes rule updates as shown in Figure 4.1. The model makes predictions under the assumption of sequentially arriving, independent and identically distributed observations $T = (t_1, \ldots, t_k)$.

The model dynamically updates posterior probability estimates $p(a(v_i) = R \mid t_k)$ for a user $v_i$ as the additional evidence $t_k$ is acquired, as defined in a general form below for any attribute $a(v_i) \in A$ given the tweets $T$ of user.

$$p(a(v_i) = x \in A \mid T) = \frac{p(T \mid a(v_i) = x) \cdot p(a(v_i) = x)}{\sum_{y \in A} p(T \mid a(v_i) = y) \cdot p(a(v_i) = y)}$$

where $y$ is the number of all attribute values, and $k$ is the number of tweets per user.

For example, to predict user political preference, we start with a balanced prior

[Given the dynamic character of online discourse it will clearly be of interest in the future to consider models that go beyond the iid assumption.]
CHAPTER 4. STREAMING ONLINE PREDICTION

\[ P(R) = P(D) = 0.5, \]

and sequentially update the posterior \( p(R | T) \) by accumulating evidence from the likelihood \( p(t_k | R) \)\(^3\)

\[
p(R | T) = \frac{\prod_k p(t_k | R) \cdot p(R)}{\prod_k p(t_k | R) \cdot p(R) + \prod_k p(t_k | D) \cdot p(D)}
\]

\[
= \frac{\prod_k p(R | t_k)p(t_k)p(R)^{-1} \cdot p(R)}{\prod_k p(R | t_k)p(t_k)p(R)^{-1} \cdot p(R) + \prod_k [p(D | t_k)p(t_k)p(D)^{-1} \cdot p(D)]}
\]

\[
= \frac{[p(R)^{-n-1}]^{-n} \prod_k p(R | t_k)}{p(R)^{-(n-1)} \prod_k p(R | t_k) + p(D)^{-(n-1)} \prod_k p(D | t_k)}
\]

\[
= \frac{\prod_k p(R | t_k)}{\prod_k p(R | t_k) + \prod_k p(D | t_k)}.
\]

In the above we can simplify based on having fixed, balanced priors in the train and test sets \( p(R) = p(D) = 0.5 \) and thus \( p(R)^{1-n} \) and \( p(D)^{1-n} \) cancel out.

When working with non-balanced priors, and further when those priors differ between train and test, we have the following:

\[
p_{test}(R | T) = \frac{\prod_k [p(R | t_k)p(t_k)p_{train}(R)^{-1}] \cdot p_{test}(R)}{Z} = \frac{p_{test}(R)p_{train}(R)^{-n} \prod_k [p(R | t_k)p(t_k)]}{Z}
\]

\[
Z = \frac{p_{test}(R)}{p_{train}(R)^n} \prod_k [p(R | t_k)p(t_k)] + \frac{p_{test}(D)}{p_{train}(D)^n} \prod_k [p(D | t_k)p(t_k)].
\]

Below \( \hat{p} \) is written to make clear that the conditional probability of a class given \( \prod_k p(R | t_k) \) and \( \prod_k p(D | t_k) \) products will lead to numeric underflow as \( k \) is large. However, [Volkova et al. (2014)] found that political preference can be often be predicted using roughly 100 tweets, depending on the context of user selection, where this could mean hours, or weeks, based on the user’s tweeting frequency.
CHAPTER 4. STREAMING ONLINE PREDICTION

A tweet is in practice some function learned from a training set, being applied to a test set, and doesn’t represent a “true” conditional probability in the test. ˆp is closer to the “true” test-time conditional probability once multiplied by the ratio of train and test priors.

We can pull out \( p(t_k) \) as before, giving us the following:

\[
p_{\text{test}}(R \mid T) = \frac{p_{\text{test}}(R)}{p_{\text{train}}(R)^n} \prod_k \hat{p}(R \mid t_k) \frac{p_{\text{test}}(D)}{p_{\text{train}}(D)^n} \prod_k \hat{p}(D \mid t_k) = \frac{1}{1 + \left[ \frac{p_{\text{test}}(R)}{p_{\text{train}}(R)^n} \prod_k \hat{p}(R \mid t_k) \right]^{-1} \frac{p_{\text{test}}(D)}{p_{\text{train}}(D)^n} \prod_k \hat{p}(D \mid t_k)} = [1 + \frac{p_{\text{test}}(D)}{p_{\text{test}}(R)} \frac{p_{\text{train}}(R)^n}{p_{\text{train}}(D)^n} \prod_k \hat{p}(R \mid t_k)]^{-1}.
\]

(4.4)

Our goal is to maximize posterior probability estimates given a stream of communications for each user in the test data over (a) time \( \tau \) and (b) the number of tweets \( T \). For that, for each user we take tweets that arrive continuously over time and apply two different streaming models:

- **User Model with Dynamic Updates**: relies exclusively on user tweets \( t_{1}^{(v_i)}, \ldots, t_{k}^{(v_i)} \) following the order they arrive over time \( \tau \), where for each user \( v_i \) we dynamically update the posterior \( p_{\text{test}}(R \mid t_{1}^{(v_i)}, \ldots, t_{k}^{(v_i)}) \).

- **User-Neighbor Model with Dynamic Updates**: relies on both neighbor \( N_r \) communications including friend, follower, retweet, user mention and user tweets \( t_{1}^{(v_i)}, \ldots, t_{k}^{(N_r)} \) following the order they arrive over time \( \tau \); here we dynamically update the posterior probability \( p_{\text{test}}(R \mid t_{1}^{(v_i)}, \ldots, t_{k}^{(N_r)}) \).
4.2 Experiments

We evaluate our models with dynamic Bayesian updates on a continuous stream of communications over time as shown in Figure 4.1. Unlike static model experiments, we are not modeling the influence of the number of neighbors or the amount of content per neighbor. Here, we order user and neighbor communication streams by real world time of posting and measure changes in posterior probabilities over time. The main purpose of these experiments is to quantitatively evaluate (1) the number of tweets and (2) the amount of real world time it takes to observe enough evidence on Twitter to make reliable predictions taking as an example political preference attribute.

We experiment with log-linear models defined in Eq. 3.1 and 3.2 and continually estimate the posterior probabilities for political preferences \( P(R \mid T) \) as defined in Eq. 4.2. We average the posterior probability results over the users in \( G_{cand}, G_{geo} \), and \( G_{active} \) graphs. We train streaming models on an attribute balanced subset of tweets for each user \( v_i \) excluding \( v_i \)'s tweets (or \( v_i \)'s neighbor tweets for a joint model). This setup is similar to leave-one-out classification. The classifier is learned using binary word n-gram features extracted from user or user-neighbor communications. We prefer binary to normalized count-based features to overcome sparsity issues caused by making predictions on each tweet individually.
CHAPTER 4. STREAMING ONLINE PREDICTION

4.3 Results

We present streaming classification results taking as an example political preference attribute only. We experiment with three graphs including $G_{\text{active}}$, $G_{\text{cand}}$ and $G_{\text{geo}}$ to further highlight significant differences in performance that arise from the underlying data collection and annotations strategies.

4.3.1 Modeling Iterative Posterior Updates from a User Stream

Figures 4.2a and 4.2b demonstrate dynamic user model prediction results averaged over users from $G_{\text{cand}}$ and $G_{\text{active}}$ graphs. Each figure outlines changes in sequential average probability estimates $p_\mu(R \mid T)$ for each individual self-authored tweet $t_k$ as defined in Eq. 4.2. The average probability estimates $p_\mu(R \mid T)$ are reported for every 5 tweets in a stream $T = (t_1, \ldots, t_k)$ as $\frac{\sum_{n=1}^{k} P(R \mid t_k)}{n}$, where $n$ is the total number of users with the same attribute $R$ or $D$. We represent $p_\mu(R \mid T)$ as a box and whisker plot with the median, lower and upper quantiles to show the variance; the length of whiskers indicate lower and upper extreme values.

We find similar behavior across all three graphs. In particular, the posterior estimates converge faster when predicting Democratic than Republican users but it has been trained on an equal number of tweets per class. We observe that average pos-

\[4\] Code for inferring user gender, age and political preference attributes using streaming and batch models can be found here: https://bitbucket.org/svolkova/attribute
Figure 4.2: Streaming classification results from user communications for $G_{\text{cand}}$ and $G_{\text{active}}$ graphs averaged over 5 tweets (red - Republican, blue - Democratic).

The posterior estimates $P_{\mu}(R \mid T)$ converge faster to 0 (Democratic) than to 1 (Republican) in Figures 4.2a and 4.2b. It suggests that language of Democrats is more expressive of their political preference than language of Republicans. For example, frequent politically influenced terms used widely by Democratic users include faith4liberty, constitutionally, pass, vote2012, terroristic.

The variance for average posterior estimates decreases when the number of tweets increases for all three datasets. Moreover, we detect that $P_{\mu}(R \mid T)$ estimates for
users in $G_{cand}$ converge 2-3 times faster in terms of number of tweets than for users in $G_{active}$. The lowest convergence is detected for $G_{geo}$ where after $t_k = 250$ tweets the average posterior estimate $P_{\mu}(R \mid t_k) = 0.904 \pm 0.044$ and $P_{\mu}(D \mid t_k) = 0.861 \pm 0.008$.

It means that users in $G_{cand}$ are more politically vocal compared to users in $G_{active}$ and $G_{geo}$. As a result, less active users in $G_{geo}$ just need more than 250 tweets to converge to a true 0 or 1 class. These results are coherent with the outcomes for our static models shown in Figures 3.4 and 3.5. These findings further confirm that differences in performance are caused by various biases present in the data due to distinct sampling and annotation approaches.

Figure 4.3a and 4.3b illustrate the amount of time required for the user model to infer political preferences estimated for 1,031 users in $G_{cand}$, 371 users in $G_{active}$ and 270 users in $G_{geo}$. The amount of time needed can be evaluated for different accuracy levels e.g., 0.75 and 0.95. Thus, with 75% accuracy we classify:

- 100 ($\sim20\%$) $R$ users in 3.6 hours and $D$ users in 2.2 hours for $G_{cand}$;
- 100 ($\sim56\%$) $R$ users in 20 weeks and 100 ($\sim52\%$) $D$ users in 8.9 weeks for $G_{active}$
  which is 800 times longer that for $G_{cand}$;
- 100 ($\sim75\%$) $R$ users in 12 weeks and 80 ($\sim60\%$) $D$ users in 19 weeks for $G_{geo}$.

Such extreme divergences in the amount of time required for classification across all graphs should be of strong interest to researchers concerned with attribute prediction, because Twitter users produce messages with extremely different frequencies. In our case, users in $G_{active}$ tweet approximately 800 times less frequently than in $G_{cand}$.
CHAPTER 4. STREAMING ONLINE PREDICTION

Figure 4.3: Time needed for the dynamic user model (a, b) and the joint-user model (c, d) to infer political preferences of Democratic (blue) and Republican (red) users at 75% (dotted line) and 95% (solid line) accuracy levels.

4.3.2 Modeling Predictions from a User-Neighbor Stream

We estimate dynamic posterior updates from a joint stream of user and neighbor communications in $G_{geo}$, $G_{cand}$ and $G_{active}$ graphs. To make a fair comparison with a streaming user model, we start with the same user tweet $t_0(v_i)$. Then instead of waiting for the next user tweet we rely on any neighbor tweets that appear until the
user produces the next tweet $t_1(v_i)$. We rely on communications from four types of neighbors that demonstrated the highest performance for the baseline static models such as friends, followers, retweets and user mentions.

The convergence rate for the average posterior probability estimates $P_\mu(R \mid T)$ depending on the number of tweets is similar to the user model results presented in Figure 4.2. However, for $G_{geo}$ the variance for $P_\mu(R \mid T)$ is higher for Democratic users; for $G_{active}$ $P_\mu(R \mid T) \rightarrow 1$ for Republicans in less than 110 tweets which is $\Delta t = 40$ tweets faster than the user model; for $G_{cand}$ the convergence for both $P_\mu(R \mid T) \rightarrow 1$ and $P_\mu(D \mid T) \rightarrow 0$ is not significantly different than the user model.

Figures 4.3c and 4.3d show the amount of time required for a joint user-neighbor model to infer political preferences estimated for users in $G_{cand}$ and $G_{active}$. We find that with 75% accuracy we can classify 100 users for:

- $G_{cand}$: Republican users in 23 minutes and Democratic users in 10 minutes;
- $G_{active}$: $R$ users in 3.2 weeks and $D$ users in 1.1 weeks which is 7 times faster.
CHAPTER 4. STREAMING ONLINE PREDICTION

on average across attributes than for the user model;

- $G_{geo}$: $R$ users in 1.2 weeks and $D$ users in 3.5 weeks which is on average 6 times faster across attributes than for the user model.

Similar or better $P_{\mu}(R \mid T)$ convergence in terms of the number of tweets and, especially, in the amount of time needed for user and user-neighbor models further confirms that neighborhood content is useful for latent user attribute inference taking as an example political preference prediction. Moreover, communications from a joint stream allow to make an inference up to 7 times faster.

To summarize, in Figure 4.4 we recap the political preference classification speed and present a summary on how much time will it take to predict political preferences for 100 random users with 75% accuracy (50 Democratic (blue) and 50 Republican (red) users). We randomly sample users from three datasets: candidate-centric, geo-centric and active and compare the prediction speed obtained using our user and user-neighbor streaming models among them.

4.4 Batch vs. Streaming Model Comparison

In Figure 4.5 we summarize our streaming model results for political preference prediction on three datasets: candidate-centric, geo-centric and active and compare them to the batch model results from Chapter 3. We find that our streaming models
CHAPTER 4. STREAMING ONLINE PREDICTION

significantly outperform static batch models by as much as +0.28 for the candidate-centric graph, 0.31 for the geo-centric graph and +0.25 for the active graph (when the predictions are made from a joint user-neighbor stream).

![Multi-dataset comparison of batch vs. online models to infer user political preferences.](chart)

**Figure 4.5:** Multi-dataset comparison of batch vs. online models to infer user political preferences.

4.5 Conclusions

In Chapter 4 we introduced models for iterative latent attribute prediction designed to model dynamic (streaming) nature of social media. We first defined models that use iterative Bayesian rule updates to dynamically update beliefs about user attributes from a stream of user or joint user-neighbor communications. We run a set of experiments to determine (a) the number of tweets and (b) time e.g., hours, days, weeks needed to infer political preferences for a given user with certain confidence.
level. We also discussed the problem of balanced e.g., gender vs. unbalanced e.g., political preference priors for the attributes on Twitter. We tested our models on three datasets, similar to the experimental setup in Chapter 3, to ensure model generalization for different datasets that contain various sampling and annotation biases. We compare the results obtained using our streaming models applied over (a) user and (b) user-neighbor streams of tweets with the results from our static (a) user and (b) neighbor models from Chapter 3.

Our key findings on streaming dynamic models for personal analytics in social media are presented below.

- **Generalization of the classifiers.** As in Chapter 3, we raise a very important but under-explored problem of the generalization of classifiers for personal analytics in social media, also recently discussed by Cohen and Ruths (2013) and Volkova et al. (2014). For instance, the existing models developed for political preference prediction are all trained on Twitter data but report significantly different results even for the same baseline models trained using bag-of-word lexical features as shown in Table 2.4. As before, we experiment with three different datasets. Our results for both static and dynamic models show that the accuracy indeed depends on the way the data was constructed. Therefore, publicly available datasets need to be released for a meaningful comparison of the approaches for personal analytics in social media.

- **Streaming models are more effective than batch models for personal
CHAPTER 4. STREAMING ONLINE PREDICTION

analytics. The predictions made using dynamic models with Bayesian updates over user and joint user-neighbor communication streams demonstrate higher performance with lower resources spent compared to the batch models. Depending on user involvement in online communications e.g., retweeting, replying, actively following other users of the same kind, expressiveness e.g., sharing more or less personal information and activeness e.g., frequency of tweeting, the perfect predictions can be made using as low as 100 - 500 tweets per user.

- More evidence is available per user more confident attribute prediction are. Regardless the type of the model – static vs. streaming, the more tweets are available per user the better classification results are (as expected). Moreover, when we estimate all models on the same amount of content per user, we found that streaming models with iterative Bayesian updates outperform static ngram models.

In Chapter 5 we apply our tweet-based streaming models that rely on Bayesian rule updates in several settings that allow both learning and inference over time. Unlike the existing batch models that degrade over time due to the extreme content shift in social media e.g., Euromaidan was frequently discussed in 2014 but rarely mentioned in 2015, we propose to actively learn models over time or iteratively retrain the existing models on newly available data to handle the data drift in social media and obtain more accurate and generalizable models over time.
Chapter 5

Iterative Learning and Prediction

With individual authors in social media as training and test instances, their associated content ("features") are made available incrementally over time, as they converse over online discussions. In this Chapter\footnote{This chapter presents “Online Bayesian Models for Personal Analytics in Social Media”, which was published in the Proceedings of the 29th Conference on Artificial Intelligence (AAAI) in 2015 and is a joint work with Ben Van Durme.} we propose various approaches to handling this dynamic data, from traditional batch training and testing, to incremental bootstrapping, and then active learning via crowdsourcing. Our underlying model adapted from Chapter\footnote{This chapter presents “Online Bayesian Models for Personal Analytics in Social Media”, which was published in the Proceedings of the 29th Conference on Artificial Intelligence (AAAI) in 2015 and is a joint work with Ben Van Durme.} relies on an intuitive application of Bayes rule, which should be easy to adopt by the community, thus allowing for a general shift towards online modeling for social media.

As we discussed earlier, the majority of work on social media predictive analytics treats the modeling task much as prior work on non-social media: construct a corpus of labeled materials, and perform supervised classification in a batch setting. This
Figure 5.1: An example of a political preference learning and prediction over a dynamic stream of communications. \( R \) stands for Republican and \( D \) for Democratic users. As time \( \tau \) goes by, both labeled and unlabeled users generate tweets \( t_1 \ldots t_m \). Boxes outline the amount of train and test data available at each timestamp \( \tau_k \).

ignores one of the primary distinguishing characteristics of social media content: it is (I) dynamically generated over time, and (II) usually centered within the context of a social network (i.e., friends or other types of associates of the author). Further, different users of the medium contribute to greater or lesser extent: a given user may send one tweet a week, or one tweet an hour, etc. Prior work tends to gloss over this fact by building controlled collections with a large, fixed amount of content assumed per user e.g., 1K tweets. (Rao et al., 2010; Zamal et al., 2012; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013; Bamman et al., 2014).

In contrast, we showed the intuitive importance of the amount of content available per user at test time in Chapters 3 and 4: the more content you have, the better your
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

predictions. This was motivated by Van Durme (2012b) who proposed a model that allowed for incremental updating of classifier predictions over time, as users continued to author new content. This model treated each user as a sort of dynamic feature vector that evolved over time, and assumed access to a pre-trained classification model based on labeled data available a priori, akin to earlier work in the purely batch setting.

Here we go beyond the existing work and propose two novel contributions in mining streaming social media:

- contrasting Van Durme (2012b), we treat each new message as independent evidence which is combined into an incremental user-prediction model as a straightforward application of Bayes Rule as discussed in Chapter 4;
- we explore model training in parallel to its application, rather than assuming a previously existing labeled dataset.

Also, distinct from Van Durme, but previously explored in the batch-setting by Zamal et al. (2012) and Volkova et al. (2014) we make use of the local user neighborhood (applying our findings from Chapter 3) and annotator rationales e.g., highly predictive features (further explored in Chapter 6) in our dynamic model.

Our approach captures the same incremental intuitions as the work by Van Durme, but we situate it within the well understood framework of Bayesian inference as described in Chapter 4. We hope this will encourage others to build upon this effort.
in constructing more complicated models. Further, by recognizing that both training as well as testing materials are dynamically generated in social media, then possibly coupled to dynamic model feedback via crowdsourcing, this suggests latent author attribute prediction as a rich source for methodological challenges in online and active learning. Here we mean to give perspective on the various ways this dynamism may be incorporated into an experimental framework.

5.1 Methodology

5.1.1 Data

Our approach is relevant generally to multi-class prediction problems in social media. Here we focus on a binary prediction task, specifically the prediction of political preference as captured by the dominant two American political parties: Democratic, and Republican. We rely on a dataset previously used for political affiliation classification by (Pennacchiotti & Popescu, 2011b), (Zamal et al., 2012) and described in details in Chapter 3 as Active graph. The original Twitter users with their political labels extracted from http://www.wefollow.com as described by (Pennacchiotti & Popescu, 2011b). The user-friend data was collected by (Zamal et al., 2012). The original data consists of 200 Republican and 200 Democratic users associated with 925 tweets on average per user. Each user has on average 6155 friends with 642 tweets
per friend. Sharing restrictions\footnote{Twitter only allows to share user and tweet IDs. The actual content e.g., tweets or user meta data can be download by querying Twitter API. However, as of Aug. 2013, a certain portion of user profiles were deleted or became private: this is a standard issue in reproducing prior results based on Twitter and is not specific to this work.} and rate limits on Twitter data collection only allowed us to recreate a subset of that collection. Based on the subset we were able to obtain we formed a balanced collection of 150 Democratic and 150 Republican users. For each user, we randomly sampled 20 friends with 200 tweets per friend.

### 5.1.2 Models

We assume a set of independent users $U = \{u_i\}$, and neighbors $N = \{n_j\}$, with $N^{(u)}$ the neighbors of $u$. In our experiments in this Chapter those neighbors will be the friends of a user on Twitter (as defined in section 3.1.3). We are concerned with models over data that changes over time: let $\tau$ be an index over discrete time-steps, where at each time-step $\tau_k$ we observe zero or more tweets from each user, and each user-neighbor, on which we base our predictions. A user is labeled at time $\tau$ if we know the value of the attribute function $A(u) \in \{a_i\}$.

For example, in our experiments we will model the (American) political preference attribute, defined as: $A(u) \in \{R, D\}$, with $R$ standing for Republican and $D$ for Democratic. Let $L_\tau \subseteq U$ be the labeled users at time $\tau$, and $\overline{L}_\tau = U \setminus L_\tau$ the unlabeled users. Our goal is to predict the attribute value for each user in $\overline{L}_\tau$ at every $\tau$ given the evidence available up to $\tau$.

Unlike previous models for latent user attribute classification, we:
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

1. consider updating the initial model learned at $\tau_0$ as new evidence becomes available at $\tau_k$;

2. reestimate decision probabilities for the unlabeled users given the updated model and new content generated by these users and their neighbors by $\tau_k$.

We define two models $\Phi(u, \tau)$ and $\Phi(n, \tau)$ learned from dynamically growing streams of tweets $T^{(U)}$ and $T^{(N)}$. The user model $\Phi(u, \tau)$ is learned exclusively from user communications to be applied to user tweets $t^{(u)}_1, t^{(u)}_2, \ldots, t^{(u)}_m \in T^{(u)}_{\tau}$. $\Phi(u, \tau)$ is then a function mapping a user to the most likely attribute value assignment at $\tau$:

$$\Phi(u, \tau) = \arg\max_a P(A(u) = a \mid T^{(u)}_{\tau}). \quad (5.1)$$

Neighbor model $\Phi(n, \tau)$ is learned from neighbor communications of Democratic and Republican users. It is defined similarly to Eq. (5.1) and is applied to classify friend tweets within friend or joint user-friend stream $t^{(n)}_1, t^{(n)}_2, \ldots, t^{(n)}_m \in T^{(n)}_{\tau}$.

A user is labeled at time $\tau$ if we predict the value of the attribute function $A(u)$. We apply Bayesian rule updates to dynamically revise posterior probability estimates of the attribute value $P(A(u) = R \mid T_{\tau})$ given a prior e.g., in our case we start with a balanced prior $P(R) = P(D) = 0.5$.

$$P(A(u) = R \mid T_{\tau}) = \frac{P(A(u) = R) \cdot P(T_{\tau} \mid A(u) = R)}{\sum_{a \in A} P(A(u) = a) \cdot P(T_{\tau} \mid A(u) = a)}, \quad (5.2)$$
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

We will assume tweets to be independent conditioned on attribute, which means our model factors across individual messages $T_\tau = (t_1, \ldots, t_m)$, allowing for simple posterior updates on a tweet by tweet basis:

$$P(A(u) = R \mid t_1 \ldots t_m) = \frac{P(A(u) = R) \cdot \prod_m P(t_m \mid A(u) = R)}{\sum_{a \in A} P(A(u) = a) \prod_m P(t_m \mid A(u) = a)}.$$  (5.3)

The conditional probability of a given tweet is determined by a log-linear model trained on observations from $L_\tau$. We show the example updated posterior probabilities for political preference prediction $P(R \mid t_1 \ldots t_m)$ in Figure 5.2.

Figure 5.2: Active learning classification setup. Nodes represent Twitter users, edges (connecting lines) stand for friend relationships between the users; dark red and blue nodes represent labeled $R$ and $D$ users; light red and blue nodes represent friends of Republican $R$ and Democratic $D$ users; $\Phi_U(\tau_1)$ and $\Phi_N(\tau_1)$ are the models trained exclusively on user or neighbor (friend) content.
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

The final decisions about label assignments can be made at any time $\tau_k$ e.g., if $P(R \mid t_1 \ldots t_m) = 0.9$ one can label the user as $R$ with an associated 90% model confidence given the evidence available by $\tau_k$. We analyze the difference in precision and recall by making decisions based on high or low probability assignments using different thresholds $\theta$: 0.55 and 0.95. When $P(A(u) = a \mid t_1 \ldots t_m)$ exceeds $\theta$ we make a final decision about the label for a user at time $\tau_k$.

5.2 Experimental Setup

We design a set of classification experiments from three types of data streams including user (U), neighbor (N) and user-neighbor (UN). We aim to explore the following prediction settings:

a. Iterative Batch (IB),
b. Iterative Batch with Rationale Filter (IBR),
c. Active without Oracle (AWOO),
d. Active with Oracle (AWO),
e. Active with Rationale Filter (AWR).

For all settings we perform 6-fold cross validation and use a balanced prior: 50 users in the train split and 250 users in the test. For all experiments we use the LibLinear package integrated in the Jerboa toolkit (Van Durme, 2012a). The framework generalizes to non-balanced priors, but does assume that the prior is known a priori; estimating class priors in social media has been briefly explored by Beller et al. (2014), and is an element of future work.
log-linear models with dynamic Bayesian updates defined in Eq.5.1 and Eq.5.3 are learned using binary word unigram features extracted from user or neighbor content (as they have demonstrated better performance compared to higher order n-grams).

5.2.1 Iterative Batch

We learn tweet-based models at each time stamp $\tau$ from the set of labeled users $L_\tau$ and their neighbors e.g., friends. We apply these models using Eq.5.3 to U, N and UN streams to label all unlabeled users $\overline{L}_\tau$ over time. The set of labeled users is constant across all values of $\tau$: we have labels on some users before hand, and no new labels are gathered; only the amount of content available for the users and their neighbors is increasing over time.

Figure 5.3: Batch setting (a) without and (b) with rationale filtering.
5.2.2 Iterative Batch With Rationale Filtering

Prior work by (O. Zaidan & Eisner, 2008) and (Yessenalina, Choi, & Cardie, 2010) explored the utility of asking annotators to provide *rationales*, explicitly highlighted words or phrases in provided content, that best justified why the annotator made their labeling decision.

For example, some Democratic and Republican rationales are shown below:

- D: *Hundreds die each year mining coal worldwide. When will we move to safer-for-everyone energy technologies?* Wow. Uganda is passing a *bill to execute all gays*.

- R: *From @washingtonpost: many of the doctor-legislators in the House oppose job-crushing healthcare law.*

Our batch setup with rationale filtering shown in Figure 5.3b is equivalent to the iterative batch setup, except at every time stamp \( \tau \) we modify our training data to include tweets with the rationales exclusively. For that, at every \( \tau \) we estimate predictive unigrams – potential rationale words \( w \in V \) for Democratic and Republican users in \( L_\tau \):

\[
V^{a \in A(u)} = \{ w \mid P(w \mid A(u) = a) \geq 0.55 \}
\] (5.4)

The conditional probabilities of each word \( P(w \mid A(u) = D) \) and \( P(w \mid A(u) = R) \) are calculated as the empirical estimates over tweets, where \( w \) was constrained to have a minimum count of three. We then ask annotators on Mechanical Turk to select...
rationales from the strongest ranked candidates for $D$ and $R$ by showing them potential rationales and a subset of tweets up to $\tau$. We ask three redundant annotations for each unigram $w$ and take the majority vote to determine if the unigram truly reveals political preferences. For example, the Democratic rationales with $P(w \mid A(u) = a) > 0.9$ and 100% annotator agreement include: *immigrants*, *students*, *unemployment* and Republican: *#teaparty*, *obamacare*.

### 5.2.3 Active Without Oracle

Unlike our batch setup applied iteratively over a stream of tweets, we propose to update the $\Phi(u, \tau)$ and $\Phi(n, \tau)$ models by moving users from the test set labeled at $\tau_k$ to the training set at $\tau_{k+1}$ as shown in Figure 5.2. The final decisions about class labels for the unlabeled users are made based on posterior probability estimates $P(A(u) \mid T, \tau)$ to exceed the threshold $\theta$. Similarly to the batch setting we experiment with two values of $\theta$ and three data streams: U, N and UN. This bootstrapping approach (aka self-training [Yarowsky 1995]) we refer to as active without oracle (AWOO).

### 5.2.4 Active With Oracle

Alternatively, the final label assignments can be judged by an oracle e.g., annotators on Mechanical Turk. For example, we might show $m$ tweets produced by the user
by time \( \tau \) to one or more annotators. And only if one or more independent annotator judgments agree with \( \Phi(u, \tau) \), then we assign a corresponding label to this user at \( \tau_k \), and move this user to the training set at \( \tau_{k+1} \). Here, since we know the labels we simulate turker judgments (so the oracle is 100% correct). Thus, this setup measures the upper bound for classification. But in the future, we would like to engage real turkers to make class label judgments in the loop. We refer to this setup as active with oracle (AWO) and show it in Figure 5.4b.

Figure 5.4: Active setting (a) with oracle, (b) without oracle and (c) with rationale filtering. Active setting without oracle (AWOO) is similar to (c) AWR except that the rationale filtering step is omitted.

**5.2.5 Active With Rationale Filtering**

The rationale filtering step used for IBR setup is also applied to AWOO setup at every \( \tau \) as shown in Figure 5.4a. The difference between batch and active models with rationale filtering is that the potential rationales are estimated on a different set of training data using Eq. 5.4. In the active case tweets from previously unlabeled users that exceed \( \theta \) at \( \tau_k \) are added to the tweets of labeled users at \( \tau_{k+1} \).
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

Data: Labeled $L_\tau$ and unlabeled $\overline{L}_\tau$ users and their neighbors $N$; user $T^{(u)}$ and neighbor $T^{(n)}$ tweets; time $\tau$; confidence threshold $\theta$

Result: Attribute labels $A(u)$ for unlabeled users in $\overline{L}_\tau$

foreach time stamp $\tau_k \in \tau$ do

foreach labeled user $u \in L_{\tau_k}$ and their immediate neighbors $N^{(u)}$ do

$T^{(u)}, T^{(n)} \leftarrow$ collect tweets up to $\tau_k$

if setup is “IBR” or “AWR” then

filter rationale tweets from $T^{(u)}, T^{(n)}$

end

$\vec{f}(T^{(u)}), \vec{f}(T^{(n)}) \leftarrow$ build user and neighbor feature vectors from $T^{(u)}, T^{(n)}$

end

train user $\Phi(u, \tau_k)$ and neighbor $\Phi(n, \tau_k)$ models

foreach unlabeled user $u \in \overline{L}_{\tau_k}$ do

$s \leftarrow$ user (U) or neighbor (N) or joint user-neighbor (UN) stream

foreach tweet $t^{(s)}_m \in T^{(s)}$ do

update $P(A(u) \mid t_m)$ using Eq. 5.3

if $P(A(u) \mid t_m) \geq \theta$ then

$A(u) \leftarrow a$

if setup is AWO then

if $A(u) = A_{oracle}(u)$ then

$L_{\tau_k+1} \leftarrow L_{\tau_k} + u$

$\overline{L}_{\tau_k+1} \leftarrow \overline{L}_{\tau_k} - u$

end

end

else if setup is AWO or AWR then

$L_{\tau_k+1} \leftarrow L_{\tau_k} + u$

$\overline{L}_{\tau_k+1} \leftarrow \overline{L}_{\tau_k} - u$

end

else if setup is IB or IBR then

$L_{\tau_k+1} \leftarrow L_{\tau_k} + \overline{L}_{\tau_k+1}$

$\overline{L}_{\tau_k+1} \leftarrow \overline{L}_{\tau_k}$

end

end

end

end

Algorithm 1: ONLINEINFER ($L_\tau, \overline{L}_\tau, T_\tau, \theta$)
5.2.6 Evaluation

We are concerned with accuracy when operating at different confidence thresholds.

Let \( \text{Acc}_{\tau,\theta} \) be the prediction accuracy at \( \tau \), when considering just users for which the posterior probability exceeds \( \theta \). At a given value of \( \tau \) and \( \theta \), let:

\[
\text{Acc}_{\tau,\theta} = \frac{TR + TD}{R + D},
\]

where \( TR = \) true Republicans, \( TD = \) true Democrats, and \( R, D \) are the number of users labeled as Republicans or Democrats, respectively. This generalizes standard language of (True) Positive and (True) Negative to allow for non-binary scenarios, such as if adding “Libertarian” (\( L \)), “Green Party” (\( G \)), etc., to the attribute set:

\[
\text{Acc} = \frac{TR + TD + TG + TL}{R + D + G + L}.
\]

We abbreviate this as: \( \text{Acc}_{\tau,\theta} = C_{\tau,\theta}/A_{\tau,\theta} \), with \( C_{\tau,\theta} \) being the number of correctly classified users, and \( A_{\tau,\theta} \) being the number of users above a given threshold \( \theta \). We also estimate \( Q_{\tau,\theta} \) which is the total number of active users who tweeted at least once by \( \tau \) (note that \( C_{\tau,\theta} \leq A_{\tau,\theta} \leq Q_{\tau,\theta} \)). The performance metric \( \text{Acc}_{\tau,\theta} \) defined in Eq. 5.5 can be effectively used for targeted online advertising where one would like to send the advertisements as early as possible to only active users at time \( \tau \) for whom labels are assigned with a reasonable confidence \( \theta \) as discussed in details in Section 5.4.
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

5.3 Experimental Results

5.3.1 Iterative Batch Results

We first confirm that our incrementally batch-trained approach performs as would be expected. In Figure 5.5 (a - b), consider model U (based only on user tweets): the difference between decision thresholds 0.95 and 0.55 shows a classic precision versus recall tradeoff; at 0.95 less users are classified (x-axis) but at higher precision (y-axis), as compared to 0.55 which instead has higher recall. This pattern repeats for all models U, N and UN, trained and tested with less data (a: Jan - Apr) as well as more data (b: Jan - Sep). With more data (b), performance improves for all scenarios. U is outperformed by N and UN: having access to the content of neighbors improves performance considerably in all cases (affirming the conclusions of Zamal et al. (2012) and Volkova et al. (2014)).

Next we contrast those results to AWOO: not only do we retrain the model each month as in batch, but now we bootstrap by taking our most confident (0.95 or 0.55) predictions for users and add them into our labeled set as if their labels were known. We found that our AWOO model yields higher performance than IB model in early months (up to 1-Jul-2011), and insignificantly lower results after that. It happens because in the active setting the model accumulates noisy predictions for some users over time. In contrast, the AWO model does not have this issue and yields consistently better results over time as we show latter.
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

(a) IB, $\tau_3=1$-Apr-2011
Train: 3K user, 6K neigh

(b) IB, $\tau_8=1$-Sep-2011
Train: 11K user, 20K neigh

(c) AWOO, $\tau_3=1$-Apr-2011
Train: 18K user, 34K neigh

(d) AWOO, $\tau_8=1$-Sep-2011
Train: 61K user, 125K neigh

Figure 5.5: Comparing the performance of batch (IB) vs. active (AWOO) settings using user U, neighbor N and user-neighbor UN data streams and different confidence thresholds: $\theta = 0.55$ (bold) and $\theta = 0.95$ (light) markers.

In Figure 5.6 we present more detailed classification results for batch and active setting for two thresholds 0.55 and 0.95. These results allow us to analyze the threshold and data stream type influence on classification performance.
Figure 5.6: Comparing iterative batch (a, b) vs. active (c, d) models with (b, d) and without (a, c) rationale filtering for political preference with $\theta = 0.55$ and $\theta = 0.95$ applied to user, neighbor and user-neighbor streams. Starting on 1-Jan-2011, at each time stamp $\tau$ (e.g., $\tau_2 = 1$-Mar-2011, ..., $\tau_{11} = 1$-Dec-2011) we measure $C_{\tau, \theta}$ = the number of correctly classified users, $A_{\tau, \theta}$ = the number of users above the threshold $\theta$, $Q_{\tau}$ = the number of active users who tweeted at least once by time $\tau$. 
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

5.3.2 Analyzing Threshold Influence

The results in Figures 5.5 and 5.6 demonstrate that for higher $\theta$, when the models are more constrained and, therefore, more confident about their predictions, less users $A$ are above the threshold $\theta$. Consequently, the number of correctly classified users $C$ is lower for 0.95 vs. 0.55. Therefore, one has to make a decision about $\theta$ taking into account this precision-recall tradeoff: models with higher $\theta$ are more precise but yield lower recall vs. models with lower $\theta$ are less precise but yield higher recall.

Moreover, for our active setting threshold $\theta$ has another important objective – to control the amount and quality of the data labeled at $\tau_k$ and used to update the model at $\tau_{k+1}$. The results in Figure 5.6 show that the active models outperform the iterative batch models in terms of recall in early months. This results are very important for an example use-case of targeted advertising task (discussed in details in section 5.4) when more ads need be sent to more users as early as possible. Moreover, $C_{\tau,\theta}/A_{\tau,\theta}$ ratio is higher for AWOO vs. IB setting when U stream is used.

5.3.3 Studying Data Stream Type Influence

We observe that in all settings when the probability estimates are updated from N, UN streams compared to U stream the number of correctly classified users $C_{\tau,\theta}$ at each $\tau$ is significantly higher. The reason for UN, N streams yielding better results is that more tweets associated with the user e.g., friend tweets that carry a substantial signal
for prediction become available. Many authors don’t tweet very often e.g., 85.3% of all Twitter users post less than one update per day\footnote{As reported at \url{http://www.sysomos.com/insidetwitter}}. Thus, less tweets are generated by random users by time $\tau$ compared to the number of tweets generated by a set of their friends. However, relative gains over time for N and UN are lower compared to U stream. It is because “less difficult to classify” users are easily classified using UN (N) streams earlier at $\tau_k$ and only “more difficult to classify” users are left to be classified later at $\tau_{k+1}$.

### 5.3.4 Comparing Performance for Batch vs Active Learning Setting

In Figure 5.6 we reported classification performance in terms of the number of correctly classified users over time and model quality calculated using the Eq. 5.5. Below we summarize the results for both batch and active setting and outline concrete differences between them as shown in Figures 5.7 and 5.8. We find that models with rational filtering (IBR, AWR) yield higher precision but the models without rationale filtering (IB, AWOO) yield higher recall. We also observe that for batch models the number of correctly classified users increases over time – for IB faster, but for IBR slower. Unlike IB and IBR models, AWOO and AWR models classify more users correctly faster (in Mar) but then plateaus.
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

Figure 5.7: Demonstrating the trade-off between model quality and classification accuracy for the batch models with (IBR: right) and without (IB: left) rationale filtering.

Figure 5.8: Demonstrating the trade-off between model quality and classification accuracy for the active learning model with (AWR: right) and without (AWOO: left) rationale filtering.

Finally, in Figure 5.9 we summarize how the data stream type (user vs. user-neighbor) influences the model quality in batch vs. active learning setting over time.
5.3.5 Evaluation in Constrained-Resource Conditions

We also study the influence of data stream types and thresholds in daily classification setup when the amount of training and test data is limited. For that we run daily experiments for the first 14 days of each month over 12 months using Ustream in our iterative batch setting. In Figure 5.10 we present the results in terms of accuracy and the number of correctly classified users daily as box and whisker plots.
Figure 5.10: Daily classification results in a batch setting. $\text{Acc}_{\tau,\theta}$ (top) and $C_{\tau,\theta}$ (bottom) for each day in the first 2 weeks of each month in a year using U stream and different $\theta$.

Our key observations on prediction quality and speed in low resource conditions:

(a) The differences in $\text{Acc}_{\tau}$ and $C_{\tau}$ between two thresholds are not statistically significant on a daily basis which suggests that both 0.55 and 0.95 models trained on limited data are equally uncertain.

(b) $C_{\tau,\theta}$ is increasing and the variance for $\text{Acc}_{\tau}$ is decreasing daily suggesting that quality for both IB models is improving over time.

(c) UN stream yields significantly higher $C_{\tau}$ and $C_{\tau}/A_{\tau}$ results compared to U stream even when limited content is available.
5.3.6 Classification with Oracle Annotations

In Figure 5.11 we demonstrate the upper bound for political preference classification performance with $\theta = 0.95$ using our active with oracle (AWO) experimental setup. Similar to other experiments, we report classification performance $\text{Acc}_{\tau,\theta}$ at every $\tau$ with the number of user and neighbor tweets available for training when predictions are made over $U$ and $N$ data streams. We find that $\text{Acc}_{\tau,\theta}$ is monotonically increasing over time and is significantly higher then for IB and AWOO settings. To give a cost estimate of requesting iterative oracle annotations, we outline the number of requests to the oracle aggregated over time in Figure 5.11 (top).

Active learning with iterative oracle annotations demonstrate the highest performance compared to all other classification settings. For instance, 226 out of 250 users (90%) are correctly classified by June using $N$ stream and 230 (92%) using $UN$ stream for AWO setup compared to 191 (76%) and 203 (81%) users using AWOO setup. Similarly, 112 (45%) users are correctly classified by June using $U$ stream using AWO model compared to 80 (32%) using AWOO setting.

5.3.7 Applying Rationale Filtering

Here we analyze the impact of rationale filtering on prediction performance in batch: IB vs. IBR and active: AWOO vs. AWR settings over time. In Figure 5.6 we report results for models with and without rationale filtering. As before, we present
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

Figure 5.11: Classification results for the active learning setting with oracle (AWO) annotations using U, N and UN streams and $\theta = 0.95$ (Numbers in the rows labeled U and N are the number of tweets, in thousands, used to train user and neighbor models).

the results for two thresholds 0.55 and 0.95 and three data streams: U, N and UN.

For IBR and AWR models with rationale filtering we observe similar precision-recall trends to IB and AWOO models shown in Figure 5.5.

During rationale filtering we select training examples with highly predictive norms (a.k.a. rationales) at every $\tau$. This filtering step reduces the number of training examples $L$, vocabulary size $V$ and feature space for both user $\Phi(u, \tau)$ and neighbor $\Phi(n, \tau)$ models over time as shown in Tables 5.1 and 5.2. We observe that the size
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

<table>
<thead>
<tr>
<th></th>
<th>Mar</th>
<th>Jun</th>
<th>Sep</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta L_{\theta}$</td>
<td>1.1</td>
<td>2.4</td>
<td>4.2</td>
<td>9.0</td>
</tr>
<tr>
<td>$\Delta V_{\theta}$</td>
<td>0.8</td>
<td>1.3</td>
<td>1.9</td>
<td>3.0</td>
</tr>
<tr>
<td>$\Delta Acc_{0.55}$</td>
<td>+21.7</td>
<td>+20.9</td>
<td>+14.2</td>
<td>+21.6</td>
</tr>
<tr>
<td>$\Delta Acc_{0.95}$</td>
<td>+27.7</td>
<td>+18.5</td>
<td>+7.8</td>
<td>+15.1</td>
</tr>
<tr>
<td>$\Delta L_{\theta}$</td>
<td>2.5</td>
<td>6.9</td>
<td>12.5</td>
<td>24.9</td>
</tr>
<tr>
<td>$\Delta V_{\theta}$</td>
<td>1.7</td>
<td>3.4</td>
<td>4.8</td>
<td>7.0</td>
</tr>
<tr>
<td>$\Delta Acc_{0.55}$</td>
<td>+5.6</td>
<td>+10.6</td>
<td>+5.8</td>
<td>+3.0</td>
</tr>
<tr>
<td>$\Delta Acc_{0.95}$</td>
<td>+13.2</td>
<td>+11.7</td>
<td>+5.6</td>
<td>+3.0</td>
</tr>
<tr>
<td>$\Delta L_{\theta}$</td>
<td>1.7</td>
<td>3.4</td>
<td>4.8</td>
<td>7.0</td>
</tr>
<tr>
<td>$\Delta V_{\theta}$</td>
<td>1.7</td>
<td>3.4</td>
<td>4.8</td>
<td>7.0</td>
</tr>
<tr>
<td>$\Delta Acc_{0.55}$</td>
<td>+19.2</td>
<td>+9.9</td>
<td>+4.4</td>
<td>+10.3</td>
</tr>
<tr>
<td>$\Delta Acc_{0.95}$</td>
<td>+21.5</td>
<td>+9.1</td>
<td>+4.0</td>
<td>+10.8</td>
</tr>
</tbody>
</table>

Table 5.1: Political preference classification accuracy for iterative batch models with and without rationale filtering.

of the training data $L$ is reduced at least in half at every time stamp. Therefore, we consider rationale filtering as a dimensionality reduction step for our batch and active models with incremental Bayesian updates. More precisely, we show the difference in terms of $D$ and $R$ tweets before and after rationale filtering in Figure 5.12.

Nevertheless, the size of the training data is significantly lower at every $\tau$ the quality of batch and active models trained with filtered data is better for IBR vs. IB and AWR vs. AWOO. In other words, selecting tweets with highly predictive feature norms for training leads to consistent performance improvements over time. We show the empirical results for the relative percentage gain $\Delta Acc,\%$ for batch and active models with vs. without rationale filtering in Tables 5.1 and 5.2 respectively. Models with rationale filtering yield higher precision but lower recall compared to the models without rationale filtering when N or UN streams are used. Except we observe higher precision but comparable or higher recall when U stream is used.
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

<table>
<thead>
<tr>
<th></th>
<th>Mar</th>
<th>Jun</th>
<th>Sep</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>∆Acc₀.₅₅</td>
<td>+20.6</td>
<td>+19.7</td>
<td>+18.5</td>
<td>+20.1</td>
</tr>
<tr>
<td>∆Acc₀.₉₅</td>
<td>+19.0</td>
<td>+18.7</td>
<td>+18.0</td>
<td>+20.6</td>
</tr>
<tr>
<td>∆L₀.₉₅</td>
<td>3.5</td>
<td>11.9</td>
<td>16.3</td>
<td>32.4</td>
</tr>
<tr>
<td>∆V₀.₉₅</td>
<td>2.5</td>
<td>3.8</td>
<td>6.3</td>
<td>11.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mar</th>
<th>Jun</th>
<th>Sep</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>∆Acc₀.₅₅</td>
<td>+14.1</td>
<td>+13.9</td>
<td>+13.7</td>
<td>+13.7</td>
</tr>
<tr>
<td>∆Acc₀.₉₅</td>
<td>+11.7</td>
<td>+11.7</td>
<td>+11.9</td>
<td>+11.9</td>
</tr>
<tr>
<td>∆L₀.₉₅</td>
<td>7.9</td>
<td>19.1</td>
<td>36.5</td>
<td>81.4</td>
</tr>
<tr>
<td>∆V₀.₉₅</td>
<td>2.7</td>
<td>4.6</td>
<td>7.3</td>
<td>13.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mar</th>
<th>Jun</th>
<th>Sep</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>∆Acc₀.₅₅</td>
<td>+11.5</td>
<td>+10.8</td>
<td>+10.7</td>
<td>+10.7</td>
</tr>
<tr>
<td>∆Acc₀.₉₅</td>
<td>+9.5</td>
<td>+9.4</td>
<td>+9.4</td>
<td>+9.4</td>
</tr>
</tbody>
</table>

Table 5.2: Classification accuracy for active learning models with (AWR) and without (AWOO) rationale filtering. ∆Accθ represents the relative percentage gain between AWR and AWOO, ∆Lθ is the difference in the number of tweets available for training, ∆Vθ is the difference in feature space (vocabulary) size at τk.

To summarize, rationale filtering significantly improves classification accuracy and can be effectively used for attribute prediction that require high precision e.g., product likes or personal interests. For batch setting, IBR setup yields much better results than IB setup as high as AccMar,₀.₉₅ = 27.7%. For active setting, AWR setup yields as high as AccMar,₀.₅₅ = 20.6% gain over AWOO using U stream. Moreover, for both batch and active setting: the higher ∆Accτ,θ reported when predictions are made from U compared to N or UN streams; the incremental relative gains for Accτ,θ are higher for 0.55 compared to 0.95 models.
Figure 5.12: Comparing the number of training examples used (tweets in thousands) for active vs. batch models over time in the user model ($\Phi(u, \tau)$, left) and the neighbor model ($\Phi(n, \tau)$, right); red – Republican tweets, blue – Democratic tweets.

### 5.4 Use-cases and Recommendations for Online Prediction

Let’s consider an example task of targeted advertising – when the goal is to send ads to the targeted population e.g., Democratic vs. Republican users, and discuss how our models can be effectively applied for this task.
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

**Use-case I:** If one needs to deliver ads fast but is ok to be less confident in user attribute predictions:

- use models with higher recall (AWOO, IB);
- apply lower decision threshold e.g., 0.55;
- for highly assortative attributes (as we defined in Chapter 3) e.g., political preference use a joint user-neighbor stream.

**Use-case II:** If one needs to deliver ads to a true target crowd but is willing to trade it off for time:

- use models with higher precision (AWR, IBR);
- apply higher decision threshold e.g., 0.95;
- models with rational filtering (IBR, AWR) require less computation (lower-dimensional feature vectors), are more accurate but annotations cost money.

### 5.5 Conclusions

In Chapter 5 we proposed novel approaches for making predictions over dynamically evolving social media streams based on incremental Bayes online updates. We studied an iterative incremental retraining in batch and active settings with and without iterative oracle annotations. Moreover, we applied interactive feature (a.k.a. rationale) annotation technique as a filter for iterative retraining of the proposed models. Finally, we incorporated social network structure information by making
CHAPTER 5. ITERATIVE LEARNING AND PREDICTION

predictions from neighbor and joint user-neighbor streams.

Our key findings on dynamic approaches that are capable of learning and making predictions of latent user properties over time include:

- **Active learning outperforms iterative batch re-training** Active retraining with correctly classified users from test data added to the training data at every time stamp significantly outperforms iterative batch retraining setup.

- **Joint user-neighbor streams are more effective than user streams** Making predictions using a joint user-neighbor or neighbor stream of communications is more effective than using only user communication stream.

- **Precision/recall trade-off** Models with higher confidence yield higher precision and models with lower confidence yield higher recall for both batch and active setting. However, in constrained-resource conditions, when the size of training and test data is very small, the difference in precision and recall between these models is not pronounced.

- **Rationale annotation significantly improves classification results** Rationale annotation and filtering during iterative retraining leads up to 27.7% relative improvement in iterative batch and 20.6% in active setting.

- **Oracle annotations yields better performance** Active retraining with oracle annotations yields the highest recall: 85% of test users are correctly classified after the second iteration using a joint user-neighbor stream.

Our approaches from Chapter 4 and 5 for making predictions over dynamically
evolving social media streams based on incremental Bayes online updates can be effectively used in:

- real-time streaming scenarios for dynamically growing social networks;
- constrained-resource training conditions e.g., iterative retraining and active learning will allow exploring new under-studied attributes e.g., religion, income etc. for which no or limited labeled data exists;
- constrained-resource prediction settings e.g., when no or limited user data is available at any given time, neighbor or user-neighbor streams can be effectively used to make predictions about the user;
- low-cost annotation models that rely on iterative instances (assigning class labels to users) or feature annotations (highlighting predictive words in tweets) via crowdsourcing. Moreover, our batch and active models with iterative rational filtering help to reduce storage and memory requirements when processing large feature vectors and iteratively re-training models for real-time predictions.

In Chapter 6 we propose another set of models for incorporating human knowledge into the learning process. In particular, extending the rationale work from Chapter 5 we further evaluate the benefits of rationales – highly predictive features annotated by humans taking gender prediction task as an example. We compare our new rationale weighting models with the state-of-the-art approaches that rely on bag-of-word features for two languages – English and Spanish.
Chapter 6

Learning with Annotator Rationales

As we showed in Chapter 2 there is a substantial prior work on characterizing users in social media. Another promising yet understudied area of research is to elicit and utilize annotator rationales, targeted annotator feedback regarding why/how they chose a particular annotation. The primary example of this approach in the NLP literature is by O. Zaidan, Eisner, and Piatko (2007), who used highlighted substrings of text as enhanced feedback to improve sentiment classification of movie reviews, with follow-on work by O. Zaidan and Eisner (2008) and Yessenalina et al. (2010). In Chapter 5 we successfully applied annotator rationale filtering techniques for demographic prediction task in social media.
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

In this Chapter\footnote{This chapter presents “Improving Gender Prediction of Social Media Users via Weighted Annotator Rationales”, which was published in the Proceedings of the NIPS Workshop on Personalization: Methods and Applications in 2014 and is a joint work with David Yarowsky.} we extend our work on using rationales for personal analytics and propose novel models that rely on rationale weighting rather than filtering approaches for gender prediction task, with additional contributions including:

- developing effective new ways to incorporate human domain knowledge by weighting tweets and elicited rationales in a (i) supervised and (ii) semi-supervised setting to improve user attribute classification in \ref{6.2};

- empirically assessing the benefits of the rationales and showing the advantages of rationale annotation and weighting over the state-of-the-art models for user attribute inference, in both English and Spanish in \ref{6.3}.

The cost efficiency of the proposed rationale annotation and weighting approach used in a semi-supervised bootstrapping setting will aid scaling of latent user attribute prediction to resource-limited domains and languages.

6.1 Data

For the experiments in this Chapter, we use three sets of data for each language:

I. a large pool of unlabeled data (1M tweets): for English 12.6k users with on average 78 tweets per user, and for Spanish 7.5k users with on average 132 tweets per user;
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

II. a small amount of training data labeled with user demographic attributes e.g.,
   gender: for English 164 male and 193 female users, and for Spanish 251 male
   and 192 female users (each user is associated with 200 tweets);

III. held out test data: 100 male and 100 female users with 200 tweets per user.

The labeled training and test data is used in a supervised classification setting. The
unlabeled data is used in semi-supervised setting to boost the performance of the
existing supervised models for latent attribute prediction.

Figure 6.1: Gender rationale n-gram distribution for English and Spanish.

To collect the data we randomly sampled users from the 1% Twitter feed and
downloaded 200 of their most recent tweets using the Twitter API. We obtained
gender labels using 3-way redundant annotation on Mechanical Turk. Each annotator
was given a link to a user profile and had to judge user demographic attributes based
on user tweets. We estimate the final label using the majority voting. The annotation
agreement among three annotators exceeds 70%, and between two annotators exceeds
90%. We also asked each annotator to highlight words or phrases \(- n\text{-}grams \leq 3\) in
user self-authored tweets that are highly indicative of user gender, and assign their
confidence in each rationale on a 4-point scale.

In addition, we asked each annotator to highlight words or phrases \(- n\text{-}grams \leq 3\)
in user self-authored tweets that are highly indicative of the targeted user demographic
attributes, in this case gender, and assign their confidence in each rationale on a 4-
point scale. The example screenshots of Mechanical Turk HITs are presented in
Appendix A Figures A and A.

Figure F.1 illustrates the most frequent male and female rationales collected for
English.\(^2\) The crowdsourced rationales resemble the results of another work that
analyses language of gender in social media (Schwartz, Eichstaedt, Kern, Dziurzynski,
Ramones, et al., 2013; Kern, Eichstaedt, Schwartz, Dziurzynski, et al., 2014; Sap et
al., 2014).\(^3\) For instance, male rationales include \textit{dude, football, money, gambling,}
and female rationales include \textit{boyfriend, sexy, phone, hair, shopping, bieber}.

We also report the distribution of rationale n-grams for both English and Spanish
in Figure 6.1. We observe that for English the overlap of crowdsourced rationales
across multiple annotators is 30.5% for unigrams, 8% for bigrams and less than 2%
for trigrams. For Spanish the trend is similar.

\(^2\)Appendix F also includes rationale word clouds for age and political preference attributes.
\(^3\)World Well Being Project http://wwbp.org/.
6.2 Methodology

Below we present our supervised and semi-supervised self-trained models with feature (rationale) weighting schemes to improve the existing attribute classification approaches described in Chapter 3.

6.2.1 Models

As input, we are given a set of users \( u \in U \) and their self-authored communications, e.g. their \( T \) tweets. Each user is associated with a set of 200 most recent tweets. As defined in earlier chapters, our goal is to predict an attribute \( a \in A \) for each user \( u \in U \), e.g. gender \( a \in \{ \text{Male, Female} \} \). For any \( t \in T \), \( a \in A \), the model defines a probability:

\[
p(a \mid t, \theta) = \frac{\exp(\theta \cdot f(t, a))}{\sum_{a' \in A} \exp(\theta \cdot f(t, a'))}
\]

(6.1)

where \( f : T \times A \rightarrow \mathbb{R}^d \) is a function that maps any attribute-communication pair \( (t, a) \) to a feature vector \( f(t, a) \). \( \theta \in \mathbb{R}^d \) is a parameter vector to learn \( (d \) is the number of features and parameters in the model); \( \theta \cdot f(t, a) = \sum_{k=1}^{d} \theta_k f_k(t, a) \) is the inner product between \( \theta \) and \( f(t, a) \). Recall that the log-linear model for such classification as defined in the Eq. 2.3 and Eq. 3.1 is:

\[
\Phi(u) = \begin{cases} 
\text{Male} & p(a \mid t, \theta) \geq 0.5, \\
\text{Female} & \text{otherwise}.
\end{cases}
\]

(6.2)
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

6.2.1.1 Direct Model

Our direct model represents a commonly-observed supervised classification setting on this task. We train our model on labeled users from Train and apply it to 200 users from Test following the Eq 6.1 and Eq 6.2. This model is learned from the labeled user tweets exclusively.

6.2.1.2 Transitivity I Model

Given a large pool of unlabeled users and their tweets, we propose to train a direct model $\Phi(u)$ and apply it to assign labels to the thousands of unlabeled users. Then, we suggest to train a new model $\Phi_{1M}(u)$ in a semi-supervised setting, and apply both $\Phi(u)$ and $\Phi_{1M}(u)$ models to classify 200 users in Test:

$$\Phi'(u) = \lambda \cdot \Phi(u) + (1 - \lambda) \cdot \Phi_{1M}(u)$$  \quad (6.3)

6.2.1.3 Transitivity II Model

Similarly to the $\Phi'(u)$ model, we train a new $\Phi_{1M}(u)$ model on a large pool of unlabeled users, and then retrain the model on its own set of users (Clark, Curran, & Osborne 2003). We apply $\Phi(u)$ and $\Phi_{1M}^{self}(u)$ models to classify 200 users in Test as defined below:

$$\Phi''(u) = \lambda \cdot \Phi(u) + (1 - \lambda) \cdot \Phi_{1M}^{self}(u)$$  \quad (6.4)
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

6.2.2 Weighting Rationales

To incorporate attribute-specific rationales into the models defined in Eq. 6.1 - 6.3 we propose three feature weighting schemes as shown in Algorithm 2.

**Data:** \( r \): a list of rationales for each attribute value \( a \in A \),
\( a \in \{\text{Male, Female}\} \); \( f \): a list of frequencies for the rationales in \( r \).

**Parameters:** \( \xi \): parameter to control rationale weights \( \xi \in \{1, \ldots, 200\} \).

**Result:** Data with the weighted annotator rationales.

**foreach** attribute value \( a \in A \) **do**

**foreach** rationale n-gram \( r_j \in r \) **do**

if scheme == I **then**

| generate \( \xi f_j \) new users with \( r_j \) rationale n-grams per tweet

else if scheme == II **then**

| generate \( \xi \) new users with \( f_j \) tweets and \( r_j \) rationales per tweet

else if scheme == III **then**

| randomly sample \( f_j \) existing users for each attribute value \( a \in A \)
| and generate \( \xi \) tweets with \( r_j \) rationale n-grams per tweet

end

end

**Algorithm 2:** WEIGHTRATIONALS \((r, f, \xi)\)

As input we are given user self-authored communications and a list of attribute-specific rationales including \( m \) male and \( n \) female rationales \( r \in R \) for gender attribute \( a \in \{\text{Male, Female}\} \). The rationales \( r \) are associated with frequency \( f \geq 1 \). We propose to incorporate rationales into the existing models for predicting author gender using three weighting schemes described below.

For **weighting scheme I** we generate \( \xi \sum_{a \in A} f \cdot r \) new data points to encode users with rationales; \( \xi \) is the parameter to be optimized. In total, we generate \( \xi (m + n) \sum_{a \in A} |f|_1 \) new users encoded using sparse feature vectors of n-grams. For instance, for the male rationale \( r = \text{"gambling"} \) with \( f = 3 \) and \( \xi = 5 \) we generate 15
new users with training instances containing the rationale n-gram “gambling”.

For weighting scheme II we generate $\xi \sum_{a \in A} r$ new data points to encode users with $f$ rationales. In total, we generate $\xi (m + n)$ new users with less sparse feature vectors compared to the scheme A. Following the example rationale “gambling”, we generate 5 new users with the training instance “gambling gambling gambling gambling”.

For weighting scheme III we modify $f$ randomly sampled existing data points by adding $\xi$ tweets with $r$ rationales per tweet for each data point. Following the example rationale “gambling”, we randomly sample 3 male users from the existing users and generate 5 training instances with the rationale n-gram “gambling”.

6.3 Experiments

6.3.1 Experimental Setup

We train logistic regression classifiers as shown in Eq.6.1 and 6.2 via LIBLINEAR (Rong et al., 2008) integrated into Jerboa toolkit (Van Durme, 2012a). We optimize the classifier regularization parameters on the development data. We randomly sample 20% of the training data as development data. The remaining disjoint 80% is used for training. We report the final results for 200 users from the test data.
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

6.3.2 Experimental Results

In Figures 6.2a and 6.2b we present accuracy results for gender classification using the baseline direct model $\Phi(u)$ defined in Eq. 6.2 for English and Spanish, respectively. In contrast, we find that using only the most confident rationales ($R'$), with annotator confidence $\geq 3$, yields lower accuracy compared to using all rationales in all other experimental variables for both languages except for some cases using weighting scheme III. Moreover, most of our rationale weighting schemes outperform the baseline supervised model by 8% for English and 6% for Spanish in accuracy.

Interestingly, we also discovered that when using rationales combined with raw tweets as user features, we could improve performance by filtering the tweets to include only those containing at least one rationale n-gram ($T' + R$) rather than using all tweets ($T + R$). As shown in Figure 6.2, ($T' + R$) $\geq$ ($T + R$) $\geq$ $T$ $\geq$ $R$. The trend is the same for using only confident rationales $R'$. For example, $\Phi(u)$ model trained for English using weighting scheme I yields the results: 0.74 $>$ 0.72 $>$ 0.66 $>$ 0.61. Similarly, $\Phi(u)$ trained for Spanish using weighting scheme I yields the results: 0.67 $>$ 0.65 $>$ 0.61 $>$ 0.60. We get these improvements because (a) more accurate data (or less noisy data) is better and (b) features are less sparse and highly discriminative features e.g., rationales are ranked higher compared to all other features.

In addition, we made a comparison with an contrastive distilled-feature resource, the list of conceptual class attributes for gender collected by Bergsma & Van Durme [2013]. The list contains 958 Male and 659 Female n-grams. In Figure 6.2 we refer to
Figure 6.2: Gender prediction accuracy using the direct model \( \Phi(u) \) with weighted annotator rationales for English and Spanish.

them as \( R^{BV} \) rationales. We find that \( R^{BV} \) features perform significantly better than confident rationales but significantly worse (schema I) or comparably (schemes II and III) to using all rationales when models are learned from rationales only. When we combine tweets with rationales, models learned from our rationale plus a tweet mix
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

Table 6.1: Gender classification results for English using $\Phi(u)$, $\Phi'(u)$ and $\Phi''(u)$ models with weighted annotator rationales. Models are trained on $T$: tweets only, $R$: rationales only, $T + R$: all tweets + rationales, $T' + R$: filtered tweets + rationales. $\Delta E_{max}$ is a relative error reduction of $T' + R$ or $T + R$ compared to $T$.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>0.66</td>
<td>0.65</td>
<td>0.66</td>
<td>0.66</td>
<td>0.65</td>
<td>0.57</td>
<td>0.63</td>
<td>0.55</td>
<td>0.63</td>
</tr>
<tr>
<td>$R$</td>
<td>0.61</td>
<td>0.55</td>
<td>0.63</td>
<td>0.65</td>
<td>0.55</td>
<td>0.57</td>
<td>0.74</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>$T \cup R$</td>
<td>0.72</td>
<td>0.71</td>
<td>0.71</td>
<td>0.74</td>
<td>0.72</td>
<td>0.68</td>
<td>0.74</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>$T' \cup R$</td>
<td>0.74</td>
<td>0.70</td>
<td>0.69</td>
<td>0.74</td>
<td>0.70</td>
<td>0.75</td>
<td>0.74</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>$\Delta E_{max}$</td>
<td>0.08</td>
<td>0.05</td>
<td>0.05</td>
<td>0.09</td>
<td>0.07</td>
<td>0.10</td>
<td>0.08</td>
<td>0.05</td>
<td>0.06</td>
</tr>
</tbody>
</table>

$T + R$ and $T' + R$ significantly outperform the models learned from the tweet plus $R^{BV}$ mix for all weighting schemes.

Finally, we report experimental results for English $\Phi(u)'$ models in Table 6.1. We find that $\Phi(u)'$ models trained in semi-supervised setting exclusively on tweets $T$ do not yield statistically significant improvements over the baseline $\Phi(u)$. However, when user tweets are combined with rationales $T + R$ the absolute gain is 2% when weighting scheme I is applied. Moreover, when the tweets filtered to only those tweets that contain rationale n-grams ($T'$) are mixed with raw rationales to train the model $\Phi(u)'$ with weighting scheme III, the absolute gain is the highest – 10% over the baseline $T$ (the error reduction is $\Delta E = 28\%$).
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

6.3.3 Comparison with Previous Work

As we described in the literature review section the majority of the existing models for latent author attribute prediction define the task as a supervised classification. They rely on thousands of user self-authored tweets (primarily in English) trained using bag-of-word (BoW) lexical features. Less works study gender prediction for languages other than English (Ciot et al., 2013; Volkova et al., 2013a). For example, Ciot et al. (2013) report comparable to our classification accuracy for Spanish – 0.76 for French and 0.63 for Japanese.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Users</th>
<th>Tweets (per user)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rao et al. (2010)</td>
<td>1k</td>
<td>0.4M (4k)</td>
<td>0.687</td>
</tr>
<tr>
<td>Burger et al. (2011)</td>
<td>184k</td>
<td>4M (22)</td>
<td>0.745</td>
</tr>
<tr>
<td>Zamal et al. (2012)</td>
<td>400</td>
<td>400k (1k)</td>
<td>0.795</td>
</tr>
<tr>
<td>Bergsma et al. (2013)</td>
<td>400</td>
<td>4B (500)</td>
<td>0.720</td>
</tr>
<tr>
<td>Volkova (Chapter 3)</td>
<td>384</td>
<td>77k (200)</td>
<td>0.82</td>
</tr>
<tr>
<td>BoW Baseline $T$ (only)</td>
<td>357</td>
<td>70k (200)</td>
<td>0.66</td>
</tr>
<tr>
<td>Rationale $T' + R$</td>
<td>357</td>
<td>70k (200)</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 6.2: Comparing our rationale-weighted models with previously published approaches to gender classification.

To compare our models with the existing approaches for gender prediction we present a brief quantitative comparison in Table 6.2. These models are all trained in a supervised setting with various features, with the comparable bag-of-words feature performances marked in bold. Our best model outperforms the BoW baseline presented by Rao et al. (2010) by an absolute 6%, as well as their other feature combinations by 3%. Moreover, we achieve comparable accuracy with the similar character n-gram model presented by Burger et al. (2011), but learned from millions of tweets.
Furthermore, our work achieves 3% absolute performance gain relative to the bootstrapped models presented by (Bergsma & Van Durme, 2013) using conceptual class attributes over the same amount of training examples (400 users). Only when their models are bootstrapped from billions of tweets does the final accuracy increase to 0.87; we assume dramatically less data.

The baseline user models proposed by Zamal et al. (2012) and in Chapter 3 yield higher accuracy compared to our rationale-weighted models, 4.5% and 7% respectively. Both Zamal’s and models from Chapter 3 have been tested on the same set of Twitter users. This dramatic difference on how the state-of-the-art BoW model perform on different samples of Twitter users is another evidence of data sampling and annotation biases present in many Twitter datasets.

### 6.4 Additional Validation Experiments

In addition to our rationale filtering techniques in Chapter 5 and rationale weighting approaches in this Chapter, we run a set of experiments to further validate the effect of annotator rationales on demographic prediction. For this validation we use a large, gender-annotated Twitter dataset from (Burger et al., 2011). Gender labels were collected by following the URLs from Twitter profiles to other blog sites where the gender information was provided. The total number of users who tweet in English according to our LID tool (Bergsma et al., 2012) is 91,621 (98,004 reported
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

in the original paper). We split this data into three parts: train $U^{train} = 71,342$
with 46% female users, development $U^{dev} = 10,161$ with 46% female users, and test
$U^{test} = 10,234$ with 47% male users (keeping the same set of users studied by Burger
et al. (2011)). We use only train and test splits for the experiments below.

For every user in the dataset, we converted raw tweets into features by removing
URLs, usermentions (e.g., @kittyfru), punctuation, stopwords, digits, and words
with fewer than two characters.

In Chapter 3 section 3.4.4 we showed how prediction performance depends on
the amount of content available per user (Burger et al., 2011; Cohen & Ruths, 2013;
Volkova et al., 2014). Thus, to understand our results better we estimate the number
of tweets per user in the dataset. We plot the distribution of users and their total
number of tweets in Figure 6.3. We found that 17.3% of users have only 1 tweet,
55.3% have $\leq 5$ tweets, 71.3% have $\leq 10$ tweets, 86.9% have $\leq 25$, 93.6% $\leq 50$,
97.1% have $\leq 100$ tweets and only 1% have at least 200 tweets per user. Note than
more than half of the users in our dataset have fewer than five tweets.

6.4.1 Removing Rationales from Tweets

To further investigate the usefulness of rationales for latent attribute prediction,
we take a similar approach to O. F. Zaidan, Eisner, and Piatko (2008). More precisely,
we remove all rationale n-grams from user tweets (in both training and test

\footnote{To remove English stopwords we used python package available at: https://pypi.python.org/pypi/stop-words}
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

Figure 6.3: The distribution of tweets per user in English portion of MITRE dataset from [Burger et al. 2011].

data) and demonstrate how it influences classification accuracy. We experiment with three rationale lists for the gender attribute which include: 314 female and 658 male rationales from [Bergsma and Van Durme 2013], our 1175 female and 873 male rationales [Volkova & Yarowsky 2014], and the lexicon from [Sap et al. 2014]. We used logistic-regression classifiers trained using binary word unigram features as described in detail in Chapter 3. We present the results in Table 6.3.

<table>
<thead>
<tr>
<th>Rationale List</th>
<th>Size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original test data</td>
<td>–</td>
<td>0.713</td>
</tr>
<tr>
<td>Bergsma and Van Durme (2013)</td>
<td>972</td>
<td>0.709</td>
</tr>
<tr>
<td>Volkova and Yarowsky (2014)</td>
<td>2,048</td>
<td>0.694</td>
</tr>
<tr>
<td>Sap et al. (2014)</td>
<td>7,139</td>
<td>0.659</td>
</tr>
</tbody>
</table>

Table 6.3: Gender classification performance when annotator rationales have been removed from tweets.
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

Similar to O. F. Zaidan et al. (2008), we found that when annotator rationales are removed from the original communications, performance falls. For our gender prediction task, accuracy drops significantly given the size of the test dataset (10,234 users). For example, accuracy decreases from 0.713 to 0.694 using our rationale lists, and it falls to 0.659 using the rationales from Sap et al. (2014). These results further confirm that rationales indeed carry a strong signal useful for user demographic attribute prediction.

6.4.2 Matching Rationales

Lastly, we test annotator rationales using a basic keyword matching approach. We count the number of rationales for every user in the train and test sets in Burger’s et al. data. We label a user as female if the number of female rationales is higher than the number of male rationales, and vice versa. The results of this rationale matching approach are presented in Table 6.4.

We report the total number of users in train and test sets (total), the total number of users with at least one annotator rationale matched (matched), and the number of users for which our predicted label is the same as the true label (correct). We also report precision and recall numbers that reflect the predictive power of the rationales as shown below:

\[
\text{Precision} = \frac{\text{correct users}}{\text{matched users}} \quad \text{and} \quad \text{Recall} = \frac{\text{correct users}}{\text{total users}}.
\]
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

Our results in Table 6.4 demonstrate that a simple rationale matching approach yields higher performance than the baseline majority class labeling (0.46). We found that 15% – 17% of all users in the train and test sets use at least one rationale from Bergsma and Van Durme (2013), whereas 88% – 89% of users in train and test generate at least one rationale from Volkova and Yarowsky (2014). We also found that the rationales of Volkova et al. yield higher recall, while those of Bergsma et al. yield higher precision. Finally, as expected, models trained using logistic regression significantly outperform the rationale matching approach, an accuracy of 0.713 for test and 0.710 for train data.

We were not able to use the gender lexicon from Sap et al. (2014) for the keyword matching approach, because they do not discriminate rationales between genders.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Total</th>
<th>Matched</th>
<th>Correct</th>
<th>( P )</th>
<th>( R )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TEST</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original test data (log-linear model) accuracy = 0.713</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergsma and Van Durme (2013)</td>
<td>10,234</td>
<td>1,732</td>
<td>1,162</td>
<td>0.671</td>
<td>0.113</td>
</tr>
<tr>
<td>Volkova and Yarowsky (2014)</td>
<td>10,234</td>
<td>9,135</td>
<td>5,216</td>
<td>0.570</td>
<td>0.514</td>
</tr>
<tr>
<td><strong>TRAIN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original train data (log-linear model, 10-fold c.v.) accuracy = 0.710</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergsma and Van Durme (2013)</td>
<td>71,342</td>
<td>11,039</td>
<td>7,271</td>
<td>0.659</td>
<td>0.101</td>
</tr>
<tr>
<td>Volkova and Yarowsky (2014)</td>
<td>71,342</td>
<td>63,056</td>
<td>36,337</td>
<td>0.576</td>
<td>0.509</td>
</tr>
</tbody>
</table>

Table 6.4: Gender classification performance when matching annotator rationales (aka basic keyword matching).
6.5 Conclusions

In Chapter 6, we proposed several readily-replicable new models for latent attribute classification of social media users for English and Spanish that outperform the state-of-the-art models learned exclusively from user data. We introduced three novel rationale weighting schemes integrated into different models with varied amount of supervision. Our key findings on the advantages of rationales for classifying user attributes in social media are listed below:

- $T$ vs. $T' + R$: incorporating rationales as additional informative features into the models is beneficial for gender prediction either in fully supervised (the largest relative error rate reduction is 24%) or semi-supervised bootstrapping setting (the largest relative error rate reduction is 28%);

- $R'$ vs $R$: using all rationales is better than using just confident rationales: 2 - 12% accuracy gain for English and 6 - 9% for Spanish;

- $T$ vs. $R$: in the common experimental setting where the collected Twitter data cannot be shared with others, distilled rationales alone can be used to train the models leading to only a 3% absolute accuracy loss for English and 1% for Spanish.

- $T + R$ vs. $T' + R$: applying rationales in combination with filtered tweets is better than mixing rationales with all tweets available for a given user – up to 3% absolute accuracy gain for English and 1.5% for Spanish.
CHAPTER 6. LEARNING WITH ANNOTATOR RATIONALES

Finally, the investment in rationale annotation is very cost-effective; a 28\% relative error reduction is achieved with only a $10 total additional Mechanical Turk cost to collect the rationales in this reported experimental setup. Furthermore, the value of using rationales to improve performance on this task is not only about money; many domains have limited raw data or severely volume limited APIs or IP constraints, making our demonstrated rationale-based performance gains with no additional raw data even more valuable.

In addition, we run two validation experiments that further demonstrate the benefits of using annotator rationales for latent attribute prediction task in social media. First, we showed that removing annotator rationales from the original communications decreases classification performance. Second, using a simple rationale matching approach demonstrates and compares the predictive power of annotator rationales collected manually (Bergsma & Van Durme, 2013) vs. using crowdsourcing (Volkova & Yarowsky, 2014).

In part II of this thesis we focus on (a) feature analysis and (b) applications for user demographic prediction task in social media.

We start by presenting some background work on the relationships between user demographics, interests, emotions and sentiments in social media in Chapter 7. We then investigate the relationships between emotions and sentiments expressed in user tweets and user attributes in Chapter 8, and user interests and latent user properties in Chapter 9.
In particular, in Chapter 8 we apply static models for personal analytics in social media presented in Chapter 3 and create a framework for automatically predicting a variety of previously unexplored latent user properties e.g., education, income, life satisfaction etc. The proposed framework allows to study correlations between user emotions and opinions and perceived psycho-demographic traits, as well as emotion and opinion differences between users and their neighbors in social networks on an unprecedented scale – 25M tweets, 123k users.
PART II: Feature Analysis and Applications
Chapter 7

Background: Affect and Demographics in Social Networks

In this chapter we give a brief overview on recent works for subjective language recognition in social networks such as opinion and emotion detection and relate it to latent user attribute classification task.

7.1 Emotion Detection

Emotions and sentiments affect the way we feel, think, and what we say or do in online social networks. Though both are affective states, there are important differences between them. Emotions are states of consciousness in which various internal sensations are experienced. They can be triggered by sights, smells, sounds
and events in the external environment. Sentiments are our likes and dislikes, and they involve a person-object relationship e.g., people express sentiments towards products, services or brands. Emotions are relatively short in duration, while sentiments display themselves over longer periods of time (Desmet, 2002).

Emotion analysis has been successfully applied to many kinds of informal and short texts including emails (S. Mohammad & Yang, 2011), blogs (Aman & Szpakowicz, 2007; Kosinski et al., 2013), and news headlines (Strapparava & Mihalcea, 2007), but emotions in social media, including Twitter and Facebook, have only been investigated recently (Bollen, Mao, & Pepe, 2011; Roberts et al., 2012; De Choudhury et al., 2012; W. Wang et al., 2012; Kim et al., 2012; Qadir & Riloff, 2013; S. M. Mohammad & Kiritchenko, 2014). In contrast, sentiment classification in social media has been extensively studied (Pang et al., 2002; Pang & Lee, 2008; Go et al., 2009; Pak & Paroubek, 2010; Hassan Saif, Miriam Fernandez & Alani, 2013; Nakov et al., 2013; S. M. Mohammad et al., 2013; Zhu et al., 2014).

Researchers have used supervised learning models trained on lexical word ngram features, synsets, emoticons, topics, and lexicon frameworks e.g., LIWC, MPQA, WordNet-Affect to determine which emotions are expressed on Twitter (W. Wang et al., 2012; Roberts et al., 2012; Kim et al., 2012; Qadir & Riloff, 2013; S. M. Mohammad & Kiritchenko, 2014). Due to the lack of social media data annotated with emotions and opinions, this work bootstraps noisy labeled data for sentiment (Go

1Mood Changes in UK Twitter Content: http://mediapatterns.enm.bris.ac.uk/mood/
et al., 2009; Pak & Paroubek, 2010), sarcasm (González-Ibáñez et al., 2011) and emotions (W. Wang et al., 2012; De Choudhury et al., 2012; Qadir & Riloff, 2013; S. M. Mohammad & Kiritchenko, 2014), and bases emotion prediction training on hashtags (e.g., #happy, #sad) or emoticons (Purver & Battersby, 2012). This bootstrapped data allows learning models using lexical or syntactic features. We use a similar technique, and build a hashtag dataset annotated with emotions, from which we train models for automatic emotion prediction.

This work focuses on capturing the 6 high-level emotions proposed by Ekman (1992): joy, anger, sadness, fear, disgust and surprise. Other papers study moods, including tension, depression, anger, vigor, fatigue, and confusion (Bollen, Mao, & Pepe, 2011) and issues such as politeness, embarrassment, formality, deception etc. (Pearl & Steyvers, 2010).

7.2 Emotion Contagion and User Attributes

Emotional contagion theory states that emotions and sentiments of two messages posted by friends are more likely to be similar than those of two randomly selected messages (Hatfield & Cacioppo, 1994). In other words, homophily leads not only to the connections between users with similar properties e.g., gender, age (Zamal et al., 2012) or political preferences (Volkova et al., 2014) but also to so called “affective” connections between the users with shared emotions and sentiments in social networks.
as the saying “birds of a feather flock together” states (McPherson et al., 2001).

There have been recent studies about emotion contagion in massively large social networks. One study shows that anger is more influential than joy in a Weibo (Twitter like) social network in China. It states that angry tweets can spread quickly and broadly in the network (R. Fan et al., 2013). Another study reveals that psychological states like happiness are assortative in a Twitter social network – happy people tend to be friends with happy people whereas unhappy users connect with unhappy users (Bollen, Gonçalves, et al., 2011; Dodds et al., 2011). It also demonstrates that the average happiness scores are positively correlated among the users and their local social circles. Our work also deals with the relation between a user’s emotions and the emotions of their social network friends, but we study how the emotional contrast between the user and its environment relates the demographic traits of the user. Finally, a controversial study (Baldridge, 2014) on emotion contagion in a Facebook social network further confirms that users are indeed susceptible to the emotion observed in their social environments online (Coviello et al., 2014).

In contrast to those efforts, we do not aim to model the spread of emotions or opinions in a social network. Instead, given both homophilic and assortative properties of a social network, we study how emotions expressed in a user’s environments correlate with that user’s emotions, and whether these correlations depend on the user’s demographic traits. Researchers have only recently started examining the relation between demographics and social network emotions. Some work examined the relation between
CHAPTER 7. BACKGROUND: AFFECT AND DEMOGRAPHICS IN SOCIAL NETWORKS

gender and sentiment (S. Mohammad & Yang, 2011; Volkova et al., 2013a), between personality and emotions (S. Mohammad & Kiritchenko, 2013) and between emotions and interests (Lewenberg, Bachrach, & Volkova, 2015). For instance, (Volkova et al., 2013a) showed that sentiment classification can be improved through learning sentiment classifiers form gender-specific sets of lexical features. Other research demonstrated that the “big-five” personality traits (extroversion, neurotism, agreeableness, conscientiousness, and openness) can not only be predicted from language in social media (Golbeck et al., 2011; Bachrach et al., 2012; Schwartz, Eichstaedt, Dziurzynski, et al., 2013; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013) but also from fine-grained emotions (S. Mohammad & Kiritchenko, 2013; S. M. Mohammad & Kiritchenko, 2014). We cover a wide range of demographic traits previously un-explored in social media, including income, education, relationship status, optimism and life-satisfaction.

7.3 Opinion Mining and User Properties

Numerous studies since the early 1970’s have investigated gender-language differences in interaction, theme, and grammar among other topics (Schiffman, 2002; Sunderland, Duann, & Bake, 2002). More recent research has studied gender differences in telephone speech (Cieri, Miller, & Walker, 2004; Godfrey, Holliman, & McDaniel, 1992b) and emails (Styler, 2011). In addition, the authors showed and
emotional word usage written across genders in a subset of emails from the Enron corpus.\textsuperscript{2} S. Mohammad and Yang \textsuperscript{(2011)} analyzed gender differences in the expression of sentiment in and emotional word usage across genders in email, love letters, hate mail \textsuperscript{4} and suicide notes \textsuperscript{5} corpora.

There has also been a considerable amount of work in subjectivity and sentiment analysis over the past decade, including, more recently, in microblogs (Barbosa & Feng \textsuperscript{2010}; Bermingham & Smeaton \textsuperscript{2010}; Pak & Paroubek \textsuperscript{2010}; Bifet & Frank \textsuperscript{2010}; Davidov et al. \textsuperscript{2010}; G. Li et al. \textsuperscript{2010}; Kouloumpis et al. \textsuperscript{2011}; Agarwal et al. \textsuperscript{2011}; X. Wang et al. \textsuperscript{2011}; Calais Guerra et al. \textsuperscript{2011}; Tan et al. \textsuperscript{2011}; Chen et al. \textsuperscript{2012}). In spite of the surge of research in both sentiment and social media, only a limited amount of work focusing on gender identification has looked at differences in subjective language across genders within social media. Thelwall et al. \textsuperscript{(2010)} found that female users express positive emoticons more than male users on MySpace.

Other researchers included subjective patterns as features for gender classification of Twitter users (Rao et al. \textsuperscript{2010}). These socio-linguistic features included emoticons \textit{e.g.}, :-( or :-*), repeated letters \textit{e.g.}, niceeee, noooo, laughter abbreviations \textit{e.g.}, LOL, repeated punctuation \textit{e.g.}, !!!?!?!!, exasperation \textit{e.g.}, ahh, grrrr, shout, excitement \textit{etc.} They found that the majority of emotion-bearing features, \textit{e.g.}, emoticons, repeated letters, exasperation, are used more by female than male users, which is consistent

\textsuperscript{2}EnronSent corpus: \url{http://verbs.colorado.edu/enronsent/}
\textsuperscript{3}Love Letter Corpus: \url{http://www.lovingyou.com/content/inspiration/loveletters}
\textsuperscript{4}Hate Mail Corpus: \url{http://www.ratbags.com}
\textsuperscript{5}Suicide Notes: \url{http://www.well.com/art/suicidenotes.html}
CHAPTER 7. BACKGROUND: AFFECT AND DEMOGRAPHICS IN SOCIAL NETWORKS

with results reported in other recent work (Garera & Yarowsky, 2009; Burger et al., 2011; Goswami et al., 2009; Argamon et al., 2007). Other related work is that of Otterbacher (Otterbacher, 2010), who studied stylistic differences between male and female reviewers writing product reviews, and Mukherjee and Liu (2010), who applied positive, negative and emotional connotation features for gender classification in microblogs.

Although previous work has investigated gender differences in the use of subjective language, and features of sentiment have been used in gender identification, to the best of our knowledge no one has yet investigated whether gender differences in the use of subjective language can be exploited to improve sentiment classification in English or any other language. In this work we seek to answer this question for the domain of social media.
Chapter 8

Learning from Emotions and User-Neighbor Emotional Contrast

In this Chapter we apply static approaches presented in Chapter 3 and create a unified real-world framework for social media predictive analytics. Our system examines information on Twitter, determining, first, various properties for a given user based on a set of aggregated tweets for that user and, second, fine-grained emotions expressed in tweets based on the words used in these tweets.

1This chapter presents and significantly extends the demo system that we built and released along with “Inferring Latent User Properties from Texts Published in Social Media”, which was published in the Proceedings of the 29th Conference on Artificial Intelligence (AAAI) in 2015 and is a joint work with Yoram Bachrach, Michael Armstrong and Vijay Sharma.

This chapter presents models used to predict emotions, opinions and psycho-demographic attributes in “Studying User Income through Language, Behaviour and Affect in Social Media”, which was published in PLOS One journal in 2015 and is a joint work with Daniel Preotiuc-Pietro, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras.

This chapter presents “On Predicting Socio-Demographic Traits and Emotions in Social Networks and Implications to Online Self-Disclosure” (to appear) in Cyberpsychology, Behavior, and Social Networking journal in 2015 and is a joint work with Yoram Bachrach.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

First, we describe our methodology that is based on building models for inferring coarse-grained emotions and perceived attributes – subjectively estimated impressions that other users will have when reading the content in the profile. We then show how to use our system to study the relationships between perceived demographic attributes, opinions and emotions on a large sample of Twitter users.

Unlike the existing systems that rely on self-reported annotations (Beller et al., 2014; Coppersmith, Harman, & Dredze, 2014; Coppersmith, Dredze, & Harman, 2014), and labels obtained using distant supervision via following activities (Zamal et al., 2012; Volkova et al., 2014; Culotta et al., 2015) or user names (Burger et al., 2011; Bergsma et al., 2013) (as we outlined in Chapter 1 section 1.3), we examine a variety of perceived attributes, including age, gender, ethnicity, income, political preferences, education, income, children and relationship status as shown in Figure 8.1. We correlate these demographics with the emotional profile emanating from the users’ tweets, as captured by Ekman’s emotion classification.

Figure 8.1: An example prediction of perceived attributes, emotions and sentiments.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Second, we examine communications in a social network to study user emotional contrast – the propensity of users to express different emotions than those expressed by their neighbors in the network as shown in Figure 8.2. Our analysis is based on a large scale Twitter dataset. Focusing on Ekman’s six basic emotions, we analyze differences between the emotional tone expressed by these users and their neighbors.

Figure 8.2: Incoming (environment emotional tone) and outgoing (user emotional tone) emotions and opinions for users with varied predicted perceived demographics.

Carrying out such an analysis requires using a large dataset consisting of thousands of users annotated with a variety of properties, and a large pool of their communications annotated with emotions and sentiments. Creating such a large dataset is difficult. Moreover, ground-truth and sensitive demographics e.g., income, age is not available for the majority of social networks including Twitter (unlike Facebook, Google+ or Youtube), or might be hidden due to privacy settings. Therefore, we start by sampling $U^L = 5,000$ Twitter user profiles taking into account data collection biases discussed by Tufekci (2014). We annotate them with a variety of
perceived attributes via crowdsourcing as described in section 8.1. We then sample $U = 123,513$ Twitter users presented in Table 8.3 and predict their perceived attributes as described in details in section 8.2.

Using crowdsourced annotations to get labeled data for previously unexplored attributes e.g., income, children, optimism level is a commonly used practice in the NLP community. It has some disadvantages (but self-reported attributes also have some shortcomings - e.g. users inflating their education level or reporting a desired image rather than truthful reports). In our crowdsourced experiments, annotators had the entire profile available (and not just the tweets) when tagging profiles, including the profile pictures, photos and videos. While there at least some attributes for which access to e.g., the user’s photo (wrt ethnicity, gender, age), could lead to high confidence labels by annotators, we stress that these annotations are based on subjective responses of annotators with no direct knowledge of the users, outside information a given user has publicly shared on the Internet. In addition, we took many precautions to obtain good quality annotations as we describe in section 8.1.1.1 and confirmed that models trained on our subjectively labeled data but tested on several publicly available datasets demonstrate similar or higher than the baseline performance in section 8.2.4.1.

Our high-level methodology is described in Figure 8.3. We train models for predicting perceived user properties from $U^L = 5,000$ annotated profiles. We predict emotions and opinions emanating from $T = 24,919,528$ user tweets extracted from
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

\[ U = 123,513 \text{ user profiles. We collect } T^L = 52,925 \text{ tweets annotated with emotions and their synonyms via distant supervision from emotion hashtags e.g., } \#\text{anger, } \#\text{fear, } \#\text{joy similar to the existing approaches (De Choudhury et al., 2012; Qadir \\& Riloff, 2013; S. M. Mohammad \\& Kiritchenko, 2014). To perform a reliable analysis we make sure the quality of inferring latent user properties is satisfactory and show that our models have an improvement over the existing systems in section 8.2.4.1.}\]

\[\text{Figure 8.3: Our approach for predicting perceived psycho-demographic attributes, opinions and emotions.}\]
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

8.1 Data

8.1.1 Perceived Demographic Data

For the experiments in this chapter we used several datasets for learning attribute, sentiment and emotion prediction models for Twitter. Unlike Facebook (Bachrach et al., 2012; Kosinski et al., 2013), Twitter profiles do not have personal information attached to the profile e.g., gender, age, education etc. work for gender (Burger et al., 2011; Van Durme, 2012b), age (D. Nguyen et al., 2013), ethnicity (Bergsma et al., 2013) and political preference (Zamal et al., 2012; Cohen & Ruths, 2013; Volkova et al., 2014) prediction. Collecting self-reports (Burger et al., 2011; Zamal et al., 2012) brings data sampling biases which makes models trained on self-reported data less generalizable (as we showed in Chapter 3 section 3.4.5) for making predictions of random Twitter users (also has been recently studied (Cohen & Ruths, 2013; Volkova et al., 2014)). Asking social media users to fill questionnaires (Kosinski et al., 2013; Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013) is time consuming and costs money (unless users do it voluntarily). Therefore, we suggest an alternative way to collect attribute annotations as has been effectively used for a variety of other NLP tasks before (Snow et al., 2008; Callison-Burch, 2009).2

Our attribute prediction dataset consists of $U^L = 5,000$ user profiles annotated with 10 attributes shown in Table 8.1. We first sample English speaking Twitter users

CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Values</th>
<th># Profiles</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnicity</td>
<td>Caucasian (59%), Afr. Amer.</td>
<td>4,114</td>
<td>0.71</td>
</tr>
<tr>
<td>Gender</td>
<td>Female (58%), Male</td>
<td>4,998</td>
<td>0.68</td>
</tr>
<tr>
<td>Children</td>
<td>No (84%), Yes</td>
<td>5,000</td>
<td>0.40</td>
</tr>
<tr>
<td>Age</td>
<td>Below 25 y.o. (65%), Above 25 y.o.</td>
<td>3,883</td>
<td>0.35</td>
</tr>
<tr>
<td>Optimism</td>
<td>Optimist (75%), Pessimist</td>
<td>3,562</td>
<td>0.30</td>
</tr>
<tr>
<td>Life Satisfaction</td>
<td>Satisfied (78%), Dissatisfied</td>
<td>3,789</td>
<td>0.30</td>
</tr>
<tr>
<td>Education</td>
<td>High School (68%), Degree</td>
<td>4,998</td>
<td>0.30</td>
</tr>
<tr>
<td>Income</td>
<td>Under $35k (66%), Over $35k</td>
<td>4,999</td>
<td>0.28</td>
</tr>
<tr>
<td>Political</td>
<td>Liberal (76%), Conservative</td>
<td>2,498</td>
<td>0.13</td>
</tr>
<tr>
<td>Relationship</td>
<td>Single (72%), In a Relations</td>
<td>4,998</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 8.1: Crowdsourced annotations of perceived user properties (attributes are ranked by the annotator agreement coefficient \( \kappa \)).

from the US and Canada\(^3\) who tweet between 4 to 10 times a day (taking into account our findings about user activeness in Appendix B), have more than 20 friends and have been active for a year to exclude celebrity accounts, loudmouths – extremely active users (Rogers, 2010; Valente, 1995) and not active users.

8.1.1.1 Crowdsourced Participant Selection

We crowdsourced perceived attribute annotations to a group of trained and screened workers who reside in the US, have 98% reputation score and have been involved in similar annotation tasks before. In addition, we embed a variety of quality control questions (both hard and easy) for which we know ground truth; the workers have to do all of them correctly for their work to be accepted. Annotators were asked to each examine several Twitter profiles. They were given access to the full profile by access-

\(^3\)We focus on the US/Canada population similarly to (Coviello et al., 2014). We realize that the results may not hold for the entire Twitter population or in other countries. However, we feel that its large and interesting population to justify specific research attention.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

ing it through their web browser, allowing them full access to user bio, tweets, photos and videos. These pieces of information were used to make judgments regarding user properties. The scale of our annotation effort is large, and we used a group of skilled annotators (who earn at least $6 hourly). This results in a high monetary cost – $1,500 for annotating 5,000 profiles (each profile was annotated by only one turker). Thus, to minimize our annotation cost we only obtained a single annotation per target profile. In order to measure the degree of agreement between raters, we collected redundant annotations for a 2% random subsample. The inter-annotator agreement measured using Cohen’s kappa \( \kappa \) is given in Table 8.1. The agreement between raters is substantial for gender and ethnicity, and fair for the remaining attributes (except for political preference and relationship status).

For many classes we collected annotations as for a multi-class problem, beyond binary as shown in Table 8.2. However for this study we performed a post-hoc collapse across the fine-grained classes into coarser bins as shown in Table 8.1. For instance, for education we binarized into two classes – high school and degree (merged undergrad and grad); for ethnicity we took the most frequent classes – Caucasian and African American; for age and income we split at the median numbers – 25 y.o. and $35k. For optimism and life satisfaction we did not take into account annotations like “neither A nor B”. Thus, these conversions led to a different number of total annotated profiles (3rd column in Table 8.1).

We realize that crowdsourcing perceived user properties is not trivial as has been
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Multi-class Attribute Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female: 2874, Male: 2124</td>
</tr>
<tr>
<td>Children</td>
<td>No: 4203, Yes: 797</td>
</tr>
<tr>
<td>Optimism</td>
<td>Optimist: 2277, Extreme optimist: 378, Pessimist: 813, Extreme pessimist: 94</td>
</tr>
<tr>
<td>Education</td>
<td>High School: 3423, Bachelor: 1414, Graduate: 161</td>
</tr>
<tr>
<td>Income</td>
<td>Under $35k: 3324, Between $35k - $75k: 1442, Over $75k: 233</td>
</tr>
<tr>
<td>Political</td>
<td>Unaffiliated: 2070, Liberal: 1903, Conservative: 595, Independent: 432</td>
</tr>
<tr>
<td>Relationship</td>
<td>Single: 3302, In a relationship: 1015, Married: 368, Other: 283, Divorced: 30</td>
</tr>
</tbody>
</table>

Table 8.2: Crowdsourced multi-class annotations of perceived user properties for 5,000 Twitter user profiles obtained via crowdsourcing. Attribute class values are sorted by frequency; class A is highlighted in bold, class B.

<table>
<thead>
<tr>
<th>Relation</th>
<th>$\subseteq U$</th>
<th>$N_{uniq}$</th>
<th>$N_{all}$</th>
<th>$T_{total}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retweet</td>
<td>9,751</td>
<td>32,197</td>
<td>48,262</td>
<td>6,345,722</td>
</tr>
<tr>
<td>Mention</td>
<td>9,251</td>
<td>37,199</td>
<td>41,456</td>
<td>7,634,961</td>
</tr>
<tr>
<td>Friend</td>
<td>10,381</td>
<td>43,376</td>
<td>51,316</td>
<td>8,973,783</td>
</tr>
<tr>
<td>TOTAL</td>
<td>10,741</td>
<td>112,772</td>
<td>141,034</td>
<td>24,919,528</td>
</tr>
</tbody>
</table>

Table 8.3: Twitter ego-network sample statistics. $U = 123,513$ nodes with $T = 24,919,528$ tweets, and $E = 141,034$ edges of different types that represent social relations between Twitter users.

Discussed in details for gender and age annotations (D.-P. Nguyen et al., 2014) and personality judgments (Youyou et al., 2015). Some attributes are more difficult to recognize than others e.g., gender vs. income.

For the main analysis we collect a significantly larger sample of Twitter users. For
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

$U = 10,741$ users we sample their neighbors $n \in N^{(u)}$ of different types including:

- friends – $u$ follows $n^{(u)}$ and $n^{(u)}$ follows $u$ (bidirectional following),
- mentions – $u$ mentions $n^{(u)}$ in his tweets e.g., “@HarrysKiss ohhhh!!thankssss!”,
- retweets – $u$ retweets $n^{(u)}$ tweets e.g., “RT @daledaleALE: I wish the answers were simple...”.

In total we sampled $N = 141,034$ neighbors for $U = 10,741$ users; on average 15 neighbors per user, 5 neighbors of each type with their 200 tweets; in total $T = 24,919,528$ tweets as reported in Table 8.3. We report the number of users with at least one neighbor of each type $\subseteq U$ and the number of unique neighbors $N_{uniq}$.

8.1.2 Emotion Data

Similar to other works that bootstrapped noisy data annotated with emotions we rely on the fact that people use the hashtag #sadness or emoticon :-( to indicate that they are sad. For that we collect tweets from the 1% Twitter feed over the last four years with hashtags that correspond to six emotions identified by Ekman (1992): #joy, #anger, #fear, #sadness, #disgust and #surprise. In addition, we compile a synonym list using WordNet-Affect (Valitutti 2004), Google Synonyms and Roget’s Thesaurus.\footnote{Roget’s Thesaurus: http://thesaurus.com/Roget-alpha-index.html} In total we expand our initial emotion hashtag set to 360 emotion hashtags including #joy: 66, #anger: 56, #fear: 78, #sadness: 72, #disgust: 38 and #surprise: 50 and collect more tweets that contain emotion synonym hashtags.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

For example, tweets with emotion-word hashtags include:

- *I got so pissed i broke the entire keyboard on the laptop....we had to buy another keyboard #anger*
- *I love when everything turns out better than you expected. #happy*
- *The fact that today is actually september 1st scares me to death #fear*
- *Monetary issues and petty arguments ruin relationships of all kinds! #sadness*
- *Nooo, tomorrow I have to take the subway TT TT #fear #horror #dismay #disgust*
- *I know you’ll read this when you wake up #surprise*

We prefer Ekman’s emotion classification over others e.g., Plutchik’s because to be able to compare the performance of our predictive models to other systems.

We decide to collect more tweets annotated with emotion hashtags compared to other systems for several reasons. First, we aimed to improve emotion classification by learning from more labeled tweets. Second, we wanted to reduce data annotation biases present in the existing datasets that rely on 6 emotion hashtags exclusively by including tweets annotated with emotion synonym hashtags in our dataset.

We exclude tweets with fewer than three words, filter out non-English tweets (following the strategy used by (S. M. Mohammad & Kiritchenko, 2014) – exclude tweets that contain three or more words not from a Roget’s Thesaurus), remove retweets and tweets where hashtags are not at the end of the tweet. Researchers show that middle-of-tweet hashtags may not be good labels (González-Ibáñez et al., 2011). Finally, we get 28,656 tweets collected using the original six emotion hashtags
and 24,269 tweets collected using 360 emotion synonym hashtags – $T_L = 52,925$ tweets total annotated with emotions. We present the distribution of tweets for each emotion in Table 8.8. Our hashtag emotion dataset is three times larger than the recently released Hashtag Emotion Corpus (S. M. Mohammad & Kiritchenko, 2014) but smaller than a prior bootstrapped corpus (W. Wang et al., 2012). However, we show that we significantly outperform all existing emotion prediction models.

### 8.1.3 External Sentiment Data

Sentiment analysis in social media has been studied a lot recently and became a well established task compared to the emotion prediction task described above (Pang et al., 2002; Pang & Lee, 2008; Go et al., 2009; Pak & Paroubek, 2010; Hassan Saif, Miriam Fernandez & Alani, 2013; Nakov et al., 2013; S. M. Mohammad et al., 2013; Zhu et al., 2014). Thus, it is possible to take advantage of the existing publicly available resources for sentiment classification on Twitter (Hassan Saif, Miriam Fernandez & Alani, 2013). For that we rely on several publicly available sentiment analysis datasets including:

- **Stanford** – [http://help.sentiment140.com](http://help.sentiment140.com) (Go et al., 2009)
- **SemEval-2013** – [http://www.cs.york.ac.uk/semeval-2013/task2/](http://www.cs.york.ac.uk/semeval-2013/task2/) (Nakov et al., 2013)
- **JHU** – [http://www.cs.jhu.edu/~svitlana/](http://www.cs.jhu.edu/~svitlana/) (Volkova et al., 2013a)
Table 8.4: Number of tweets annotated with positive, negative and neutral sentiment recovered from public Twitter sentiment datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>182</td>
<td>177</td>
<td>139</td>
<td>498</td>
</tr>
<tr>
<td>Sanders</td>
<td>466</td>
<td>509</td>
<td>2,145</td>
<td>3,120</td>
</tr>
<tr>
<td>SentiStrength</td>
<td>632</td>
<td>1,388</td>
<td>–</td>
<td>2,020</td>
</tr>
<tr>
<td>SemEval-2013</td>
<td>3,456</td>
<td>1,443</td>
<td>1,915</td>
<td>6,814</td>
</tr>
<tr>
<td>JHU</td>
<td>1,214</td>
<td>705</td>
<td>1,684</td>
<td>3,603</td>
</tr>
<tr>
<td>Debate</td>
<td>658</td>
<td>1,318</td>
<td>477</td>
<td>2,453</td>
</tr>
<tr>
<td>HealthCare</td>
<td>190</td>
<td>358</td>
<td>500</td>
<td>1,048</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>6,798</strong></td>
<td><strong>5,897</strong></td>
<td><strong>6,860</strong></td>
<td><strong>19,555</strong></td>
</tr>
</tbody>
</table>

8.2 Predicting User Emotions, Opinions and Demographics from Tweets

In this section we describe how to build models for user-level (attributes) and tweet-level (emotion and sentiment) predictions. We then discuss how to measure emotion and opinion distributions for Twitter users.

---

As we described in Appendix C, Twitter data sharing policy does not allow sharing the actual tweets (only tweetIDs). However, some profiles became private or deleted over time and their tweets were not available when we tried to download them.
8.2.1 Perceived Attribute Prediction

We assume a set of independent users $U$. A user $u \in U$ is labeled if we know the value of the attribute function $A(u) : U \rightarrow \{a_0, a_1\}$, for example the ethnicity attribute defined as: $A_{ethnicity}(u) = \{\text{African American}; \text{Caucasian}\}$, children attribute $A_{children}(u) = \{\text{Yes, No}\}$ etc.

Similar to log-linear models defined in Eq. 2.3, 6.1 and 6.2 for gender and Eq. 3.1 and 5.1 for political preference classification in Chapters 3–6, we define a set of user-based models $\Phi_A(u)$ for classifying the 10 user attributes presented in Table 8.1. These models are learned from user self-authored content (200 tweets per user $T^{(u)}$) and return the most likely user-level attribute value assignment as described in Chapter 3 and revisited below:

$$\Phi_A(u) = \arg\max_a P(A(u) = a \mid T^{(u)})$$ (8.1)

Our models output the probability (classifier confidence) for every user attribute being one or another class as shown in Figure 8.4 e.g., Male: 0.8, Caucasian: 0.95. In total we classify 10 demographic attributes, 5 personality traits (only two outlined in Table 8.1) and 3 types of controlled impression behavior. However, for the rest of the chapter we focus only on 10 attributes shown in Table 8.1 due to low inter-annotator agreement for the rest of the attributes.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.4: An example of psycho-demographic attribute predictions for a random user.

8.2.2 Emotion and Opinion Detection

We assume a set of independent tweets $T = \{t_i\}$. A tweet is labeled if we know the value of the emotion function $E(t): T \rightarrow \{\text{joy, anger, disgust, fear, surprise, sadness}\}$ and sentiment function $S(t): T \rightarrow \{\text{positive, negative, neutral}\}$.

We define two tweet-based models $\Phi_E(t)$ and $\Phi_S(t)$ for emotion and sentiment classification learned from an independent set of labeled tweets described in the data section. These functions map each tweet to the most likely tweet-level attribute value assignments as shown for the emotion attribute below:

$$\Phi_E(t) = \arg\max_e P(E(t) = e \mid t) \quad (8.2)$$
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

8.2.3 Estimating Emotion & Opinion Distributions

Given a set of tweets $T^{(u)}$ with predicted emotions, we estimate the proportion or normalized frequency of each emotion $e$ per user. The example emotion distribution for a random Twitter user is shown in Figure 8.5a.

![Emotions Distribution](image1)

![Sentiment Distribution](image2)

Figure 8.5: An example of emotion and sentiment distributions for a random user.

Here we introduce user positive emotion score $E^+(u)$ estimated from the normalized distribution of emotions for each user. To estimate $E^+(u)$ we subtract four negative emotions – anger, sadness, fear and disgust from one positive emotion – joy:

$$E^+(u) = e_{\text{joy}} - e_{\text{anger}} - e_{\text{sad}} - e_{\text{disg}} - e_{\text{fear}}.$$ (8.3)

We exclude the ‘surprise’ emotion from the $E^+(u)$ score because it can be both positive and negative and it is hard to evaluate out of context. However, the remaining 5 emotions can be easily disambiguated.

We estimate the proportion or normalized frequency of sentiment expressed by each user. The example sentiment distribution for a random Twitter user is shown in
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.5b. Given the proportion of sentiment per user, we estimate the user positive sentiment score $S^+(u)$. For that we subtract the proportion of negative opinions from positive opinions as shown below:

$$S^+(u) = s_{pos} - s_{neg}. \quad (8.4)$$

We exclude neutral sentiment from $S^+(u)$ because we are only interested in measuring the difference between the most extreme sentiment polarities.

### 8.2.4 Experimental Setup and Results

As shown in Figure 8.3, to perform our analysis we develop two machine learning components. The first component is a user-level demographic classifier $\Phi_A(u)$, which can examine a set of tweets produced by any Twitter user and output a set of predicted demographic traits for that user, including age, education etc. Each demographic classifier relies on features extracted from user content. The second component is a tweet-level emotion and sentiment classifier $\Phi_E(t)$, which can examine any tweet to predict the emotion and sentiment expressed in the tweet. We build models for user attribute, opinion and emotion prediction using the scikit-learn toolkit (Pedregosa et al., 2011).

---

6Scikit-learn toolkit: [http://scikit-learn.org/stable/]
8.2.4.1 Perceived Attribute Prediction Quality

8.2.4.1.1 Cross-Dataset Comparison

Unfortunately, we could not compare our attribute prediction models with all approaches outlined in Tables 2.1 – 2.3. We could not access their datasets to make a direct comparison due to Twitter data sharing policy as we describe in Appendix C.

However, we are able to make a cross dataset comparison for the models learned from subjective perceived annotations and other public datasets used in Chapter 3 (also used in Zamal et al., 2012; Cohen & Ruths, 2013) annotated with gender and political preferences. We could not do the same for the age attribute due to differences in annotations e.g., above or below 25 y.o. vs. 18-23, 23-25 y.o.

In Tables 8.5 and 8.6 we present cross-dataset comparison results for gender and political preference prediction. Accuracy numbers on the diagonals are obtained using 10-fold cross validation. We found that our model learned from perceived annotations yield better accuracy on Burger’s data than a model learned from a gender graph (0.58 vs. 0.47). Moreover, when we classify users in our gender graph using Burger’s data (learned from 71k profiles), our model learned from perceived annotations (5,000 profiles) yields 66% accuracy whereas Burger’s model 72%.

For political preference classification, our model learned from perceived annotations yields (a) higher performance (0.56) than a candidate-centric graph (0.54) when classifies users from a geo-centric graph; (b) higher performance (0.61) than both geo-centric (0.53) and candidate-centric (0.58) graphs when classifies users from an active
## CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Table 8.5: Cross-dataset comparison of accuracy for the gender attribute using log-linear models trained using binary word unigram features. Models trained on perceived attribute annotations were tested on other datasets that had been labeled with gender via user names (Volkova et al., 2014; Zamal et al., 2012) and URL following (Burger et al., 2011).

<table>
<thead>
<tr>
<th>Train → Test</th>
<th>Users</th>
<th>Gender graph, Ch. 3</th>
<th>Perceived, Ch. 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burger et al. (2011)</td>
<td>71,312</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Gender graph, Ch. 3</td>
<td>383</td>
<td>0.47</td>
<td>0.79</td>
</tr>
<tr>
<td>Perceived, Ch. 8</td>
<td>4,998</td>
<td>0.58</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 8.6: Cross-dataset comparison of accuracy for the political preference attribute using log-linear models trained with binary word unigram features. Models trained on perceived attribute annotations were tested on other datasets that had been labeled via following political candidates (the cand-centric graph from Ch. 3, the active graph from Ch. 3 and (Zamal et al., 2012; Cohen & Ruths, 2013)), and self-reports (the geo-centric graph from Ch. 3).

<table>
<thead>
<tr>
<th>Train → Test</th>
<th>Users</th>
<th>Geo-centric graph</th>
<th>Cand-centric graph</th>
<th>Active graph</th>
<th>Perceived annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geo-centric graph</td>
<td>270</td>
<td>0.66</td>
<td>0.47</td>
<td>0.53</td>
<td>0.31</td>
</tr>
<tr>
<td>Cand-centric graph</td>
<td>1,031</td>
<td>0.54</td>
<td>0.76</td>
<td>0.58</td>
<td>0.76</td>
</tr>
<tr>
<td>Active graph</td>
<td>371</td>
<td>0.60</td>
<td>0.63</td>
<td>0.87</td>
<td>0.76</td>
</tr>
<tr>
<td>Perceived</td>
<td>2,498</td>
<td>0.56</td>
<td>0.49</td>
<td>0.61</td>
<td>0.76</td>
</tr>
</tbody>
</table>

For previously unexplored attributes e.g., income, education, optimism etc. we present the ROC AUC numbers\(^7\) obtained using baseline log-linear models with 10-fold cross validation in Figure 8.6.

\(^7\)We have not reported accuracy because classes are imbalanced.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.6: User demographic attribute classification performance (ROC AUC).

Moreover, we demonstrate the top ranked features by the classifier for gender, relationship status, and income attributes in Figure 8.7.

8.2.4.1.2 Perceived vs. True Attribute Annotation Comparison

In addition to our cross-dataset comparison experiments, we attempt to contrast perceived attribute annotations for our 5k users with the “true” (or the best approximation of the ground-truth) labels. We take gender as an example attribute and obtain “true” labels by matching user first names (extracted from user profile metadata) against the US Census name lists. We were able to obtain “true” gender labels for 2,124 users out of 5,000 users in our demographic data described in section 8.1.

CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.7: Word cloud illustration of the most predictive lexical markers (features) for income, gender, and relationship status attributes. Our models for gender prediction resemble the results of other works that analyze language of gender and personality in social media, e.g., [http://wwbp.org/gender-wc.html](http://wwbp.org/gender-wc.html), as well as the rationales presented in Chapter 6.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Train → Test

<table>
<thead>
<tr>
<th></th>
<th>Users</th>
<th>True</th>
<th>Perceived</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>2,124 (44% Male)</td>
<td>0.750</td>
<td>0.760</td>
</tr>
<tr>
<td>Perceived</td>
<td>2,911 (48% Male)</td>
<td>0.767</td>
<td>0.810</td>
</tr>
</tbody>
</table>

Table 8.7: Comparing model accuracy trained on “true” vs. perceived annotations for predicting the gender attribute. We use log-linear models trained using binary word unigram features. The numbers on a diagonal are obtained using 10-fold c.v.

We next perform a set of experiments to measure and compare the power of “true” vs. perceived annotations for the task of predicting user demographics. In particular, we (1) measure the agreement between “true” vs. perceived labels, (2) train models on “true” labels to make predictions for the users annotated with perceived attributes and vise versa, and (3) contrast top-ranked (the most predictive) features for both types of annotations.

We found that the Pearson correlation between “true” and perceived gender labels is 0.72, which indicates a strong agreement between “true” vs. perceived labels, and further confirms the quality of perceived annotations. In Table 8.7 we compare performance for models trained on “true” vs. perceived annotations. We use a sample of users annotated with perceived attributes by excluding users with “true” annotations (to make sure there is no train/test overlap). We found that models trained on “true” gender labels and tested on perceived annotations yield comparable accuracy (0.760) to the models trained on perceived annotations and tested on users with “true” labels (0.767). Moreover, models trained exclusively on perceived attribute annotations using 10-fold c.v. achieve higher accuracy (0.81). These results further confirm that the quality of perceived attribute annotations is acceptable.
Finally, in Figure 8.8 we present word clouds of the top-ranked features identified by our models after training on “true” gender labels vs. perceived annotations. We observe some feature overlap between two models, such as *bro, lebrons* for male users, and *haircut, nails, omgg, baby, cute* for female users.

![Word Clouds](image)

(a) True Male  
(b) True Female  
(c) Perceived Male  
(d) Perceived Female

Figure 8.8: Word cloud illustration of the most predictive lexical markers for gender learned from true vs. perceived annotations. The size of the word reflects the feature weight learned by a classifier.

8.2.4.2 Emotion and Sentiment Prediction Quality

For emotion and opinion classification we train tweet-based classifiers defined in Eq. 8.2 using features extracted from tweets annotated with positive, negative or neu-
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

neutral sentiment $T^S_S$ and six basic emotions $T^S_E$. To get user-level emotion distributions we aggregate predicted emotions over user tweets as described in section 8.2.2.

In addition to lexical features we used for predicting user attributes in section 8.2.4.1, for emotion and sentiment prediction we extract a set of stylistic features including positive and negative emoticons, elongated words, capitalization, repeated punctuation, number of hashtags and take into account the clause-level negation (Pang et al., 2002) as outlined below.

- **Lexical**: binary unigram bag-of-word features (using higher order n-grams or normalized frequency count-based features does not improve prediction results).
- **Stylistic**: elongated words e.g., Yaaay, noooo; capitalization e.g., COOL, MAD; pos/neg emoticons punctuation e.g., !!!!, ????: number of hashtags.
- **Negation**: append a _NEG suffix to every word appearing between a negation and a clause-level punctuation mark (Pang et al. 2002, Das & Chen 2007).
- **Lexicon**: presence and score for unigram features from the Hashtag Emotion Lexicon (S. M. Mohammad & Kiritchenko 2014) and manually created emotion lexicon (S. M. Mohammad & Turney 2013).
- **PoSTags**: part-of-speech tags obtained using Twitter POS tagger.

Adding other linguistic features e.g., higher order n-grams, part-of-speech tags or lexicons did not improve either sentiment or emotion classification performance.

We demonstrate our prediction quality using 10-fold cross validation on our hashtag

\[\text{http://sentiment.christopherpotts.net/tokenizing.html}\]
\[\text{http://sentiment.christopherpotts.net/lingstruc.html}\]
\[\text{http://www.ark.cs.cmu.edu/TweetNLP/}\]
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#anger</td>
<td>457,972</td>
<td>0.72</td>
<td>583</td>
</tr>
<tr>
<td>#disgust</td>
<td></td>
<td>0.64</td>
<td>400</td>
</tr>
<tr>
<td>#fear</td>
<td>11,156</td>
<td>0.74</td>
<td>922</td>
</tr>
<tr>
<td>#joy</td>
<td>567,487</td>
<td>0.68</td>
<td>222</td>
</tr>
<tr>
<td>#sadness</td>
<td>489,831</td>
<td>0.69</td>
<td>493</td>
</tr>
<tr>
<td>#surprise</td>
<td>1,991</td>
<td>0.61</td>
<td>324</td>
</tr>
<tr>
<td>All:</td>
<td>1,991,184</td>
<td>0.67</td>
<td>3,777</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#Emotion</th>
<th>Mohammad et al. (2014)</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>#anger</td>
<td>1,555</td>
<td>4,963</td>
</tr>
<tr>
<td>#disgust</td>
<td>761</td>
<td>12,948</td>
</tr>
<tr>
<td>#fear</td>
<td>2,816</td>
<td>9,097</td>
</tr>
<tr>
<td>#joy</td>
<td>8,240</td>
<td>15,559</td>
</tr>
<tr>
<td>#sadness</td>
<td>3,830</td>
<td>4,232</td>
</tr>
<tr>
<td>#surprise</td>
<td>3849</td>
<td>8,244</td>
</tr>
<tr>
<td>All:</td>
<td>21,051</td>
<td>52,925</td>
</tr>
</tbody>
</table>

Table 8.8: Emotion classification F1 results (one vs. all for each emotion and 6 way for all) using our models compared to other systems. The last column reports F1 results when we combine our data with Mohammad’s ($T = 73,976$ total tweets) and apply 10 fold c.v.

emotion dataset and compare it to other existing datasets annotated with emotions in Table 8.8. Our results significantly outperform the existing approaches for emotion detection on Twitter and are comparable with the state-of-the-art system for Twitter sentiment classification (S. M. Mohammad et al., 2013; Zhu et al., 2014) (evaluated on the official SemEval-2013 test set). Finally, in Figure 8.9, we demonstrate the top ranked features by the classifier for six Ekman’s emotions.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.9: Word cloud illustration of the most predictive lexical markers for joy, sadness, fear, anger, surprise and disgust emotions.
8.3 Correlating User Emotions and Perceived Demographics

8.3.1 Estimating Emotion and Opinion Differences

We first evaluate whether Twitter users with different attribute values e.g., $a_0 =$ Male vs. $a_1 =$ Female express emotions and sentiment on the same or different levels. For that we group emotion $e \in E$ and opinion $s \in S$ distributions for $a_0$ and $a_1$ users as well as $E^+$ and sentiment $S^+$ positive scores calculated using the Eq. 8.3 and 8.4. We apply a non-parametric Mann-Whitney U test (Mann & Whitney, 1947) to test whether two samples come from the same population or not. Our null hypothesis $H_0$ is that $a_0$ emotion and sentiment proportions tend to be distributed similarly to $a_1$ values; $H_a$ is that they tend to be distributed differently.

We then quantitatively measure emotion and opinion differences between $a_0$ and $a_1$ users. For that we estimate the averaged $a_0$ and $a_1$ distributions over emotions and opinions as $\mu_k^{(a_0)} = \frac{\sum_{a_0, e_k} |e_k|}{U}$ and $\mu_k^{(a_1)} = \frac{\sum_{a_1, e_k} |e_k|}{U}$, $e_k \in E$ (similarly $\mu_l^{(a_0)}$ and $\mu_l^{(a_1)}$, $s_l \in S$). We then take the difference between emotion means $\Delta \mu_k = \mu_k^{(a_0)} - \mu_k^{(a_1)}$ and opinion means $\Delta \mu_l = \mu_l^{(a_0)} - \mu_l^{(a_1)}$. 
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

8.3.2 Experimental Results

We now analyze emotional differences between the groups of users with subjectively predicted contrastive attributes $a_0$ and $a_1$, and present the results in Figure 8.10. We outline opinion differences in Table 8.9. Our key findings are discussed in detail below. All the differences in emotion and opinions for users with contrastive demographics are statistically significantly different (p-value < 0.001, using the Mann-Whitney U test).

For brevity, we refer to a user predicted to be male as a male, and a tweet predicted to contain surprise as a simply containing surprise. However, it is important to recall that a major contribution of this work is that these results are based on automatically predicted subjective properties, as compared to ground truth. We argue that while such automatically predicted annotations may not be perfect (as shown using the ROC AUC in Figure 8.10 in addition, subjective biases might be introduced during the annotation process e.g., annotators might have a bias that females use more emotions in their tweets, or users with more positive tweets are satisfied with life and optimists) at the individual user or tweet level, they (1) provide for meaningful analysis when done on the aggregate. Moreover, note that our (2) models could only use the most recent 200 tweets, far less than all the data available to annotators (all the tweets, photos, profile picture etc.), and (3) our training dataset with $U^L = 5,000$ users is disjoint from the dataset of $U = 123,513$ users we use for the main analysis.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

- **Gender** Female users generate more happy as well as sad tweets, while male users produce more surprise and fear tweets. Female users have higher positive emotion scores. Our findings on female users being more emotional online are in line with other work (S. Mohammad & Yang, 2011; Volkova et al., 2013a).

- **Age** Older (above 25 y.o.) users are 7.5% more positive, generate 4% more joy and 4% fewer sad tweets. Younger users produce more disgust, anger and surprise tweets. Our results are in line with the recently explored “aging positivity effect” in social media that states that older people are happier than younger people (Kern, Eichstaedt, Schwartz, Park, et al., 2014).

- **Relationship** Users in a relationship produce 4% more positive emotions, generate 2.5% more joy and 1.4% fewer sad tweets compared to single users. Single users produce more surprise, anger and disgust tweets.

- **Children** Users without children produce 3.5% more sad and 3.6% fewer joy tweets. Users with children have 6.4% higher positive emotion scores. They produce fewer disgust, anger and surprise tweets but more fear tweets.

- **Education** Users with a college degree produce 4.7% more joy tweets and have 8.7% higher positive emotion score. In contrast, users with only high school education generate 4.4% more sad, disgust and anger tweets.

- **Political** Conservative users produce 3% more joy and 1.8% more fear tweets. Liberal users generate 2.6% more sad tweets and have a 7% lower $E^+$. 

174
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.10: Emotion differences $\Delta \mu_k$ among users with contrastive demographics.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR
EMOTIONAL CONTRAST

<table>
<thead>
<tr>
<th>Attribute</th>
<th>$s_{pos}$</th>
<th>$s_{neut}$</th>
<th>$s_{neg}$</th>
<th>$S^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male+, Female–</td>
<td>-3.7</td>
<td>+7.2</td>
<td>-3.5</td>
<td>-0.3</td>
</tr>
<tr>
<td>Below 25+, Above 25–</td>
<td>-1.6</td>
<td>+5.3</td>
<td>+6.9</td>
<td>-8.4</td>
</tr>
<tr>
<td>Single+, Relationship–</td>
<td>+1.1</td>
<td>+0.7</td>
<td>+1.8</td>
<td>-3.0</td>
</tr>
<tr>
<td>Kids+, No kids–</td>
<td>+0.8</td>
<td>+3.3</td>
<td>-4.1</td>
<td>+4.9</td>
</tr>
<tr>
<td>Degree+, School–</td>
<td>+3.3</td>
<td>+4.1</td>
<td>-7.5</td>
<td>+10.8</td>
</tr>
<tr>
<td>≤$35k+$, &gt;$35k–</td>
<td>-3.0</td>
<td>-4.5</td>
<td>+7.5</td>
<td>-10.6</td>
</tr>
<tr>
<td>Liberal+, Conservative–</td>
<td>-1.1</td>
<td>-3.4</td>
<td>+4.5</td>
<td>-5.6</td>
</tr>
<tr>
<td>≤Average+, &gt;Average–</td>
<td>-1.9</td>
<td>-4.5</td>
<td>+6.4</td>
<td>-8.2</td>
</tr>
<tr>
<td>Afr. Amer.+, Caucasian–</td>
<td>-5.3</td>
<td>+1.2</td>
<td>+4.1</td>
<td>-9.4</td>
</tr>
<tr>
<td>Christian+, Unaffiliated–</td>
<td>+1.2</td>
<td>-0.0</td>
<td>-1.2</td>
<td>+2.3</td>
</tr>
<tr>
<td>Optimist+, Pessimist–</td>
<td>+5.3</td>
<td>+2.3</td>
<td>-7.6</td>
<td>+12.8</td>
</tr>
<tr>
<td>Satisfied+, Dissatisfied–</td>
<td>+5.2</td>
<td>+2.6</td>
<td>-7.8</td>
<td>+13.0</td>
</tr>
</tbody>
</table>

Table 8.9: Sentiment differences $\Delta \mu_l$ among users with contrastive demographics. The most pronounced differences are highlighted in bold.

- **Income** Users with higher income (above $35k) tweet 4.9% more joy tweets and have an 8.9% higher positivity score. Users with lower income tweet 4.5% more sad tweets, almost 1% more anger and disgust tweets.

- **Ethnicity** Caucasian users produce 4.2% more joy tweets and have 7.5% higher positive emotion score. African American users generate 2% more sad, 1% more disgust and surprise and about 1% more fear and anger tweets.

- **Optimism** Optimists produce 7% more joy tweets and have 13% higher $E^+$ whereas pessimists generate 3.4% more sad, about 1% more anger and disgust.

- **Life Satisfaction** Users satisfied with life produce 6% more joy and 3% fewer sad tweets, and 11% higher $E^+$.

Sentiment divergence results in Table 8.9 (similar to emotion differences) demon-
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

strate that users who are female, above 25 years old, in a relationship, with kids, have a degree, get over $35k a year, are conservative, with above average intelligence, Caucasian and Christian produce significantly more positive opinions.

8.4 Inferring Perceived Demographics from Emotions and Opinions

In Figure 8.11 we present demographic classification results using emotion and opinion distributions as features. We show that some emotions and opinions are predictive of one attribute value (red), some of an opposite value (blue). For instance, negative sentiment and sadness are predictive for users with no children and users below 25 years old; anger for non Christian users; surprise for single users; neutral sentiment vs. joy, positive sentiment and sadness for gender etc. In addition, we show a dendrogram for attributes (rows) and emotions (columns). It groups data based on row and column similarities using a hierarchical clustering algorithm. We observe that children and age attributes, life satisfaction and optimism are the most similar; negative sentiment and sadness, positive sentiment and joy are quite similar too.

Finally, we compare prediction performance using emotions and opinions vs. lexical features only. We observe that some attributes are more linguistically expressed, and therefore, are better predicted using lexical features e.g., gender ($-0.16$), race ($-0.19$), relationship ($-0.05$), children ($-0.04$), political beliefs ($-0.05$), religion ($-0.05$).
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.11: Correlations when predicting hidden user demographics (rows) from user-user emotion or opinion features (columns). Color represents regression coefficients for each feature. Red stands for male, satisfied, single, non-religious, liberal, no children, younger than 25, annual income over $35k, Caucasian, college degree, and above average intelligence. Blue stands for the opposite attribute values.

However, some attributes are better predicted from emotions and opinions extracted from tweets e.g., age (+0.9), income (+0.05), education (+0.02), optimism (+0.06), and life satisfaction (+0.06). The latter performance improvements for some attributes confirm that the emotions and sentiment expressed in user tweets is more predictive than language and topics used in their communications.
8.5 Correlating User-Environment Emotional Contrast and Demographics

8.5.1 Motivation

People vary in the ways they respond to the emotional tone of their environment in social media. Some people tend to send out messages with a positive emotional tone, while others tend to express more negative emotions such as sadness or fear. Some of us are likely to share peer messages that are angry, whereas others filter out such messages sent by peers. We focus on the problem of predicting a user’s demographic traits by examining the emotions expressed by the user and their network neighbors, shown in Figure 8.2. We first define the user’s emotional tone, the environment’s emotional tone, and the user-environment emotional contrast.

Definition 2. Environment emotional tone is the proportion of tweets with a specific emotion produced by the user’s neighbors. For example, if the majority of tweets sent by the user’s neighbors express joy, that user has a positive environment. In contrast, a user is in a negative environment if most of his neighbors express anger.

Definition 3. User emotional tone is the proportion of tweets with a specific emotion produced by a user. If a user mostly sends sad messages, he generates a sad emotional tone, while a user who sends mostly joyful messages has a joyful tone.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Definition 4. User-environment emotional contrast is a degree to which user emotions differ from the emotions expressed by their environment. We say that users amplify an emotion when they express it more frequently than their environment, and say they dampen an emotion when they express it less frequently.

In this section we analyze how a user’s psycho-demographic traits are predictive of the way they respond to the emotional tone of their environment in a social network. One hypothesis stipulates that the emotional response is a universal human trait, regardless of the specific demographic background (Wierzbicka, 1986; Cuddy et al., 2009). For example, men and women or young and old people should not be different in the way they respond to their emotional environment. An opposite hypothesis is a demographic dependent emotional contrast hypothesis, stipulating that our demographic background is predictive of our emotional contrast with the environment. For example, one might expect users with lower income to express negative emotion even when their environment expresses mostly positive emotions (high degree of emotional contrast), while users with higher income are more likely to express joy even if their environment expresses negative emotions (Kahneman & Deaton, 2010). We provide an empirical analysis based on a large dataset from a Twitter network, supporting the demographic dependent emotional contrast hypothesis.
8.5.2 Estimating User-Environment Emotional Contrast

We perform our user-environment emotional contrast analysis on a set of users $U$ and neighbors $N$, where $N^{(u)}$ are the neighbors of $u$. For each user we define the set of incoming $T^{in}$ and outgoing $T^{out}$ tweets. We then classify $T^{in}$ and $T^{out}$ tweets containing a sentiment $s \in S$ or emotion $e \in E$, e.g. $T^{in}_e$, $T^{out}_e$ and $T^{in}_s$, $T^{out}_s$ where $E \rightarrow \{\text{anger, joy, fear, surprise, disgust, sad}\}$ and $S \rightarrow \{\text{positive, negative, neutral}\}$.

We measure the proportion of incoming and outgoing tweets containing a certain emotion or sentiment e.g., $p^{in}_{sad} = |T^{in}_{sad}|/|T^{in}|$ for a given user. Next, we estimate user-environment emotional contrast using normalized difference between the incoming $p^{in}_e$ and outgoing $p^{out}_e$ proportions for each emotion:\footnote{12Sentiment differences $\Delta s$ are measured similarly.}

$$\Delta e = \frac{p^{out}_e - p^{in}_e}{p^{out}_e + p^{in}_e}, \forall e \in E. \quad (8.5)$$

We estimate user environment emotional tone and user emotional tone from the distributions over the incoming and outgoing sentiments and emotions for a user e.g., $D^{in}_e = \{p^{in}_{joy}, p^{in}_{anger}, \ldots, p^{in}_{fear}\}$ and $D^{out}_e = \{p^{out}_{joy}, p^{out}_{anger}, \ldots, p^{out}_{fear}\}$. We evaluate user environment emotional tone ($D^{in}_e$, $D^{in}_s$ - proportions of incoming emotions and sentiments) on either a combined set of friend, mentioned and retweeted user tweets or independent subsets of their tweets; and user emotional tone ($D^{out}_e$, $D^{out}_s$ - proportions
of outgoing emotions and sentiments) from user tweets.

We measure similarity between user emotional tone and environment emotional tone via Jensen Shannon Divergence (JSD). It is a symmetric and finite KL divergence that measures the difference between two probability distributions e.g., $D_{in}^e$ and $D_{out}^e$ [Plackett 1983]:

$$\text{JSD}(D_{in}^e || D_{out}^e) = \frac{1}{2} I(D_{in}^e || D) + \frac{1}{2} I(D_{out}^e || D), \quad (8.6)$$

where $D = \frac{1}{2} I(D_{in}^e || D_{out}^e)$, $I = \sum_e D_{in}^e \ln \frac{D_{in}^e}{D_{out}^e}$.

Next, we compare emotion and sentiment differences for the groups of users with different attribute values $A = \{a_0; a_1\}$ e.g., $a_0 = \text{Optimist}$ and $a_1 = \text{Pessimist}$ using a non-parametric Mann-Whitney U test as shown below. For example, we measure $\mu_{\Delta e=\text{joy}}^{\text{Optimists}}$ and $\mu_{\Delta e=\text{joy}}^{\text{Pessimists}}$ within the group of users predicted to be Optimists and Pessimists, and estimate whether these two means are significantly different.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Finally, we use logistic regression models to infer twelve attributes for $U = 10,741$ users using different features below.

- outgoing emotional tone $p_{e}^{out}, p_{s}^{out}$ — the overall emotional profile of a user (regardless the emotions projected in his environment);
- user-environment emotional contrast $\Delta e, \Delta s$ — show whether a certain emotion $\Delta e$ or sentiment $\Delta s$ is being amplified or dampened by the user given the emotions he has been exposed to within his social environment;
- lexical features extracted from user content — represent the distribution of word unigrams over the vocabulary.

8.5.3 Experimental Results

We measure user-environment emotional contrast using Eq. 8.5 between users with different demographics shown in Table 8.1. We present our key findings that confirm the demographic dependent emotional contrast hypothesis in Figure 9.3. We show that users of certain demographics dampen ↓ or amplify ↑ certain emotions as discussed in Definition 3. We demonstrate that regardless of demographics, users tend to amplify sadness, disgust, joy and neutral opinions and dampen surprise, fear, anger, positive and negative opinions compared to their environments except some exclusions below. We present more detailed analysis on user-environment emotional contrast for different attribute-emotion combinations and neighbor types in Figures 8.13 – 8.21.
8.5.3.1 User-Environment Emotional Contrast

Users predicted to be older and having kids dampen sadness whereas younger users and user without kids amplify it. It is also known as the aging positivity effect recently picked up in social media (Kern, Eichstaedt, Schwartz, Park, et al., 2014). It states that older people are happier than younger people (Carstensen & Mikels, 2005). Users predicted to be stressed and pessimists dampen joy whereas users that are not stressed and optimists amplify joy tweets. Users predicted to be dissatisfied with life and not excited and easygoing amplify anger tweet compared to their environment whereas users predicted to be satisfied with life and excited dampen anger tweets.

Figure 8.12: User-environment emotional and opinion contrast given predicted perceived demographic traits.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

8.5.3.2 User-Environment Opinion Contrast

Users predicted to be older, with a degree and higher income dampen neutral opinions compared to their environment (otherwise express more subjective opinions) whereas users predicted to be younger, with lower income and high school education amplify neutral opinions. Users predicted to be male and having kids amplify positive opinions whereas female users and users without kids dampen them.

We report similarities between user emotional and environment emotional tones for different groups of Twitter users using Jensen Shannon Divergence defined in the Eq. 8.6. We present the mean JSD values estimated over the sets of users with two contrasting attributes e.g., predicted to be $a_0=$Male vs. $a_1=$Female in Tables 8.10 and 8.11. We estimate user environment emotional tone over different user-neighbor environments e.g., retweet, friend, and all neighborhoods including user mentions. We found that if user environment emotional tones are estimated from mentioned or retweeted neighbors the JSD values are lower compared to the friend neighbors. these numbers may suggest that users are more emotionally similar to the users they mention or retweet than to their friends (users they follow).

We show that user incoming and outgoing sentiment tones $D_{s}^{\text{in}}$ and $D_{s}^{\text{out}}$ are significantly different for the majority of attributes except gender (friend neighborhood), ethnicity (friend and all), optimism and life satisfaction (retweet). The divergences are consistently pronounced across all neighborhoods for income, age, education and children attributes (p-value < 0.01). When the incoming and outgoing emotional
Chapter 8. Learning from Emotions and User-Neighbor Emotional Contrast

Table 8.10: Mean Jensen Shannon Divergences (JSD, displayed as percentages) between incoming $D_{\text{in}}$ and outgoing $D_{\text{out}}$ sentiment distributions for users with different attributes $a_0$ and $a_1$. Mann-Whitney test results for differences between $a_0$ and $a_1$ JSD values are shown in blue and regular font (p-value $\leq 0.01$) or green and italic (p-value $\leq 0.05$).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>$a_0, a_1$</th>
<th>Retweet</th>
<th>Friend</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td>$\geq 35\text{k}, &lt; 35\text{k}$</td>
<td>22.1</td>
<td>19.4</td>
<td>23.7</td>
</tr>
<tr>
<td>Age</td>
<td>$&lt; 25\text{ y.o}, \geq 25\text{ y.o.}$</td>
<td>19.0</td>
<td>22.7</td>
<td>20.2</td>
</tr>
<tr>
<td>Educat.</td>
<td>[School, Degree]</td>
<td>19.4</td>
<td>22.1</td>
<td>21.1</td>
</tr>
<tr>
<td>Relations</td>
<td>[Single, Relat.]</td>
<td>19.6</td>
<td>22.5</td>
<td>21.5</td>
</tr>
<tr>
<td>Children</td>
<td>[Yes, No]</td>
<td>24.2</td>
<td>19.9</td>
<td>28.4</td>
</tr>
<tr>
<td>Intellig.</td>
<td>[$\leq$ Aver., $&gt;$Aver.]</td>
<td>19.9</td>
<td>22.2</td>
<td>21.9</td>
</tr>
<tr>
<td>Gender</td>
<td>[Male, Female]</td>
<td>19.7</td>
<td>20.5</td>
<td>22.0</td>
</tr>
<tr>
<td>Race</td>
<td>[Cauc., Afr.Amer.]</td>
<td>20.5</td>
<td>19.4</td>
<td>21.7</td>
</tr>
<tr>
<td>Optimism</td>
<td>[Pes., Opt.]</td>
<td>19.9</td>
<td>20.3</td>
<td>23.1</td>
</tr>
<tr>
<td>LifeSatis.</td>
<td>[Dissat., Satisf.]</td>
<td>19.4</td>
<td>20.3</td>
<td>21.6</td>
</tr>
<tr>
<td>Political</td>
<td>[Liberal, Conservative]</td>
<td>19.9</td>
<td>21.9</td>
<td>21.8</td>
</tr>
<tr>
<td>Religion</td>
<td>[Christian, Unaffiliated]</td>
<td>20.6</td>
<td>18.9</td>
<td>22.2</td>
</tr>
</tbody>
</table>

Tones $D_{\text{in}}$ and $D_{\text{out}}$ are estimated over all neighbors, they are significantly different for all attributes except gender and income (retweet), age and optimism (friend), and education and life satisfaction (all).

We next measure user-environment emotional contrast using Eq. 8.5 for different demographics outlined in Table 8.1. We say that users dampen emotions from their social environment if users’ incoming emotional tone is not reflected in users’ outgoing emotional tone on the same level $\Delta e < 0$. We say that users amplify emotions from their social environment if users’ outgoing emotional tone is reflected in users’ incoming emotional tone $\Delta e > 0$. We present our key findings on user-environment emotion $\Delta e$ and opinion $\Delta s$ differences in Figures 8.13 – 8.21. We report p-values for MannWhitney test on $\Delta e$ and $\Delta s$ as p-values $\leq 0.01^{***}$, $\leq 0.05^{**}$.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

<table>
<thead>
<tr>
<th>Attribute [(a_0, a_1)]</th>
<th>Retweet</th>
<th>Friend</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children [Yes, No]</td>
<td>20.9</td>
<td>17.8</td>
<td>35.6 33.2 22.6 18.0</td>
</tr>
<tr>
<td>Race [Cauc., Afr.Amer.]</td>
<td>17.2</td>
<td>19.8</td>
<td>32.5 35.2 17.5 20.1</td>
</tr>
<tr>
<td>Relations [Single, Relat.]</td>
<td>18.1 17.9</td>
<td>33.2 34.3 18.2 19.0</td>
<td></td>
</tr>
<tr>
<td>Religion [Christ., Unaffil.]</td>
<td>18.1 17.7</td>
<td>33.8 32.2 18.7 17.5</td>
<td></td>
</tr>
<tr>
<td>Gender [Male, Female]</td>
<td>18.3</td>
<td>17.9</td>
<td>31.6 34.6 18.2 18.5</td>
</tr>
<tr>
<td>Age [&lt; 25 y.o, ≥ 25 y.o.]</td>
<td>17.2 19.9</td>
<td>32.8 34.7 17.0 21.1</td>
<td></td>
</tr>
<tr>
<td>Intellig. [≤Aver., &gt;Aver.]</td>
<td>18.1 17.4</td>
<td>33.8 30.4 18.5 17.4</td>
<td></td>
</tr>
<tr>
<td>Income [≥ $35K, &lt; $35K]</td>
<td>18.7</td>
<td>17.8</td>
<td>33.6 33.3 20.0 17.6</td>
</tr>
<tr>
<td>Educat. [School, Degree]</td>
<td>18.0 18.1</td>
<td>33.9 32.1 18.1 18.9</td>
<td></td>
</tr>
<tr>
<td>Political [Liberal, Conservative]</td>
<td>18.0 18.2</td>
<td>33.6 32.2 18.4 18.2</td>
<td></td>
</tr>
<tr>
<td>Optimism [Pes., Opt.]</td>
<td>18.9</td>
<td>17.9</td>
<td>33.6 33.3 18.6 18.3</td>
</tr>
<tr>
<td>LifeSatis. [Dissat., Satisf.]</td>
<td>18.6 18.0</td>
<td>33.1 33.4 18.5 16.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 8.11: Mean Jensen Shannon Divergences (JSD, displayed as percentages) between the incoming \(D_{in}\) and outgoing \(D_{out}\) emotion distributions for users with different attributes. Mann-Whitney test results for differences between JSD values are shown in blue and regular font (p-value \(\leq 0.01\)) or green and italic (p-value \(\leq 0.05\)).

**Gender:** Female users have a stronger tendency to dampen surprise and fear from their environment but amplify sadness compared to male users, supporting the claim that female users are more emotionally driven than male users in social media (Volkova et al., 2013a). Male users have a stronger tendency to dampen anger compared to female users. Female users tend to dampen significantly more negative opinions from their environment compared to male users.

**Age:** Younger users amplify sadness but older users neither dampen nor amplify sadness from their environment. It is also known as the *aging positivity effect* recently picked up in social media (Kern, Eichstaedt, Schwartz, Park, et al., 2014). It states that older people are happier than younger people (Carstensen & Mikels, 2005). They have a stronger tendency to dampen anger and amplify disgust compared to younger
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.13: User-environment emotional contrast between male and female users. Error bars show standard deviations from the mean (variance estimator from data) for every $e$ and $s$; p-values are shown as $\leq 0.01^{***}$, $\leq 0.05^{**}$ and $\leq 0.1^*$. 

Figure 8.14: User-environment emotional contrast between older (above age 25) and younger (below age 25) users. 

users. Younger users have a stronger tendency to dampen fear and negative sentiment compared to older users. 

Education: Users with a college degree have a weaker tendency to amplify sadness but stronger tendency to amplify disgust from their environment compared to
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Users with high school education. They have a stronger tendency to dampen anger but weaker tendency to dampen fear. Users with high school education are likely to amplify neutral opinions whereas users with a college degree dampen them.

Children: Users with children have a stronger tendency to amplify joy and dampen surprise and fear from their environment compared to users without kids. Users with kids dampen sadness whereas users without kids amplify it. Users without kids tend to amplify positive opinions whereas users with kids dampen them.

Income: Users with higher annual income (over $35k) have a weaker tendency to amplify sadness and have a stronger tendency to amplify disgust, dampen anger and fear from their environment. They tend to dampen neutral opinions whereas users with lower income amplify them.

Relationships: Users in a relationships have a stronger tendency to amplify joy and disgust from their environment but have a weaker tendency to amplify sadness
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.16: User-environment emotional contrast between users with and without children.

Figure 8.17: User-environment emotional contrast between users with higher and lower annual income.

compared to single users. Single users have a stronger tendency to dampen anger, surprise and fear from their environment. Positive and negative opinion differences are less pronounced between single and not single users.

**Optimism:** Optimists amplify joy from their environment whereas pessimists do
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.18: User-environment emotional contrast between single users and users in a relationship.

not. Instead, pessimists have a stronger tendency to amplify sadness and disgust compared to optimists. Optimists tend to dampen fear. Pessimists tend to dampen positive and amplify neutral opinions.

Figure 8.19: Opinion and emotion differences for optimism.

**Life Satisfaction:** User-environment emotional contrast for life satisfaction attribute highly correlates with the optimism attribute. *Users dissatisfied with life* have
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

a weaker tendency to amplify joy but a stronger tendency to amplify sadness↑ and disgust↑ from their environment compared to users satisfied with life. They amplify anger↑ whereas users satisfied with life dampen anger↓. Users satisfied with life have a stronger tendency to dampen fear but weaker tendency to dampen positive and negative opinions.

![Figure 8.20: User-environment emotional contrast between users predicted to be satisfied and dissatisfied with life.](image)

Figure 8.20: User-environment emotional contrast between users predicted to be satisfied and dissatisfied with life.

It is not surprising that people perceived to be optimistic or satisfied with life are people who produce a lot of joyful emotions. One might assume that it is because annotators were likely to label the users as “satisfied” or “dissatisfied” based on whether they posted more positive or more negative tweets. However, we think it is unlikely that participants who examine a profile in order to tag a target person as an optimist or satisfied with life are scanning the friends of that target. Thus, our findings that dampening and amplifying certain emotions (i.e. emotional contrast features) are correlated with user demographics are important findings.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.21: Opinion and emotion contagion differences between African American and Caucasian users.

**Ethnicity:** Caucasian users have a stronger tendency to amplify sadness and disgust whereas African American users have a stronger tendency to amplify joy and dampen disgust. African American users have a stronger tendency to dampen anger and surprise, but a weaker tendency to dampen fear. Caucasian users are likely to dampen more tweets with negative and fewer tweets with positive sentiment.

**Political Preferences:** We did not find significant differences in user-environment emotional contrast between two groups of users with different political orientations.

**Religion:** Christian users have a stronger tendency to amplify joy, but weaker tendency to amplify sadness and disgust; they have a stronger tendency to dampen anger and fear from their environment compared to unaffiliated users.

In addition to our analysis on user-environment emotional contrast and demographics, we discovered which users are more “opinionated” relative to their environment on Twitter. In other words, users in which demographic group amplify less
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR 
EMOTIONAL CONTRAST

neutral but more subjective tweets e.g., positive, negative. As shown in Figures 8.13 to 8.21, male users are significantly more opinionated than female users, users with kids are significantly more opinionated than users without kids, users with a college degree are significantly more opinionated than users with high school education, older users are significantly more opinionated than younger users, users in a relationship are significantly more opinionated than single users, users with higher income are significantly more opinionated than users with lower income, optimists are significantly more opinionated than pessimists, satisfied users are significantly more opinionated than dissatisfied users, users with higher intelligence are significantly more opinionated than users with lower intelligence, and African American users are significantly more opinionated than Caucasian users.

8.5.4 Discussion

We examined the expression of emotions in social media, an issue that has also been the focus of recent work which analyzed emotion contagion using a controlled experiment on Facebook (Coviello et al., 2014). That study had important ethical implications, as it involved manipulating the emotional tone of the messages users viewed in a controlled way. It is not feasible for an arbitrary researcher to reproduce that experiment, as it was carried on the proprietary Facebook network. Further, the significant criticism of the ethical implications of the experimental design of that study (McNeal, 2014) indicates how problematic it is to carry out research on emotions in social networks using a controlled/interventional technique.

Our methodology for studying emotions in social media thus uses an observational method, focusing on Twitter. We collected subjective judgments on a range of previously unexplored latent user properties, and trained machine learning models to
predict those properties for a large sample of Twitter users. We proposed a concrete
quantitative definition of the emotional contrast between users and their network en-
vIRONMENT, based on the emotions emanating from the users versus their neighbors
(friends, retweeted users or mentioned users). We related users’ multidimensional
emotional contrast to their perceived demographic traits.

Our results indicate that various dimensions of a user’s emotional contrast are cor-
related with several demographic traits, and that it is possible to accurately predict
many traits of a user based solely on the emotions that they and their environment
generate. We showed that various demographic traits correlate with the emotional
contrast between users and their environment, supporting the \textit{demographic-dependent
emotional contrast hypothesis}. We also demonstrated that it is possible to accurately
predict many perceived demographic traits of Twitter users based solely on the emo-
tional contrast between them and their friends. This suggests that the way in which
the emotions we radiate differ from those expressed in our environment reveals a lot
about our identity. For example, women tend to dampen negative sentiment while
men amplify it, older users tend to dampen sad emotions while younger users amplify
them and users who are satisfied with life tend to dampen angry emotions while those
dissatisfied with life tend to amplify them.

Moreover, we found that joy, disgust and sad emotions (except for children, age,
gender, income pessimist and life satisfaction attributes) are likely to be amplified
by the users from their environment. In contrast, anger (except for life satisfac-
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

tion attribute), surprise and fear are likely to be dampened by the users from their environment. Latter results contradict earlier reports about anger being the most contagious emotion in a social network (R. Fan et al., 2013). We think that these discrepancies might be due to the cultural differences between Chinese Weibo users and Twitter users from the US and Canada as well as due to the biases present in both data samples e.g., time period when the data sampled that reflects the concept drift issue in social media.

8.5.4.1 Limitations and Caveats

We note that our analysis and methodology have several limitations.

• First, we only study correlations between emotional contrast and demographics. As such we do not make any causal inference regarding these parameters.

• Second, our labels regarding demographic traits of Twitter users were the result of subjective reports obtained using human annotations. Thus, these labels are the impression of people regarding the traits of the profile owner, rather than the true traits. Clearly, some attributes are harder to estimate than others given a user’s profile and tweets. This leads to noisy annotations, which could skew the results of our analysis.

• Additionally, our labels regarding emotions contained in tweets were obtained by bootstrapping emotion hashtags to train an initial model. Again, if these
hashtags are inaccurate, for example due to sarcasm or users using multiple or contradicting hashtags for the same tweet, our models may infer incorrect emotions, which would skew our analysis.

• Finally, we note that we have not restricted the neighbor tweets to include only those that precede user tweets. Rather, we crawled both user and neighbor tweets within a short time frame (less than a week). To measure user-environment emotional contrast we made sure that user and neighbor tweets were produced at the same time.

Despite these limitations, our results in Figure 8.22 and Table 8.8 do indicate higher performance compared to earlier work. Due to the large size of our dataset, we believe our conclusions regarding correlations between emotional contrast and perceived demographics or regarding predicting demographics based on emotional contrast features are correct.

8.6 Inferring Perceived Attributes from User-Neighbor Emotional Contrast

Our findings in previous sections indicate that predicted demographic traits correlate with the emotional contrast between users and their environment in social media. It suggests that the emotional contrast is a multidimensional property of a user. For
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

each of Ekman’s six emotions, one may examine the outgoing tone for that emotion: the proportion of tweets made by that user carrying that emotional tone. Similarly, for each such emotion, one may examine the incoming tone for that emotion: the proportion of tweets produced by user friends. Then, for each emotion one can observe the emotional contrast between the user and the environment by examining the difference between the incoming and outgoing tone for that emotion. So far we examined correlations between different dimensions of user emotional contrast with various demographic traits. We now show that by using all the dimensions of the emotional contrast, we can quite accurately predict many demographic properties of the user. In other words, we show that the emotional contrast of a given user allows us to infer user demographic traits without relying on any additional information.

Figure [8.22] presents the quality of demographic predictions based on different feature sets. These results indicate that most user traits can be quite accurately predicted using solely the emotional tone (EmoSentOut) and emotional contrast (EmoSentDiff) features of the users. That is, given the emotions expressed by a user, and contrasting these with the emotions expressed by user environment, one can accurately infer many interesting properties of the user without using any additional information. Additional lexical features based on the text in user tweets do improve classification accuracy, but the improvement is moderate at best. We note that the emotional features have a strong influence on the prediction quality, resulting in significant absolute ROC AUC improvements over the lexical only fea-
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

Figure 8.22: Psycho-demographic attribute prediction different features including EmoSentOut (user emotional tone), EmoSentDiff (user-environment emotional contrast), and a combination of these two variables with additional lexical features extracted from user content.

The feature set shown in Figure 8.6 for all attributes, including ethnicity (+0.04), gender (+0.05), education (+0.11), income (+0.12), life satisfaction (+0.12), optimism (+0.11), children (+0.08) and age (+0.17). We found that for some attributes e.g., age, life satisfaction, income, optimism user emotional tone and user-environment emotional contrast features are more important than for others e.g., gender and ethnicity.
Figure 8.23 summarizes correlations found between users’ emotional-contrast features and their demographic traits. It indicates that differences between users and their environment in sadness, joy, anger and disgust could be used for predicting whether these users have children or not. Similarly, negative sentiments, joy and disgust emotions could be used to discern between user in a relationship versus single users. Further, negative and neutral opinions, as opposed to joy, fear and surprise emotions can be predictive of users with higher education or intelligence level. While

Figure 8.23: Correlations between user emotional-contrast features and demographic traits when predicting user demographics (rows) solely from user-environment emotional contrast $\Delta e, \Delta s$ (columns). Low values (dark blue) represent regression coefficients for features predictive of one attribute value vs. higher (coral) for an opposite value; light green shows non-discriminate features.
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

these correlations are interesting by themselves, we perform further analysis that indicates how well we can predict various demographic properties of the users based on their emotional tone and user-environment emotional contrast.

In addition, we show a dendrogram for attributes (rows), user-environment emotional contrast (columns). It groups data based on row and column similarities using a hierarchical clustering algorithm. We observe that optimism and life satisfaction attributes, intelligence and education are the most similar; negative sentiment and sadness emotions, anger and disgust emotional contrasts are quite similar too.

8.7 Conclusions

In Chapter 8 we presented an approach to automatically infer demographic attributes of Twitter users and detect their opinions and emotions on a large scale, analyzing 123k users and 25M tweets. We outline our key findings on emotion, opinion, user-environment emotional contrast and attribute correlations below.

- **Detectable emotions and opinions correlate with demographic attributes.** We found that users in a relationship, with children, or with above-average levels of income or education tend to express significantly more joy and less sadness in their tweets. Female users tend to be more opinionated, whereas male users tend to be more neutral. Finally, users who are younger, liberal and with a below average intelligence tend to project more negative opinions and
CHAPTER 8. LEARNING FROM EMOTIONS AND USER-NEIGHBOR EMOTIONAL CONTRAST

emotions.

- **Many perceived user demographic traits correlate with the emotional contrast between users and their neighbors.** We showed that many perceived user demographic traits correlate with the emotional contrast between users and their neighbors. We show that users predicted to be with lower income tend to amplify sad tweets, but older users or those with children dampen them; users satisfied with life dampen angry tweets whereas users dissatisfied with life amplify them; optimists amplify joyful tweets while pessimists dampen them.

- Finally, we showed that it is possible to accurately predict a wide range of perceived demographics for a given user based solely on the emotions expressed by that user and his social environment.

In Chapter 9 we propose static models for inferring latent user properties that rely on user interests extracted from a social network structure rather than on content extracted from user tweets (as we proposed in Chapters 2, 3, 6).
Chapter 9

Learning from User Interests

As we discussed in Chapters 2–8, most works on predicting user attributes rely on user content e.g., messages extracted from user profiles. However, many users do not produce any tweets or tweet extremely sporadically e.g., listeners vs. talkative users\(^1\) as shown in Figure 1.1. Burger et al. (2011) and Volkova et al. (2014) showed that attribute prediction quality dramatically decreases when users have limited content. This has motivated researchers to rely on a social network structure (Pennacchiotti & Popescu, 2011b; Conover, Gonçalves, et al., 2011) or content e.g., language from user neighbors (Zamal et al., 2012) or joint user-neighbor streams (Volkova et al., 2014; Volkova & Van Durme, 2015) to predict user demographics.

However, less work investigated either the relationships between user interests and attributes, or the predictive power of user interests – popular accounts users follow

\(^1\)Only 5% of Twitter users are loudmouth, 50% produce less than 1 tweet per week and 20% have empty accounts.
CHAPTER 9. LEARNING FROM USER INTERESTS

for latent attribute classification. On Twitter people show what they are interested in by following popular news, sport accounts, celebrities, restaurants etc. For example, recent work demonstrated that Facebook likes (Bachrach et al., 2012; Kosinski et al., 2013) and Youtube video likes (Filippova, 2012) can be effectively used to predict demographic attributes.

Unlike the existing works that mainly focus on Twitter interest prediction and recommendation (Bhattacharya et al., 2014; Kapanipathi et al., 2014), in this Chapter we propose to study the relationships and the predictive power of user interests for automatically inferring a variety of latent user attributes. Our main contribution is to qualitatively and quantitatively evaluate user interests to predict user psychodemographics, and compare the results with the state-of-the-art models that rely on user communications.

To analyze the relationships between user interests and predicted perceived psychodemographics on Twitter we collect all followers for 5,000 users from a dataset described in Chapter and determine their interest distributions using the Twitter “who to follow” account hierarchy (Gupta et al., 2013). We then build models to infer latent user properties including previously unexplored user attributes e.g., narcissism, optimism solely from user interests and compare our results with the state-of-the-art approaches that rely on user tweets.

---

2This chapter significantly extends “Using Emotions to Predict User Interest Areas in Online Social Networks”, which was published in the Proceedings of IEEE Conference on Data Science and Advanced Analytics (DSAA) in 2015 and is a joint work with Yoad Lewenberg and Yoram Bachrach.


---

204
CHAPTER 9. LEARNING FROM USER INTERESTS

9.1 Data

9.1.1 Psycho-Demographic Attributes

We use a dataset of 5K users and their 200 tweets annotated with a variety of psycho-demographic attributes as described in section 8.1. In Table 9.1 we report attribute class distributions in the data for a subset of users we were able to determine interests. Thus, the amount of annotated users in Table 9.1 is a subset of users in Table 8.1.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male: 1,797, Female: 2,329</td>
</tr>
<tr>
<td>Age</td>
<td>Below 25: 2,017, Above 25: 1,178</td>
</tr>
<tr>
<td>Political</td>
<td>Conservative: 514, Liberal: 1,577</td>
</tr>
<tr>
<td>Religion</td>
<td>Christian: 2,822, Unaffiliated: 1,305</td>
</tr>
<tr>
<td>Education</td>
<td>High school: 2,723, Degree: 1,403</td>
</tr>
<tr>
<td>Relationship</td>
<td>Single: 2,979, In a relationship 1,147</td>
</tr>
<tr>
<td>Children</td>
<td>Yes: 657, No: 3,471</td>
</tr>
<tr>
<td>Income</td>
<td>Under $35K: 1,480, Over $35K: 2,648</td>
</tr>
<tr>
<td>Optimism</td>
<td>Pessimist: 704, Optimist: 2,247</td>
</tr>
<tr>
<td>Life Satisf</td>
<td>Dissatisfied: 648, Satisfied: 2,484</td>
</tr>
<tr>
<td>Excited</td>
<td>Yes: 2,568, No: 650</td>
</tr>
<tr>
<td>Stressed</td>
<td>Yes: 908, No: 2,332</td>
</tr>
<tr>
<td>Narcissist</td>
<td>Yes: 1,505, No: 1,773</td>
</tr>
</tbody>
</table>

Table 9.1: Annotated attribute class distributions for 4,129 Twitter users for whom we were able to collect followers and determine interests.
CHAPTER 9. LEARNING FROM USER INTERESTS

9.1.2 Interests

To extract user interests, we first crawl all followers for 5,000 users from a dataset described above. We then collect “who to follow” profile IDs categorized into 26 Twitter interest categories as shown in Figure 9.1. For example, there are 48 profiles in the News category including CNN, MSNBC, FoxNews etc. We then match user follower IDs towards “who to follow” profile IDs to determine user interests. We found that 871 of 5,000 users do not follow any accounts from 26 Twitter interest categories. Thus, we remove these profiles from the dataset.

9.2 Methodology

Similarly to the methodological setup in Chapters 3 and 8, we have a set of independent users $U = \{u\}$ for whom we want to predict a variety of latent properties.
We say that we know an attribute for a given user if we know the value of the attribute function \( A(u) \in \{a\} \) e.g., Male vs. Female, Optimist vs. Pessimist, Stressed vs. Not Stressed etc.

![Pie chart](image)

**Figure 9.2:** Distribution of mean interests for male vs. female users.

To predict latent user properties we learn attribute classifiers \( \Phi(u) \) using two feature types: I. Graph-based features learned from user interests \( f^{(i)} \), II. Content-based features learned from user tweets \( f^{(t)} \). Recall from the Eq. 5.1, the model \( \Phi(u) \) is a function that maps a user to the most likely attribute value assignment:

\[
\Phi(u) = \arg\max_a p(A(u) = a \mid f).
\] (9.1)
CHAPTER 9. LEARNING FROM USER INTERESTS

To get interest features $f^{(i)}$ we get the distribution over 26 interest categories for each user. The mean distributions over interests for male vs. female users are shown in Figure 9.2. More detailed analysis on whether the differences in interests between groups of users with contrasting attributes are statistically significant is done using a non-parametric Mann Whitney U test.

To get lexical features $f^{(t)}$, we aggregate 200 tweets per user and treat them as one document (as the majority of previous work), remove urls, punctuation, @mentions and tokens that appears less than 5 times. For each user we build a feature vector using binary word unigram features similar to the models in Chapters 3 – 8. We have also experimented with other features including POS tags, punctuation, LIWC lexicon, network structure and communication behavior features, user metadata etc. However, we did not find any statistically significant improvements over the BOW features.

We use log-linear models $\Phi(u)$ from scikit-learn toolkit (Pedregosa et al., 2011). We prefer logistic regression over reasonable alternatives e.g., SVM or perceptron following previous work on predictive analytics and text classification in social media (Smith, 2004; Volkova et al., 2014).
### Table 9.2: Mean interest values in percentages for the eight most-followed interest categories (followed by more than 1/4 of users in the dataset) averaged within the groups of users with contrastive demographics (e.g., male vs. female). A negative sign (−) represents no statistical significance. Highlighted cells have a p-value ≤ 0.05; in all other cases the p-value was highly significant (p ≤ 0.01).
CHAPTER 9. LEARNING FROM USER INTERESTS

9.3 Experimental Results

We first measure the differences in interests between groups of users with contrastive demographics as shown in Table 9.2. We report mean values for every interest category and determine whether two means for the users with attribute $a_0$ and $a_1$ are statistically significantly different using Mann-Whitney U test. We found that all the differences reported in Table 9.2 are statistically significant with p-value $\leq 0.001$ (or p-value $\leq 0.05$ for some highlighted interest-attribut combinations).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Graph</th>
<th>Content</th>
<th>Content†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.76</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>Age</td>
<td>0.61</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>Political</td>
<td>0.67</td>
<td>0.68</td>
<td>0.72</td>
</tr>
<tr>
<td>Religion</td>
<td>0.57</td>
<td>0.57</td>
<td>0.63</td>
</tr>
<tr>
<td>Education</td>
<td>0.70</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>Relationship</td>
<td>0.57</td>
<td>0.57</td>
<td>0.63</td>
</tr>
<tr>
<td>Children</td>
<td>0.58</td>
<td>0.66</td>
<td>0.72</td>
</tr>
<tr>
<td>Income</td>
<td>0.67</td>
<td>0.68</td>
<td>0.73</td>
</tr>
<tr>
<td>Optimism</td>
<td>0.57</td>
<td>0.68</td>
<td>0.72</td>
</tr>
<tr>
<td>Life Satisf</td>
<td>0.60</td>
<td>0.67</td>
<td>0.72</td>
</tr>
<tr>
<td>Excited</td>
<td>0.55</td>
<td>0.61</td>
<td>0.68</td>
</tr>
<tr>
<td>Stressed</td>
<td>0.58</td>
<td>0.62</td>
<td>0.69</td>
</tr>
<tr>
<td>Narcissist</td>
<td>0.60</td>
<td>0.61</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Table 9.3: Attribute classification results reported as ROC AUC using graph (interest) vs. content (tweet) features.

As we confirm that users with different demographics tend to have different interest distributions e.g., Males are 9% more interested in NBA compared to Females, we proceed with the attribute classification task. In Table 9.3 we report classification performance for each attribute in terms of the area under the ROC curve. ROC AUC
CHAPTER 9. LEARNING FROM USER INTERESTS

is equivalent to the probability of correctly classifying two randomly selected users one from each class e.g., Liberal vs. Conservative. We present the results obtained using Network (interest) vs. Content (tweet) features. We also compare our results with the state-of-the-art performance achieved using Content† features (binary word unigrams extracted from user tweets) reported by Volkova, Van Durme, et al. (2015). Note that our dataset is a subset of the original dataset. Therefore, classification results in columns 3 and 4 obtained using the same BOW features are slightly different.

Our results show that when users do not produce any content but follow some popular accounts on Twitter, we are still able to predict latent user properties with quality comparable to state-of-the-art approaches e.g., for age, education, relationship, income etc. However, for highly verbose attributes – attributes that are being expressed through language rather than through the network structure, such as gender, optimism, excitement or stress content features yield better results compared to network features.

In Figure 9.3 we visualize the predictive power of user interests using the regression coefficients learned by our models that classify psycho-demographic attributes. We show that some interests are predictive of one attribute value (red), some of an opposite value (blue) as discussed for some attributes below.

- **Life satisfaction**: MLB, travel and NASCAR predict users to be satisfied with life; funny, news and music – dissatisfied.
CHAPTER 9. LEARNING FROM USER INTERESTS

- **Optimism**: health and family interests predict users to be optimists; news, science and music – pessimists.

- **Excitement**: PGA and government interests predict users to be not excited and easygoing; science and Twitter – excited and easygoing.

- **Age**: family, technology and business interest categories correlate with older users; funny, gaming and sports with younger users.

- **Political**: PGA, business and NASCAR correlated with conservative users; government, music and gaming with liberal.

- **Religion**: gaming and health interest categories correlate with non-religious users; faith and religion with religious.

- **Education**: technology and staff picks categories correlate with users with a degree; music and television with high school education.

- **Relationship**: family category strongly correlates with users in a relationship; gaming and photography with single users.

- **Gender**: users predicted to be female are interested in fashion and family; male in technology and gaming.

- **Children**: users following family accounts are predicted to have kids; gaming, sports, music funny interests – do not have kids.

- **Income**: books, travel, health and technology correlate with higher income; music, funny, and entertainment with lower income.

- **Narcissism**: science and news correlate with narcissism; MLB, family and
NASCAR have negative correlation with narcissism.

- **Stress**: funny, music and entertainment are predictive of lower stress; technology and business – higher stressed.

In addition, we show a dendrogram for attributes (rows) and interests (columns). It groups data based on row and column similarities using a hierarchical clustering algorithm. We observe that the most similar interests are gaming and sports, funny and music, PGA and business, health and books, Twitter and travel etc. We find that the most similar attributes are income and stress, age and education, life satisfaction and optimism.

### 9.4 Conclusions

In Chapter 9 we studied the relationships between user interests and users’ predicted perceived psycho-demographics in social media. We found significant differences in what users like given their demographics. For example, conservative users are more interested in sports while liberal users preferred television; users with a degree were interested in government and news. Pessimists shared this interest in news, while optimists were interested in health, food and drink. Users with higher income took an interest in travel and technology; those with lower income, music and photography. Narcissists tended to be interested in science and stressed users in business.
CHAPTER 9. LEARNING FROM USER INTERESTS

We then applied user interests as features to automatically infer latent user properties. Our models that rely on interest features offer a strong alternative to existing approaches that rely on user communications. For instance, it is not feasible to make predictions about a user based on his own words and language if he has little or no content associated with his account. However, our models can infer this user properties just from interests or likes extracted from his following behavior.

In the following chapters, we apply our approaches for user demographic inference proposed in Chapters 3 – 9 and demonstrate the benefits of extra-linguistic features (e.g., user personal attributes) for downstream applications such as sentiment classification (Chapter 10) and demographic- and emotion-specific language generation (Chapter 11).
Figure 9.3: Relationship between user interests and psycho-demographic attributes on Twitter. Every attribute (row) has two values, with the first value associated with red color-coding and the second one with blue. For example, for the Political attribute, conservative scores are coded in red and liberal ones in blue. The dendrograms shown on the top and on the left cluster attributes (rows) and interests (columns) by similarity using a hierarchical clustering algorithm.
Chapter 10

Exploring Demographic Language Variations for Sentiment Analysis

Different demographics, e.g., gender or age, can demonstrate substantial variation in their language use, particularly in informal contexts such as social media. In this Chapter\footnote{This chapter presents “Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media”, which was published in the Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP) in 2014 and is a joint work with Theresa Willson and David Yarowsky.} we demonstrate that user demographics or otherwise called extra-linguistic features are very important for a variety of downstream NLP tasks e.g., sentiment analysis. We focus on learning gender differences in the use of subjective language in English, Spanish, and Russian Twitter data, and explore cross-cultural differences in emoticon and hashtag use for male and female users. We show that gender differences in subjective language can effectively be used to improve sentiment classification.


10.1 Motivation

Sociolinguistics and dialectology have been studying the relationships between language at the phonological, lexical and morphosyntactic levels and social identity for decades [Picard 1997, Gefen & Ridings 2005, Macaulay 2006, Tagliamonte 2006]. Recent studies have focused on exploring demographic language variations in personal email communication, blog posts, and public discussions [Boneva et al. 2001, S. Mohammad & Yang 2011, Eisenstein et al. 2010, O’Connor et al. 2010, Bamman et al. 2012]. For instance, they suggest that women are more likely than men to express their concern, support others, and share feelings in personal emails; men write short, precise, more confrontational and challenging emails whereas women write more emotional emails [Cockcroft 2009]. However, one area that remains largely unexplored is the effect of demographic language variation on subjective language use, and whether these differences may be exploited for automatic sentiment analysis. With the growing commercial importance of applications such as personalized recommender systems and targeted advertising [T.-K. Fan & Chang 2009], detecting helpful product review [Ott et al. 2011], tracking sentiment in real time [Resnik 2013], and large-scale passive polling [O’Connor et al. 2010], we believe that sentiment analysis guided by user demographics is an important direction for research.

In this chapter we focus on gender demographics and language in social media to investigate differences in the language used to express opinions in Twitter for three languages: English, Spanish, and Russian. We focus on Twitter data because of
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

its volume, dynamic nature, and diverse population worldwide\(^2\) We find that some words are more or less likely to be positive or negative in context depending on the the gender of the author. For example, the word *weakness* is more likely to be used in a positive way by women (*Chocolate is my weakness!* but in a negative way by men (*Clearly they know our weakness. Aryyyyy*). The Russian word достичь (achieve) is used in a positive way by male users and in a negative way by female users.

Our goals of this work are to (1) explore the gender bias in the use of subjective language in social media, and (2) incorporate this bias into models to improve sentiment analysis for English, Spanish, and Russian. Specifically, in this chapter we:

- investigate multilingual lexical variations in the use of subjective language, and cross-cultural emoticon and hashtag usage on a large scale in Twitter data\(^3\)
- show that gender bias in the use of subjective language can be used to improve sentiment analysis for multiple languages in Twitter.
- demonstrate that simple, binary features representing author gender are insufficient; rather, it is the combination of lexical features, together with set-count features representing gender-dependent sentiment terms that is needed for statistically significant improvements.

To the best of our knowledge, this work is the first to show that incorporating gender leads to significant improvements for sentiment analysis, particularly subjectivity and polarity classification, for multiple languages in social media.

\(^2\)As of July 2014, Twitter has 1 billion registered users from more than 100 countries.
\(^3\)Gender-dependent and independent lexical resources of subjective terms in Twitter for Russian, Spanish and English can be found here: [http://www.cs.jhu.edu/~svitlana/](http://www.cs.jhu.edu/~svitlana/)
10.2 Data

For the experiments in this chapter, we use three sets of data for each language: a large pool of data (800K tweets) labeled for gender but *unlabeled* for sentiment, plus 2K development data and 2K test data labeled for both sentiment and gender. We use the unlabeled data to bootstrap Twitter-specific lexicons and investigate gender differences in the use of subjective language. We use the development data for parameter tuning while bootstrapping, and the test data for sentiment classification.

For English, we download tweets from the corpus created by Burger et al. (2011). This dataset contains 2,958,103 tweets from 184K users, excluding retweets. Retweets are omitted because our focus is on the sentiment of the person tweeting; in retweets, the words originate from a different user. All users in this corpus have gender labels, which Burger et al. automatically extracted from self-reported gender on Facebook or MySpace profiles linked to by the Twitter users.

English tweets are identified using a compression-based language identification (LID) tool (Bergsma et al., 2012). According to LID, there are 1,881,620 (63.6%) English tweets from which we select a random, gender-balanced sample of 0.8M tweets. Burger’s corpus does not include Russian and Spanish data on the same scale as English. Therefore, for Russian and Spanish we construct a new Twitter corpus by downloading tweets from followers of region-specific news and media Twitter feeds.

---

4The LID tool has models trained in 65 languages using a variety of news sources, legislative texts, parliamentary discourse, and other available NLP corpora. In a manually constructed test set using confusable languages in several alphabets, single tweet LID accuracy is 98% (vs. 99.4% in more formally written text).
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

We use LID to identify Russian and Spanish tweets, and remove retweets as before. In this data, gender is labeled automatically based on user first and last name morphology with a precision above 0.98 for all languages.

Sentiment labels for tweets in the development and test sets are obtained using Amazon Mechanical Turk. For each tweet we collect annotations from five workers and use majority vote to determine the final label for the tweet. Snow et al. (2008) show that for a similar task, labeling emotion and valence, on average four non-expert labelers are needed to achieve an expert level of annotation. The task we designed gives a short description of the concepts, examples of labeled tweets, and specific instructions for carrying out the task. Finally, the worker is asked to label 12 tweets\[5\]. Specifically, for each tweet the user is asked:

- whether the tweet is in the desired language - yes or no,
- whether the person tweeting is expressing a sentiment - yes or no,
- if there is a sentiment, what is its polarity - positive, negative, both.

Because this is a subjective and difficult task, we accept all annotations unless it truly seems as if the worker isn’t trying (e.g., gives nearly the same label for every tweet) or has an exceptionally low score on average based on the controls. Generally, we aim to reject only work that is completed incorrectly. Workers who seem to be trying in good faith are paid for their work. The question then is how to determine whose work to use. Each assignment is scored using the controls, and then

\[5\] Of the 12 tweets in each assignment, four are controls used to ensure annotation quality.
each worker is evaluated based on the average score over all tweets that he or she completed. Workers then are ranked, and only data from workers who score above threshold are used. Below are the example English tweets labeled for sentiment:

- **Positive:** Happy Thanksgiving to all my friends and family! Thank you for your friendship!

- **Negative:** Yeah... she’s so not thrilled. BTW date night soon. It haz more boy drama;P

- **Both:** I am ahead on homework!!! I have studied, written essays, taken tests I am tired!

- **Neutral:** @bagussoo u mean u have no bahasa class for a week?

Below are the example Russian tweets labeled for sentiment:

- **Positive:** Как же приятно просто лечь в постель после тяжелого дня... (It is a great pleasure to go to bed after a long day at work...)

- **Negative:** Уважаемый господин Прохоров купите эти выборы! (Dear Mr. Prokhorov just buy the elections!)

- **Both:** Затолкали меня на местном рынке! но зато закупилась подарками для всей семьи :) (It was crowded at the local market! But I got presents for my family:-))

- **Neutral:** Киев очень старый город (Kiev is a very old city).
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

<table>
<thead>
<tr>
<th>Data</th>
<th>Positive</th>
<th>Negative</th>
<th>Both</th>
<th>Neutral</th>
<th>Female ♀</th>
<th>Male ♂</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDev</td>
<td>617</td>
<td>357</td>
<td>202</td>
<td>824</td>
<td>1,176</td>
<td>824</td>
</tr>
<tr>
<td>ETest</td>
<td>596</td>
<td>347</td>
<td>195</td>
<td>862</td>
<td>1,194</td>
<td>806</td>
</tr>
<tr>
<td>SDev</td>
<td>358</td>
<td>354</td>
<td>86</td>
<td>1,202</td>
<td>768</td>
<td>1,232</td>
</tr>
<tr>
<td>STest</td>
<td>317</td>
<td>387</td>
<td>93</td>
<td>1,203</td>
<td>700</td>
<td>1,300</td>
</tr>
<tr>
<td>RDev</td>
<td>452</td>
<td>463</td>
<td>156</td>
<td>929</td>
<td>1,016</td>
<td>984</td>
</tr>
<tr>
<td>RTest</td>
<td>488</td>
<td>380</td>
<td>149</td>
<td>983</td>
<td>910</td>
<td>1,090</td>
</tr>
</tbody>
</table>

Table 10.1: Gender and sentiment label distribution in the development and test sets for all languages.

Table 10.1 gives the distribution of tweets over sentiment and gender labels for the development and test sets for English (EDev, ETest), Spanish (SDev, STest), and Russian (RDev, RTest).

10.3 Subjective Language and Gender

Instead we use large multilingual sentiment lexicons developed specifically for Twitter as described below. Using these lexicons we can begin to explore the relationship between subjective language and gender in the large pool of data labeled for gender but unlabeled for sentiment. We also look at the relationship between gender and the use of different hashtags and emoticons. These can be strong indicators of sentiment in social media, and in fact are sometimes used to create noisy training data for sentiment analysis in Twitter (Pak & Paroubek 2010; Kouloumpis et al. 2011).
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS
FOR SENTIMENT ANALYSIS

10.3.1 Bootstrapping Subjectivity Lexicons

Recent work by Mihalcea et al. (2012) classifies methods for bootstrapping subjectivity lexicons into two types: corpus-based and dictionary-based. Corpus-based methods extract subjectivity lexicons from unlabeled data using different similarity metrics to measure the relatedness between words, e.g., Pointwise Mutual Information (PMI). Corpus-based methods have been used to bootstrap lexicons for English (Turney, 2002) and other languages, including Romanian (Banea, Mihalcea, & Wiebe, 2008) and Japanese (Kaji & Kitsuregawa, 2007).

Dictionary-based methods rely on relations between words in existing lexical resources. For example, Rao and Ravichandran (2009) construct Hindi and French sentiment lexicons using relations in WordNet (Miller, 1995), Rosas et. al. (Perez-Rosas et al., 2012) bootstrap a Spanish lexicon using SentiWordNet (Baccianella et al., 2010) and OpinionFinder, Clematide and Klenner (2010), Chetviorkin et al. (Chetviorkin & Loukachevitch, 2012) and Abdul-Mageed et al. (2011) automatically expand and evaluate German, Russian and Arabic subjective lexicons.

We use the corpus-based, language-independent approach proposed by Volkova et al. (2013b) to bootstrap Twitter-specific subjectivity lexicons. To start, the new lexicon is seeded with terms from the initial lexicon \( L_I \). On each iteration, tweets in the unlabeled data are labeled using the current lexicon. If a tweet contains one or more terms from the lexicon it is marked subjective, otherwise neutral. Tweet polarity

---


\(^7\)OpinionFinder – [www.cs.pitt.edu/mpqa/opinionfinder](http://www.cs.pitt.edu/mpqa/opinionfinder)
is determined in a similar way, but takes into account negation. For every term not in the lexicon with a frequency threshold, the probability of that word appearing in a subjective sentence is calculated. The top $k$ terms with a subjective probability are then added to the lexicon. Bootstrapping continues until there are no more new terms meeting the criteria to add to the lexicon. The parameters are optimized using a grid search on the development data using F-measure for subjectivity classification. In Table 10.2 we report size and term polarity from the initial $L_I$ and the bootstrapped $L_B$ lexicons. Although more sophisticated bootstrapping methods exist, this approach has been shown to be effective for atomically learning subjectivity lexicons in multiple languages on a large scale without any external, rich, lexical resources, e.g., WordNet, or advanced NLP tools, e.g., syntactic parsers (Wiebe, 2000) or information extraction tools (Riloff & Wiebe, 2003).

For English, seed terms for bootstrapping are the strongly subjective terms in the MPQA lexicon (T. Wilson, Wiebe, & Hoffmann, 2005). For Spanish and Russian, the seed terms are obtained by translating the English seed terms using a bi-lingual dictionary, collecting subjectivity judgments from MTurk on the translations, filtering out translations that are not strongly subjective, and expanding the resulting word lists with plurals and inflectional forms.

To verify that bootstrapping does provide a better resource than existing dictionary-expanded lexicons, we compare our Twitter-specific lexicons $L_B$ to the corresponding

---

8We append a NEG suffix to every word appearing between a negation and a clause-level punctuation mark: [http://sentiment.christopherpotts.net/lingstruc.html](http://sentiment.christopherpotts.net/lingstruc.html)
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

<table>
<thead>
<tr>
<th></th>
<th>English $L^E$</th>
<th>Spanish $L^S$</th>
<th>Russian $L^R$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L^E_B$</td>
<td>$L^S_B$</td>
<td>$L^R_B$</td>
</tr>
<tr>
<td>Pos</td>
<td>2.3</td>
<td>2.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Neg</td>
<td>2.8</td>
<td>5.2</td>
<td>2.3</td>
</tr>
<tr>
<td>Total</td>
<td>5.1</td>
<td>8.1</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Table 10.2: Initial $L_I$ and the bootstrapped $L_B$ (highlighted) lexicon term count ($L_I \subset L_B$) with polarity across languages (thousands).

initial lexicons $L_I$ and the existing state-of-the-art subjective lexicons including:

- 8K subj. English terms from SentiWordNet $\chi^E$ (Baccianella et al., 2010);
- 1.5K full strength terms from the Spanish lexicon $\chi^S$ (Perez-Rosas et al., 2012);
- 5K terms from the Russian lexicon $\chi^R$ (Chetviorkin & Loukachevitch, 2012).

For that we apply rule-based subjectivity classification on the test data.\[9\] This subjectivity classifier predicts that a tweet is subjective if it contains at least one or two subjective terms from the lexicon. To make a fair comparison, we automatically expand $\chi^E$ with plurals and inflectional forms, $\chi^S$ with the inflectional forms for verbs, and $\chi^R$ with the inflectional forms for adverbs, adjectives and verbs. We report precision, recall and F-measure results in Table 10.3 and show that our bootstrapped lexicons outperform the corresponding initial lexicons and the external resources.

10.3.2 Lexical Evaluation

With our Twitter-specific sentiment lexicons, we can now investigate how the subjective use of these terms differs depending on gender for our three languages.

\[9\] A similar rule-based approach using terms from the MPQA lexicon is suggested by (Riloff & Wiebe, 2003).
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

\[
\begin{array}{ccc|ccc}
\text{Subj} \geq 1 & \text{Subj} \geq 2 \\
\hline
P & R & F & P & R & F \\
\hline
\chi^E & 0.67 & 0.49 & 0.57 & 0.76 & 0.16 & 0.27 \\
L^E_I & 0.69 & 0.73 & 0.71 & 0.79 & 0.34 & 0.48 \\
L^E_B & 0.64 & 0.91 & 0.75 & 0.7 & 0.74 & 0.72 \\
\chi^S & 0.52 & 0.39 & 0.45 & 0.62 & 0.07 & 0.13 \\
L^S_I & 0.50 & 0.73 & 0.59 & 0.59 & 0.36 & 0.45 \\
L^S_B & 0.44 & 0.91 & 0.59 & 0.51 & 0.71 & 0.59 \\
\chi^R & 0.61 & 0.49 & 0.55 & 0.74 & 0.17 & 0.29 \\
L^R_I & 0.72 & 0.34 & 0.46 & 0.83 & 0.07 & 0.13 \\
L^R_B & 0.64 & 0.58 & 0.61 & 0.74 & 0.23 & 0.35 \\
\end{array}
\]

Table 10.3: Precision, recall and F-measure results for subjectivity classification using the external \( \chi \), initial \( L_I \) and bootstrapped \( L_B \) lexicons for all languages.

Figure 10.1 illustrates what we expect to find. \( \{ F \} \) and \( \{ M \} \) are the sets of subjective terms used by females and males, respectively. We expect that some terms will be used by males, but never by females, and vice-versa. The vast majority, however, will be used by both genders. Within this set of shared terms, many words will show little difference in their subjective use when considering gender, but there will be some words for which gender will have an influence. Of particular interest for our work are words in which the polarity of a term as it is used in context is gender-

\[
t_i \in \{ M \} \wedge t_i \notin \{ F \} \\
t_i \notin \{ M \} \wedge t_i \notin \{ F \}
\]

Figure 10.1: Gender-dependent vs. independent subjectivity terms (+ and - indicates term polarity).
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

Figure 10.2: Distribution of gender-dependent (GDep) and gender-independent (GInd) sentiment terms.

influenced, the extreme case being terms that flip their polarity depending on the gender of the user. Polarity may be different because the concept represented by the term tends to be viewed in a different light depending on gender. There are also words like *weakness* in which a more positive or more negative word sense tends to be used by men or women. In Figure 10.2 we show the distribution of gender-specific and gender-independent terms from the $L_B$ lexicons for all languages.

To identify gender-influenced terms in our lexicons, we start by randomly sampling 400K male and 400K female tweets for each language from the data. Next, for both genders we calculate the probability of term $t_i$ appearing in a tweet with another subjective term (Eq.10.1), and the probability of it appearing with a positive or negative term (Eq.10.2–10.3) from $L_B$.

$$p_{t_i}(subj|g) = \frac{c(t_i, P, g) + c(t_i, N, g)}{c(t_i, g)},$$  \hspace{1cm} (10.1)
We introduce a novel metric $\Delta p^+_t$ to measure polarity change across genders. For
every subjective term $t_i$ we want to maximize the difference\footnote{One can also maximize $\Delta p_{t_i}^- = |p_{t_i}(-|F) - p_{t_i}(-|M)|$.}

$$\Delta p_{t_i}^+ = |p_{t_i}(+|F) - p_{t_i}(+|M)|$$

\[ s.t. \quad 1 - \frac{tf_{t_i}^{subj}(F)}{tf_{t_i}^{subj}(M)} \leq \lambda, \quad tf_{t_i}^{subj}(M) \neq 0, \quad (10.4) \]

where $p(+|F)$ and $p(+|M)$ are probabilities that term $t_i$ is positive for females and males respectively; $tf_{t_i}^{subj}(F)$ and $tf_{t_i}^{subj}(M)$ are corresponding term frequencies (if $tf_{t_i}^{subj}(F) > tf_{t_i}^{subj}(M)$ the fraction is flipped); $\lambda$ is a threshold that controls the level of term frequency similarity\footnote{\(\lambda = 0\) means term frequencies are identical for both genders; \(\lambda \to 1\) indicates increasing gender divergence.}. The terms in which polarity is most strongly gender-influenced are those with $\lambda \to 0$ and $\Delta p_{t_i}^+ \to 1$.

Table 10.4 shows a sample of the most strongly gender-influenced terms from the initial $L_I$ and the bootstrapped $L_B$ lexicons for all languages. A plus (+) means that the term tends to be used positively by women and minus (−) means that the term tends to be used positively by men. For instance, in English we found that *perfecting* is used with negative polarity by male users but with positive polarity by female users; the term *dogfighting* has negative polarity for women but positive polarity for men.
10.3.3 Hashtags

People may also express positive or negative sentiment in their tweets using hashtags. From our balanced samples of 800K tweets for each language, we extracted 611, 879, and 71 unique hashtags for English, Spanish, and Russian, respectively. As we did for terms in the previous section, we evaluated the subjective use of the hashtags. Some of these are clearly expressing sentiment (#horror), while others seem to be topics that people are opinionated about (#baseball, #latingrammy, #spartak).

<table>
<thead>
<tr>
<th>English</th>
<th>$\Delta p^+$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#parenting</td>
<td>+ 0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>#vegas</td>
<td>− 0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>#horror</td>
<td>− 0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>#baseball</td>
<td>− 0.6</td>
<td>0.9</td>
</tr>
<tr>
<td>#wolframalpha</td>
<td>− 0.7</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spanish</th>
<th>$\Delta p^+$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#rafaelnarro</td>
<td>+ 1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>#amores</td>
<td>+ 0.2</td>
<td>1.0</td>
</tr>
<tr>
<td>#britneyspears</td>
<td>+ 0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>#latingrammy</td>
<td>− 0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>#metallica</td>
<td>− 0.5</td>
<td>0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Russian</th>
<th>$\Delta p^+$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#совет (advise)</td>
<td>+ 1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>#спартак (soccer team)</td>
<td>− 0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>#укр (law)</td>
<td>+ 1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>#сны (dreams)</td>
<td>− 1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>#офис (office)</td>
<td>− 1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 10.5: Hashtag examples with opposite polarity across genders for English, Spanish, and Russian.

For example, in Figure 10.3 we show subjective language differences in terms of polarity chance between male and female users on Twitter, as defined in the Eq.10.4.

Table 10.5 gives the hashtags, correlated with subjective language, that are most
strongly gender-influenced. Analogously to \( \Delta p^+ \) values in Table 10.4, a plus (+) means the hashtag is more likely to be used positively by women, and a minus (−) means the hashtag is more likely to be used positively by men. For example, in English we found that male users tend to express positive sentiment in tweets mentioning 

\#baseball, while women tend to be negative about this hashtag. The opposite is true for the hashtag \#parenting.
10.3.4 Emoticons

We investigate how emoticons are used differently by men and women following the work by [Bamman et al., 2012]. For that we rely on the lists of emoticons from Wikipedia\textsuperscript{12} and present the cross-cultural and gender emoticon differences in Figure 10.4. The frequency of each emoticon is given on the right of each language chart, with probability of use by a male user in that language given on the $x$-axis. The top 8 emoticons are the same across languages and sorted by English frequency.

![Figure 10.4: Probability of gender and emoticons for English, Spanish and Russian.](image)

We found that emoticons in English data are used more overall by female users, which is consistent with previous findings in Schnoebelen’s work\textsuperscript{13}. In addition, we

\textsuperscript{12}List of emoticons from Wikipedia \url{http://en.wikipedia.org/wiki/List_of_emoticons}

\textsuperscript{13}Language and emotions \url{www.stanford.edu/~tylers/emotions.shtml}
found that some emoticons like :-) (smile face) and :-o (surprised) are used equally by both genders, at least in Twitter. When comparing English emoticon usage to other languages, there are some similarities, but also some clear differences. In Spanish data, several emoticons are more likely to be used by male than by female users, e.g., :-o (surprised) and :-& (tongue-tied), and the difference in probability of use by males and females is greater for the emoticons, as compared to the same emoticons for English. Interestingly, in Russian Twitter data emoticons tend to be used more or equally by male users rather than female users.

10.4 Experiments

The previous section showed that there are gender differences in the use of subjective language, hashtags, and emoticons in Twitter. We aim leverage these differences to improve subjectivity and polarity classification for the informal, creative and dynamically changing multilingual Twitter data. For that we conduct experiments using gender-independent GInd and gender-dependent GDep features and compare the results to evaluate the influence of gender on sentiment classification.

We experiment with two classification approaches: (I) rule-based classifier which uses only subjective terms from the lexicons designed to verify if the gender differences in subjective language create enough of a signal to influence sentiment classification;
(II) state-of-the-art supervised models which rely on lexical features as well as lexicon set-count features. Moreover, to show that the gender-sentiment signal can be learned by more than one classifier we apply a variety of classifiers implemented in Weka (Hall et al., 2009). For that we do 10-fold cross validation over English, Spanish, and Russian test data (ETest, STest and RTest) labeled with subjectivity (pos, neg, both vs. neut) and polarity (pos vs. neg) as described in Section 8.1.

10.4.1 Models

For the rule-based $GInd_{subj}^{RB}$ classifier, tweets are labeled as subjective or neutral as follows:

$$GInd_{subj}^{RB} = \begin{cases} 1 & \text{if } \vec{w} \cdot \vec{f} \geq 0.5, \\ 0 & \text{otherwise} \end{cases}$$ (10.5)

where $\vec{w} \cdot \vec{f}$ stands for weighted set features, e.g., terms from $L_I$ only, emoticons $E$, or different part-of-speech tags (POS) from $L_B$ weighted using $w = p(subj) = p(subj|M) + p(subj|F)$ subjectivity score as shown in Eq. 10.1. We experiment with the POS tags to show the contribution of each POS to sentiment classification.

Similarly, for the rule-based $GInd_{pol}^{RB}$ classifier, tweets are labeled as positive or negative:

\footnote{A set-count feature is the number of instances from a set of terms that appears in a tweet.}

\footnote{We also experimented with repeated punctuation (!!, ??) and letters (nooo, reealy), which are often used in sentiment classification in social media. However, we found these features to be noisy and adding them decreased performance.}

234
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

\[
GInd_{\text{pol}}^{RB} = \begin{cases} 
1 & \text{if } \vec{w}^+ \cdot \vec{f}^+ \geq \vec{w}^- \cdot \vec{f}^-, \\
0 & \text{otherwise}
\end{cases}
\]  
\tag{10.6}

where \(\vec{f}^+, \vec{f}^-\) are feature sets that include only positive and negative features from \(L_I\) or \(L_B\); \(w^+\) and \(w^-\) are positive and negative polarity scores estimated using Eq. 10.2 – 10.3 such as: 
\(w^+ = p(+|M) + p(+|F)\) and \(w^- = p(-|M) + p(-|F)\).

The gender-dependent rule-based classifiers are defined in a similar way. Specifically, \(\vec{f}\) is replaced by \(\vec{f}^M\) and \(\vec{f}^F\) in Eq. 10.5 and \(\vec{f}^-, \vec{f}^+\) are replaced by \(\vec{f}^{M-}, \vec{f}^{F-}\) and \(\vec{f}^{M+}, \vec{f}^{F+}\) respectively in Eq. 10.6. We learn subjectivity \(\vec{s}\) and polarity \(\vec{p}\) score vectors using Eq. 10.1-10.3. The difference between \(GInd\) and \(GDep\) models is that \(GInd\) scores \(\vec{w}, \vec{w}^+\) and \(\vec{w}^-\) are not conditioned on gender.

For gender-independent classification we build feature vectors using lexical features \(V\) represented as term frequencies, together with the lexicon set-count features:

\[
\vec{f}^{GInd}_{\text{subj}} = [L_I, L_B, E, V];
\]
\[
\vec{f}^{GInd}_{\text{pol}} = [L_I^+, L_B^+, E^+, L_I^-, L_B^-, E^-, V].
\]

Finally, for gender-dependent models, we try different feature combinations. (A) We extract set-count features for gender-dependent terms from \(L_I, L_B,\) and \(E\) jointly:
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS
FOR SENTIMENT ANALYSIS

\[ \vec{f}_{subj}^{GDep - J} = [L_I^M, L_B^M, E^M, L_I^F, L_B^F, E^F, V]; \]
\[ \vec{f}_{pol}^{Dep - J} = [L_I^{M+}, L_B^{M+}, E^{M+}, L_I^{F+}, L_B^{F+}, E^{F+} \]
\[ L_I^{M-}, L_B^{M-}, E^{M-}, L_I^{F-}, L_B^{F-}, E^{F-}, V]. \]

(B) We extract disjoint (prefixed) gender-specific features (in addition to lexical features \( V \)) by relying only on female set-count features when classifying female tweets; and only male set-count features for male tweets. We refer to the joint features as \( GInd - J \) and \( GDep - J \), and to the disjoint features \( GInd - D \) and \( GDep - D \).

10.4.2 Results

Figures [10.5a] and [10.5b] show performance improvements for subjectivity and polarity classification under the rule-based approach when taking into account gender. The left figure shows precision-recall curves for subjective vs. neutral classification, and the middle figure shows precision-recall curves for pos vs. neg classification.

We measure performance starting with features from \( L_I \), and then incrementally add emoticon features \( E \) and features from \( L_B \) (one POS at a time to show the contribution of each POS for sentiment classification).\(^{17}\) This experiment shows that there is a clear improvement for the models parameterized with gender.

\(^{17}\)Part-of-speech tags obtained using the Twitter POSTagger (Owoputi et al. 2013).
For the supervised models we experiment with a variety of learners for English to show that gender differences in subjective language improve sentiment classification for many learning algorithms. We present the results in Figure 10.6. For subjectivity classification, Support Vector Machines (SVM), Naive Bayes (NB) and Bayesian Logistic Regression (BLR) achieve the best results, with improvements in F-measure ranging from 0.5 - 5%. The polarity classifiers overall achieve much higher scores, with improvements for GDep features ranging from 1-2%. BLR with Gaussian prior is the top scorer for polarity classification with an F-measure of 82%. We test our results for statistical significance using McNemar’s Chi-squared test (p-value < 0.01) as suggested by (Dietterich 1998). Only three classifiers, J48, AdaBoostM1 (AB) and Random Forest (RF) do not always show significant improvements for GDep features over GInd features. However, for the majority of classifiers, GDep models outperform GInd models for both tasks, demonstrating the robustness of GDep features for sentiment analysis.

In Table 10.6 we report results for subjectivity and polarity classification using the best performing classifiers (as shown in Figure 10.5c): 


- SVM model with radial-based kernel for GInd − D and GDep − D models. We use LibSVM implementation ([EL-Manzalawy & Honavar 2005]). Each ∆R(%) row shows the relative percent improvements in terms of precision P,
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

<table>
<thead>
<tr>
<th>P</th>
<th>R</th>
<th>F</th>
<th>A</th>
<th>A\text{rand}</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>A</th>
<th>A\text{rand}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English subj vs. neut</strong> p(subj)=0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>English pos vs. neg</strong> p(pos)=0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GInd\text{LR}</td>
<td>0.62</td>
<td>0.58</td>
<td>0.60</td>
<td>0.66</td>
<td>-</td>
<td>0.78</td>
<td>0.83</td>
<td>0.80</td>
<td>0.71</td>
</tr>
<tr>
<td>GDep−J</td>
<td>0.64</td>
<td>0.62</td>
<td>0.63</td>
<td>0.68</td>
<td>0.66</td>
<td>0.80</td>
<td>0.83</td>
<td>0.82</td>
<td>0.73</td>
</tr>
<tr>
<td>ΔR, %</td>
<td>+3.23</td>
<td>+6.90</td>
<td>+5.00</td>
<td>+3.03</td>
<td>+3.03</td>
<td>+2.56</td>
<td>0.00</td>
<td>+2.50</td>
<td>+2.82</td>
</tr>
<tr>
<td>GInd\text{SVM}</td>
<td>0.66</td>
<td>0.70</td>
<td>0.68</td>
<td>0.72</td>
<td>-</td>
<td>0.79</td>
<td>0.86</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td>GDep−D</td>
<td>0.66</td>
<td>0.71</td>
<td>0.68</td>
<td>0.72</td>
<td>0.70</td>
<td>0.80</td>
<td>0.87</td>
<td>0.83</td>
<td>0.78</td>
</tr>
<tr>
<td>ΔR, %</td>
<td>-0.45</td>
<td>+0.71</td>
<td>0.00</td>
<td>-0.14</td>
<td>2.85↓</td>
<td>+0.38</td>
<td>+0.23</td>
<td>+0.24</td>
<td>+0.41</td>
</tr>
<tr>
<td><strong>Spanish subj vs. neut</strong> p(subj)=0.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Spanish pos vs. neg</strong> p(pos)=0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GInd\text{LR}</td>
<td>0.67</td>
<td>0.71</td>
<td>0.68</td>
<td>0.61</td>
<td>-</td>
<td>0.71</td>
<td>0.63</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>GDep−J</td>
<td>0.67</td>
<td>0.72</td>
<td>0.69</td>
<td>0.62</td>
<td>0.61</td>
<td>0.72</td>
<td>0.65</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td>ΔR, %</td>
<td>0.00</td>
<td>+1.40</td>
<td>+0.58</td>
<td>+0.73</td>
<td>1.64↓</td>
<td>+2.53</td>
<td>+3.17</td>
<td>+1.49</td>
<td>0.00</td>
</tr>
<tr>
<td>GInd\text{SVM}</td>
<td>0.68</td>
<td>0.79</td>
<td>0.73</td>
<td>0.65</td>
<td>-</td>
<td>0.66</td>
<td>0.65</td>
<td>0.65</td>
<td>0.69</td>
</tr>
<tr>
<td>GDep−D</td>
<td>0.68</td>
<td>0.79</td>
<td>0.73</td>
<td>0.66</td>
<td>0.65</td>
<td>0.68</td>
<td>0.67</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>ΔR, %</td>
<td>+0.35</td>
<td>+0.21</td>
<td>+0.26</td>
<td>+0.54</td>
<td>1.54↓</td>
<td>+2.43</td>
<td>+2.44</td>
<td>+2.51</td>
<td>+2.08</td>
</tr>
<tr>
<td><strong>Russian subj vs. neut</strong> p(subj)=0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Russian pos vs. neg</strong> p(pos)=0.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GInd\text{LR}</td>
<td>0.66</td>
<td>0.68</td>
<td>0.67</td>
<td>0.67</td>
<td>-</td>
<td>0.66</td>
<td>0.72</td>
<td>0.69</td>
<td>0.62</td>
</tr>
<tr>
<td>GDep−J</td>
<td>0.66</td>
<td>0.69</td>
<td>0.68</td>
<td>0.67</td>
<td>0.66</td>
<td>0.68</td>
<td>0.73</td>
<td>0.70</td>
<td>0.64</td>
</tr>
<tr>
<td>ΔR, %</td>
<td>0.00</td>
<td>+1.47</td>
<td>+0.75</td>
<td>0.00</td>
<td>1.51↓</td>
<td>+3.03</td>
<td>+1.39</td>
<td>+1.45</td>
<td>+3.23</td>
</tr>
<tr>
<td>GInd\text{SVM}</td>
<td>0.67</td>
<td>0.75</td>
<td>0.71</td>
<td>0.70</td>
<td>-</td>
<td>0.64</td>
<td>0.73</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td>GDep−D</td>
<td>0.67</td>
<td>0.76</td>
<td>0.71</td>
<td>0.70</td>
<td>0.69</td>
<td>0.65</td>
<td>0.74</td>
<td>0.69</td>
<td>0.63</td>
</tr>
<tr>
<td>ΔR, %</td>
<td>-0.30</td>
<td>+1.46</td>
<td>+0.56</td>
<td>+0.14</td>
<td>1.44↓</td>
<td>+0.93</td>
<td>+1.92</td>
<td>+1.46</td>
<td>+1.49</td>
</tr>
</tbody>
</table>

Table 10.6: Sentiment classification results obtained using gender-dependent and gender-independent joint and disjoint features in Logistic Regression (LR) and SVM models. Our results show that differences in subjective language across genders can be exploited to improve sentiment analysis, not only for English but for multiple languages. For Spanish and Russian results are lower for subjectivity classification, we suspect, because lexical features $V$ are already inflected for gender and set-count features are down-weighted by the classifier. For polarity classification, on the other hand, gender-dependent features provide consistent, significant improvements (1.5-2.5%) across all languages. To summarize, in Figure 10.7 we show statistically significant relative recall $R$, F-measure $F$ and accuracy $A$ for $GDep$ compared to $GInd$ models.
accuracy improvement over gender-independent baseline \( G_{Ind} \), as well as gender-dependent baseline with randomly shuffled male and female features \( G_{Dep} (Rand) \), 0% and 1.49% for Russian, 0.73% and 2.08% for Spanish, and 3.03% and 2.82% for English for polarity and subjectivity classification.

As a reality check, Table 10.6 also reports accuracies (in \( A_{rand} \) columns) for experiments that use random permutations of male and female subjective terms, which are then encoded as gender-dependent set-count features as before. We found that all gender-dependent models, \( G_{Dep} − J \) and \( G_{Dep} − D \), outperformed their random equivalents for both subjectivity and polarity classification (as reflected by relative accuracy decrease \( ↓ \) for \( A_{rand} \) compared to \( A \)). These results further confirm the existence of gender bias in subjective language for any of our three languages and its importance for sentiment analysis.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Spanish</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P )</td>
<td>0.73</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>( R )</td>
<td>0.93</td>
<td>0.63</td>
<td>0.74</td>
</tr>
<tr>
<td>(a)</td>
<td>0.72</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>( P )</td>
<td>0.94</td>
<td>0.64</td>
<td>0.74</td>
</tr>
<tr>
<td>( R )</td>
<td>0.71</td>
<td>0.63</td>
<td>0.72</td>
</tr>
<tr>
<td>(c)</td>
<td>0.78</td>
<td>0.71</td>
<td>0.66</td>
</tr>
<tr>
<td>( P )</td>
<td>0.83</td>
<td>0.62</td>
<td>0.76</td>
</tr>
<tr>
<td>( R )</td>
<td>0.93</td>
<td>0.65</td>
<td>0.76</td>
</tr>
<tr>
<td>(d)</td>
<td>0.69</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>( P )</td>
<td>0.93</td>
<td>0.62</td>
<td>0.76</td>
</tr>
<tr>
<td>( R )</td>
<td>0.72</td>
<td>0.65</td>
<td>0.73</td>
</tr>
<tr>
<td>(e)</td>
<td>0.80</td>
<td>0.72</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Table 10.7: Precision (\( P \)) and recall (\( R \)) results for polarity classification, encoding gender as a binary feature vs. gender-dependent features \( G_{Dep} − J \). See text for row definitions (a) through (e).

Finally, we check whether encoding gender as a binary feature would be sufficient to improve sentiment classification. For that we encode features such as: (a) unigram term frequencies \( V \), (b) term frequencies and gender binary \( V + GBin \), (c)
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

gender-independent $G{Ind}$, (d) gender-independent and gender binary $GBin + G{Ind}$, and (e) gender-dependent $G{Dep} – J$. We train logistic-regression model for polarity classification and report precision and recall results in Table 10.7. We observe that including gender as a binary feature does not yield significant improvements compared to $G{Dep} – J$ for all three languages.

10.5 Conclusions

In Chapter 10 we analyzed substantial and interesting differences in subjective language between male and female users on Twitter, including hashtag and emoticon usage. We showed that incorporating user demographics (e.g., gender) as a model component can improve subjectivity and polarity classification for the three languages in our study: English (2.5% and 5%, respectively), Spanish (1.5% and 1%) and Russian (1.5% and 1%). Note that the most recent follow-up work has demonstrated similar results (Hovy, 2015).

In Chapter 11 we propose to model attribute-affect topic variations in social media using affects expressed as sentiments or emotions in user tweets and latent user attributes. For example, we can consider relations between attribute-affect combinations such as positive sentiment and gender, political preference and disgust, or life satisfaction and joy.
(a) Rule-based subjectivity

(b) Rule-based polarity

Figure 10.5: Rule-based (RB) sentiment classification results for English. $L_I$ is the initial lexicon, $E$ is emoticons. $A, R, V, N$ are adjectives, adverbs, verbs, and nouns from $L_B$. 
Figure 10.6: Subjectivity and polarity classification results using supervised learning (SL) for English.
CHAPTER 10. EXPLORING DEMOGRAPHIC LANGUAGE VARIATIONS FOR SENTIMENT ANALYSIS

Figure 10.7: Relative accuracy (y-axis) improvements for subjectivity (subjective vs. neutral tweets) and polarity (positive vs. negative) classification results across languages.
Chapter 11

Modeling Attribute-Affect Topic Variations in Social Media

In Chapters 3–9 we showed that user properties such as attributes and emotions can be effectively predicted from texts published in social media. However, there has been limited work investigating the relationships between emotions and attributes in social media, as we discussed in section 7.2 of Chapter 7. The majority of the existing methods for relating language, emotions and attributes are discriminative (including our techniques presented in Chapter 8). In this chapter, we propose a departure from these previous methods, instead using generative approaches to model attribute and affect-specific language among social media users.

Other researchers have briefly explored generative mixture models to infer latent user attributes and study lexical variations in social media for gender, ethnicity
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN
SOCIAL MEDIA

and geo-location (Eisenstein et al. 2010; O’Connor et al. 2010; Rao et al. 2011).

Topic models have been used to model sentiment and language jointly (Mei et al.,
2007; Titov & McDonald 2008; F. Li et al. 2010). In addition, there are generative
approaches that model multi-dimensional aspects of text including cross-collection
mixture models (Zhai et al. 2004; Griffiths & Steyvers 2004; Paul & Girju 2009)
and multi-dimensional text models (Paul & Girju 2010; Paul & Dredze 2012, 2013).

However, none of the existing approaches have modeled (a) fine-grained emotions and
(b) user attributes and affect in combination, which is the main contribution of this
work.

In this chapter we present attribute-affect mixture models designed to discover
topics in social media texts produced by users with varied demographics and asso-
ciated with specific emotional or opinionated tones (affect), and to find differences
among them. These models are capable of discovering topics associated with fear for
users perceived to be females vs. males, positive topics for users perceived to be older
or topics associated with anger for users perceived to be dissatisfied with life. We
believe there are a number of applications for using attribute-affect mixture models
trained on language from social media. One application that has emerged recently is
virtual personal assistants (aka personalized conversational agents), e.g., Microsoft’s
Cortana, Google’s Ok, Google interface, or Apple’s Siri. These applications could
potentially be adapted with little difficulty to respond to a user’s attributes or mood.
11.1 Basic Topic Modeling

Naive Bayes (NB) is the basic generative model which assumes one topic per document and associates each topic with a distribution over words \([\text{Mitchell}, 1997]\). This model assumes that each document is generated by (a) randomly choosing a category and (b) randomly choosing words from that category’s word distribution. It is similar to a topic model since each category represents a topic. Probabilistic Latent Semantic Indexing (pLSI) is designed to cluster semantically related words into topics and, unlike NB, assumes each document to be a mixture of topics \([\text{Hofmann}, 1999]\). The probability of seeing the \(i\)th word in a document \(d\) where \(z\) represents the topic is defined as:

\[
P(w_i \mid d) = \sum_{z \in Z} P(w_i \mid z)P(z \mid d).
\]

Unlike Naive Bayes, pLSI does not allow labeling of previously unseen documents and suffers from overfitting.

Latent Dirichlet Allocation (LDA) deals with the issue of labeling new documents and relies on Dirichlet priors \(\alpha\) and \(\beta\) to reduce overfitting \([\text{Blei, Ng, & Jordan}, 2003]\). Under this model each document is generated as follows:

- Draw a multinomial distribution over words \(\phi_z \sim \text{Dir}(\beta)\) for each topic \(z\).
- For each document \(d\) draw a topic distribution \(\theta(d) \sim \text{Dir}(\alpha)\).
- For each word \(w_i \in d\) :
– Sample a topic $z$ from $\theta^{(d)}$.
– Sample a word $w_i$ from $\phi_z$.

As we previously discussed in the related work section, topic modeling has been successfully applied to a variety of tasks and domains. However, it has been sparsely used for (a) latent user attribute prediction, (b) fine-grained emotion and sentiment inference, and (c) in the social media domain.

### 11.2 Data

For the experiments in this chapter we use a dataset from Chapter 8, section 8.1, as shown in Table 8.1. This dataset comprises 5,000 user profiles annotated with perceived attributes via crowdsourcing, together with the users’ 1M tweets (200 tweets

<table>
<thead>
<tr>
<th>Affect</th>
<th>Attribute</th>
<th>Tweets per User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>Degree</td>
<td>RT @AJandAshley: If You Could Get Away With It, Would You Punch Your Boss In the Face</td>
</tr>
<tr>
<td>Anger</td>
<td>High School</td>
<td>People in general are pissing me off I can’t stand a moody person it irritates me. Like how long does it take for a Mind to develop ??</td>
</tr>
<tr>
<td>Positive</td>
<td>Female</td>
<td>RT @gregg_yolanda: So ready to find out if I need to shop for pink or blue #excited @gregg_yolanda pink pink pink!!!</td>
</tr>
<tr>
<td>Positive</td>
<td>Male</td>
<td>2 job interviews tomorrow whoo hoo!!! Martinez is a fool hahah. He’s gonna get dropped</td>
</tr>
<tr>
<td>Sadness</td>
<td>Above 25 y.o.</td>
<td>Looking at my lawn this AM, used #Scotts plus 2 on it 2 weeks ago. Weeds are going well. Like using nothing. Wasted money.</td>
</tr>
<tr>
<td>Sadness</td>
<td>Below 25 y.o.</td>
<td>@kidrauhlstates @iBieberThought I just want to be someone’s friend. I’m sorry I ever reached out to him</td>
</tr>
</tbody>
</table>

Table 11.1: A set of example tweets annotated with perceived attributes and affect.
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

per user) annotated with perceived six Ekman’s emotions and three types of sentiments. We present an example sets of tweets annotated with both perceived attributes and affect in Table 11.1.

11.3 Models

We propose two models - a baseline Attribute-Affect LDA (AALDA) model, and an Attribute-Affect Mixture (AAM) model - to study exical variations in social media associated with different sentiments. For example, we might consider connections between positive sentiment and emotions like anger or joy in tweets authored by users with differing gender, age, or education levels.

11.3.1 Baseline Attribute-Affect LDA Model

We first propose a variation of a basic LDA model for comparing lexical variations in social media associated with different sentiments and emotions (e.g., positive vs. negative) authored by users with contrastive demographics (e.g., male vs. female). Under this model each topic is associated with the unique word distribution that is specific to attribute-sentiment or attribute-emotion in a user-based collection of tweets. To generate words under this model, we first sample both attribute and sentiment or emotion values (which are observable in the data); we then draw a topic $z$ from the topic’s attribute-sentiment specific distribution. In Figure 11.1 we
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

Figure 11.1: A diagram of the Baseline Attribute-Affect LDA Model (AALDA). Observed variables include Words $w_i \sim \text{Mult}(\phi_{a,s})$, Sentiments $s \in \{\text{Positive, Negative}\}$, and Attributes $a \in \{a_0, a_1\}$. Model hyperparameters (priors): $\alpha, \beta$. Marginalized variables: $\theta, \phi_{a,s}$. Sampled variables: topics $z$. $T$ is the number user-specific sets of tweets, $N_t$ is the length of a set of tweets per user, $K$ is the number of topics, $S$ is the number of affect values and $A$ is the number of attribute values.

present the model; define observed, marginalized and sampled variables; and outline a generative story.

The generative process is thus:

I. Draw attribute-sentiment specific multinomial word distributions

$$\phi_{a,s} \sim \text{Dirichlet}(\beta)$$

for each topic $z_k$ given a specific attribute $a$ and sentiment $s$ values.

II. For each set of tweets $T$ per user:

Choose a value for attribute $a^{(t)}$ and sentiment $s^{(t)}$ and draw a topic mixture $\theta^{(t)} \sim \text{Dirichlet}(\alpha)$.

For each word $w_i \in t$:

- Sample a topic $z_k \sim \text{Mult}(\theta^{(t)})$. 

249
– Sample a word $w_i \sim \text{Mult}(\phi_{a,s})$.

The uncollapsed joint likelihood of the data is defined as:

$$P(z, a, s, w, \theta, \phi | \alpha, \beta) = P(w | z, a, s, \phi) P(\phi | \beta) P(z | \theta) P(\theta | \alpha). \quad (11.1)$$

After integrating out the $\phi, \theta$ collapsed joint likelihood of the data is defined as:

$$P(z, a, s, w | \alpha, \beta) = P(w | z, a, s, \phi) P(z | \alpha) \frac{\Gamma(\sum_k \alpha) \prod_k \frac{\Gamma(n_t^k + \alpha)}{\Gamma(\sum_k n_t^k + \alpha)}}{\Gamma(\sum_w n_{a,s,k}^w + \alpha) \prod_w \frac{\Gamma(n_{a,s,k}^w + \beta)}{\Gamma(\sum_w n_{a,s,k}^w + \beta)}} \cdot (11.2)$$

The full conditional probability for topics $z$ (sampled variables) is defined as:

$$P(z_{t,i} | z - z_{t,i}, a, s, w; \alpha, \beta) \propto \frac{P(z, a, s, w; \alpha, \beta)}{P(z, a, s, w; \alpha, \beta)} \frac{\Gamma(\sum_{w} n_{a,s,k}^w + \beta)}{\Gamma(\sum_{w} n_{a,s,k}^w + \beta) \Gamma(\sum_{w} n_{a,s,k}^w + \beta)} \cdot (11.3)$$

Using the property of the Gamma function $\Gamma(x + 1) = x \Gamma(x)$ we rearrange the equation above. Therefore, topics are sampled proportional to:

$$P(z_{t,i} = k | z - z_{t,i}, a, s, w; \alpha, \beta) \propto \left( \frac{n_t^k + \alpha}{n_s^t + K\alpha} \right) \left( \frac{n_{a,s,k}^w + \beta}{n_{a,s,k}^w + V\beta} \right). \quad (11.4)$$

Maximum a posteriori (MAP) estimates of the multinomial parameters $\phi_{a,s}$ and
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN
SOCIAL MEDIA

θ are defined as:
\[
\theta_{tk} = \frac{n_{tk}^{t} + \alpha}{n_k^{t} + K\alpha}, \phi_{w,k}^{a,s} = \frac{n_{w,k}^{a,s,k} + \beta}{n_k^{a,s,k} + V\beta}.
\] (11.5)

11.3.2 Attribute-Affect Mixture Model

We next propose an updated model that takes into account the fact that there
are shared topic-words across user attributes and affects, and there are attribute-
affect specific topic-words (as we previously demonstrated for gender and sentiment
in Chapter [10]). In Figure [11.2] we present the graphical notation for the Attribute-
Affect Mixture Model (AAM) that takes into account these changes. We introduce a
new binary variable x to determine whether to draw words from a shared distribution
or from attribute-affect specific topic-word distributions. The probability of x being
0 or 1 comes from a Beta distribution \( \gamma \sim Beta(\lambda_0, \lambda_1) \).

The generative process is thus:
I. Draw an attribute-affect independent multinomial word distribution
\( \phi \sim Dirichlet(\beta) \) for each topic \( z_k \).

II. Draw an attribute-affect specific multinomial word distribution
\( \phi_{a,s} \sim Dirichlet(\beta) \) for each topic \( z_k \) given an attribute \( a \) and affect \( s \) value.

III. Draw a Bernoulli distribution \( \gamma \) from \( Beta(\lambda_0, \lambda_1) \) for each topic \( z_k \), specific
attribute \( a \) and sentiment \( s \) value.

II. For each set of tweets \( T \) per user:
Choose a value for attribute \( a^{(t)} \) and sentiment \( s^{(t)} \) and draw a topic mix-
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

\[ \theta^{(t)} \sim \text{Dirichlet} (\alpha). \]

For each word \( w_i \in t: \)

- Sample a topic \( z_k \sim \text{Mult}(\theta^{(t)}). \)
- Sample \( x_i \sim \text{Ber}(\gamma). \)
- If \( x = 0 \) sample a word \( w_i \sim \text{Mult}(\phi); \) else if \( x = 1 \) sample a word \( w_i \sim \text{Mult}(\phi_{s,a}). \)

Similarly to Eq. 11.1 – 11.5 the uncollapsed and collapsed joint likelihoods of the

\[
\begin{align*}
\alpha & \quad \theta \\
\beta & \quad \phi \\
\gamma & \quad \lambda_0 \\
& \quad \lambda_1 \\
\end{align*}
\]

\[
\begin{align*}
\lambda_0 & \quad \lambda_1 \\
& \quad \text{KSA} \\
\end{align*}
\]

\[
\begin{align*}
s & \quad a \\
\end{align*}
\]

\[
\begin{align*}
w & \quad x \\
\end{align*}
\]

\[
\begin{align*}
z & \quad \text{N}_i \\
\end{align*}
\]

\[
\begin{align*}
T & \quad \text{K} \\
\end{align*}
\]

Figure 11.2: A diagram of the Attribute-Affect Mixture Model (AAM). Observed variables include Words \( w_i \sim \text{Mult}(\phi_{a,s}), \) Sentiments \( s \in \{\text{Positive, Negative}\}, \) and Attributes \( a \in \{a_0, a_1\}. \) Priors: \( \alpha, \beta, \lambda_0, \lambda_1. \) Marginalized variables: \( \theta, \phi, \gamma. \) Sampled variables: topics \( z \) and a binary switch \( x. \) \( T \) is the number user-specific sets of tweets, \( N_i \) is the length of a set of tweets per user, \( K \) is the number of topics, \( S \) in the number of affect values (e.g., positive, negative) and \( A \) is the number of attribute values (e.g., male, female).
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

data are defined as:

\[ P(z, a, s, w, \phi, \gamma | \alpha, \beta, \lambda_0, \lambda_1) \]

\[ = P(w | z, a, s, \phi) P(\phi | \beta) P(z | \theta) P(\theta | \alpha) P(x | \gamma) P(\gamma | \lambda_0, \lambda_1). \] (11.6)

\[ P(z, a, s, w, x | \alpha, \beta, \lambda_0, \lambda_1) = P(w | z, a, s, \phi) P(z | \alpha) P(x | \gamma). \] (11.7)

The full conditional probability for \( z_{t,i} \) after canceling all terms that do not involve \( z_{t,i} \) in the collapsed joint likelihood becomes:

\[ P(z_{t,i} | z - z_{t,i}, x, a, s, w; \alpha, \beta, \lambda_0, \lambda_1) = \frac{P(z, x, a, s, w | \alpha, \beta, \lambda_0, \lambda_1)}{P(z - z_{t,i}, x, a, s, w | \alpha, \beta, \lambda_0, \lambda_1)}. \]

\[ P(z_{t,i} | z - z_{t,i}, x = 0, a, s, w; \alpha, \beta, \lambda_0, \lambda_1) = \frac{\frac{\Gamma(n_k^t + 1 + \beta)}{\Gamma(\sum n_w^{a,s,k} + \beta)} \frac{\Gamma(n_k^w + 1 + \alpha)}{\Gamma(\sum n_d^k + \alpha)} \frac{\Gamma(n_a^s + 1 + \alpha)}{\Gamma(\sum n_d^s + \alpha)} \frac{\Gamma(n_{w,a}^{a,s,k} + 1 + \beta)}{\Gamma(\sum n_{w,a}^d + \alpha)} \frac{\Gamma(n_{w,a}^w + 1 + \alpha)}{\Gamma(\sum n_{w,a}^w + \alpha)} \frac{\Gamma(n_{w,a}^{a,s,k} + 1 + \beta)}{\Gamma(\sum n_{w,a}^{a,s,k} + \alpha)} \frac{\Gamma(n_{w,a}^{a,s,k} + 1 + \alpha)}{\Gamma(\sum n_{w,a}^{a,s,k} + \alpha)}}{\frac{\Gamma(n_k^t + 1 + \beta)}{\Gamma(\sum n_w^{a,s,k} + \beta)} \frac{\Gamma(n_k^w + 1 + \alpha)}{\Gamma(\sum n_d^k + \alpha)} \frac{\Gamma(n_a^s + 1 + \alpha)}{\Gamma(\sum n_d^s + \alpha)} \frac{\Gamma(n_{w,a}^{a,s,k} + 1 + \beta)}{\Gamma(\sum n_{w,a}^d + \alpha)} \frac{\Gamma(n_{w,a}^w + 1 + \alpha)}{\Gamma(\sum n_{w,a}^w + \alpha)} \frac{\Gamma(n_{w,a}^{a,s,k} + 1 + \beta)}{\Gamma(\sum n_{w,a}^{a,s,k} + \alpha)} \frac{\Gamma(n_{w,a}^{a,s,k} + 1 + \alpha)}{\Gamma(\sum n_{w,a}^{a,s,k} + \alpha)}}. \] (11.8)

Using the same property of the Gamma distribution as we did before, we find that topics are sampled proportional to:

\[ P(z_{t,i} = k | z - z_{t,i}, x = 0, a, s, w; \alpha, \beta) \propto \left( \frac{n_k^t + \alpha}{n_a^s + K\alpha} \right) \left( \frac{n_k^w + \beta}{n_{w,a}^w + V\beta} \right), \] (11.9)

\[ P(z_{t,i} = k | z - z_{t,i}, x = 1, a, s, w; \alpha, \beta) \propto \left( \frac{n_k^t + \alpha}{n_a^s + K\alpha} \right) \left( \frac{n_{w,a}^{a,s,k} + \beta}{n_{w,a}^{a,s,k} + V\beta} \right). \]
Similarly, the full conditional probability for \( x_{t,i} \) is defined as:

\[
P(x_{t,i} \mid z, x - x_{t,i}, a, s, w; \alpha, \beta, \lambda_0, \lambda_1) = \frac{P(z, x, a, s, w \mid \alpha, \beta, \lambda_0, \lambda_1)}{P(z - x_{t,i}, a, s, w \mid \alpha, \beta, \lambda_0, \lambda_1)}.
\] (11.10)

New values of \( x_{t,i} \) are sampled proportional to the following distribution:

\[
P(x_{t,i} = 0 \mid x - x_{t,i}, z, a, s, w; \gamma, \beta) \propto \left( \frac{n^{a,s,k}_{x=0} + \lambda_0}{n^{a,s,k} + \lambda_0 + \lambda_1} \right) \left( \frac{n^k + \beta}{n^k + V\beta} \right),
\]

\[
P(x_{t,i} = 1 \mid x - x_{t,i}, z, a, s, w; \gamma, \beta) \propto \left( \frac{n^{a,s,k}_{x=1}}{n^{a,s,k} + \lambda_0 + \lambda_1} \right) \left( \frac{n^{a,s,k}_w + \beta}{n^{a,s,k}_w + V\beta} \right).
\] (11.11)

MAP estimates of model parameters \( \phi, \phi^{a,s}, \theta \) and \( \gamma_{x=0}, \gamma_{x=1} \) are defined as:

\[
\theta_{tk} = \frac{n^t_k + \alpha}{n^t_s + K\alpha}, \phi^{a,s}_{wk} = \frac{n^{a,s,k}_w + \beta}{n^{a,s,k}_w + V\beta}, \phi_{wk} = \frac{n^k_w + \beta}{n^k_w + V\beta},
\]

\[
\gamma_{x=0} = \frac{n^{a,s,k}_{x=0} + \lambda_0}{n^{a,s,k} + \lambda_0 + \lambda_1}, \gamma_{x=1} = \frac{n^{a,s,k}_{x=0} + \lambda_1}{n^{a,s,k} + \lambda_0 + \lambda_1}.
\] (11.12)

### 11.3.3 Comparing Attribute-Affect Models to Other Approaches

The ccMix model (Zhai et al., 2004) is designed to compare text collections by extracting what is common and specific for each collection using probabilistic latent semantic indexing (pLSI) (Hofmann, 1999). Model parameters are estimated using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). Unlike the ccMix model, our attribute-affect models rely on LDA (Blei et al., 2003) rather than
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

pLSI and do not rely on a pre-defined parameter $\lambda_c$ (which represents the probability of drawing a word from collection-independent instead of collection-specific word distributions). Thus, our model can fit the data better by modeling the fact that some topics might be shared across attributes.

LDA-Collocation (Griffiths et al., 2007) and Topical N-Grams (X. Wang et al., 2007) assume that a word can come from two distributions: a topic-word distribution or a distribution associated with a previous word. Unlike these approaches, our attribute-affect models assume that a word is drawn from attribute-affect specific distribution rather than a distribution associated with a previous word.

The most similar to our attribute-affect models is a cross-collection LDA model (ccLDA) (Paul & Girju, 2009). It is designed to compare multiple text collections. It assumes that each word comes from either collection-specific or collection-independent word distributions. Unlike the ccLDA model, in this work we are trying to model extra-linguistic aspects of language (e.g., attributes and affects) jointly rather than trying to elicit differences between two collections. Moreover, the ccLDA model assumes non-uniform Dirichlet priors that depend on a document’s collection.
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

11.4 Experimental Results

11.4.1 Baseline Attribute-Affect LDA Model Results

The exact inference for the AALDA model is intractable. Thus, to estimate model parameters we use an approximate inference and implement a collapsed Gibbs sampler \cite{Heinrich2005}. Gibbs sampling \cite{Gilks1996} is a type of Markov Chain Monte Carlo (MCMC) algorithm that has been successfully applied for inference in LDA models \cite{Griffiths2004}. It is comparable to other estimators in speed such as variational EM \cite{Blei2003} and can be easily derived as we demonstrate in section 11.3.

We ran a sampler for 250 iterations and average model parameters over the last 25 iterations\footnote{This approach is a standard procedure while making approximate inference, e.g., using Gibbs sampling.}. At every iteration we sample new assignments of hidden variables (e.g., topics $z$) as defined in Eq. 11.4 by drawing from the distributions conditioned on a previous state of the model.

We experiment with $\alpha$ and $\beta$ priors to maximize the log-likelihood on the held-out data. We observed that if $\alpha \geq 1$ then tweets contain a majority of the topics, whereas if $\alpha < 1$ then tweets contain a mixture of a few topics or a single topic. Similarly, if $\beta \geq 0.01$ then each topic contains a mixture of all of the words; if $\beta < 0.01$ then each
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

...topic contains a mixture of just a few words.

Below we present shared $S =$ negative topics across $A =$ genders:

- Topic 0: shit fuck ass niggas nigga bitch aint got damn [profanity]
- Topic 7: just like people hate fuck get bad really shit fucking [frustration]

Below we present shared $S =$ positive topics across $A =$ genders:

- Topic 3: happy birthday love day thank hope miss much best [birthday wishes]
- Topic 2: thanks thank much great hope amazing beautiful well [appreciation]

In Table 11.2 we present example topics with positive and negative sentiment for some example attributes.

11.4.2 Preliminary Attribute-Affect Model Results

In an approach similar to the AALDA model, we used an approximate inference for the AAM model and implemented a collapsed Gibbs sampler [Heinrich 2005]. At each Gibbs sampling iteration, we sampled hidden variables – topics $z$ and a switch variable $x$ following the derivations in Eq. 11.9 and Eq. 11.11 respectively. We ran our Gibbs sampler for 4000 burn-in iterations. By minimizing the log-likelihood on a held-out dataset we found the following optimum values for four key model parameters: the number of topics $z = 10, \alpha = 1, \beta = 0.01, \lambda_0 = \lambda_1 = 1$.

We pre-processed data by converting text to lowercase and removing the following features: stopwords, urls, punctuation, words with frequency $\leq 2$, and usermentions, e.g., @obama. We also converted hashtags so that e.g., #love became hashtaglove.
### CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

<table>
<thead>
<tr>
<th>Attribute-Affect</th>
<th>Example Topics (one topic per line)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male+</td>
<td>brazil world cup germany game goal win #worldcup argentina via fitness loving babes follow daily fire entertainment #fit via obama job #okcbarons president data growth exciting #mmccon #smemchat</td>
</tr>
<tr>
<td>Female+</td>
<td>#aries scope free via delicious check money healthy recipe deals connected #stylechat closet #partydownsouth #digifem #etiquette #propertyvirgins #foodfri #wellnesswed #love #family #happy #amazing #mmva #baby #beautiful #yeg #photography #mcm</td>
</tr>
<tr>
<td>Male−</td>
<td>sorry service customer please hear experience airport follow i due accident blocked east lane congestion right west left daughter nascar race blast locate help missing indycar #nascar</td>
</tr>
<tr>
<td>Female−</td>
<td>obama via women war #tcot law gop rape bush shannon #happydays photo physiology yhu #psych fuh #day</td>
</tr>
<tr>
<td>Kids+</td>
<td>love just can new connect connected partners listening find beautiful photo morning child food fun #love today best can news may support kids help might world something interesting hand making summer story</td>
</tr>
<tr>
<td>Kids−</td>
<td>just got like hot little dog gt mom hair going car put told house old stop feel one eating miss really like just sad much omg hate feel summer actually crying bad seriously school okay ugh wait life</td>
</tr>
<tr>
<td>Stressed+</td>
<td>good new song music cool like best think free fun pretty way yeah wow around hey yes job looks kids one best movie made video watch liked fun getting taking</td>
</tr>
<tr>
<td>Stressed−</td>
<td>obama will gop usa #morningjoe kids killed cavs years america via must news american law talk can country work get just need home wait tired like hate day going want really back night now sleep wish lol still</td>
</tr>
<tr>
<td>Narcissist+</td>
<td>ford vegas #fordchatca total goal poker june steps #fantasyfootball #jenniferfalls photo honor #summerreading lakota photoset brand enchanted order client #libra #idol #bgc wally jena decode download retweets looool</td>
</tr>
<tr>
<td>Narcissist−</td>
<td>#collegietteproblems boating traveling #happydays market fooled superior nuggets #herconference hayley cancer today obama may might obamacare via president leo #tcot</td>
</tr>
</tbody>
</table>

Table 11.2: The most representative example topics with positive e.g., Male+ and negative Male− sentiment for gender, age, children, stress and narcissism attributes. Tokens that start with # are hashtags e.g., #worldcup, #morningjoe.
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN
SOCIAL MEDIA

In Tables 11.3 and 11.4 we present positive and negative topics for male vs. female
users in social media. For example, we observed several negative topics – topic 2
(politics), topics 4 (swear words), topic 9 (work) – which demonstrate significant
topic variations across genders. Similarly, positive topics 1 and 4 (entertainment) as
well as topic 6 (birthday wishes) highlight language variations across genders.

11.5 Conclusions

In Chapter 11 we demonstrated preliminary results on how extra-linguistic features
such as latent user attributes, sentiments and emotions (affect) can be successfully
used for a downstream NLP application, attribute-affect specific language modeling.
We developed attribute-sentiment specific topic models to study subjective language
variations for social media users with contrastive demographics (e.g., male vs. female,
older vs. younger). We derived and implemented inference for the proposed models,
and ran exploratory experiments.

Moreover, our models are easily extendable. For instance, they could be applied if
one needed to model (a) sentiment or emotion annotation uncertainty, (b) differences
in affective language of users vs. their neighbors (e.g., friends, followers), or (c)
user-neighbor attribute-affect change over time [Griffiths & Steyvers, 2004].
TABLE 11.3: Top 10 words for 10 negative topics for male M and female F users. For every topic, the first line shows the topic shared across genders. On the lines below, M is a male-specific topic and F is a female-specific topic.
CHAPTER 11. MODELING ATTRIBUTE-AFFECT TOPIC VARIATIONS IN SOCIAL MEDIA

<table>
<thead>
<tr>
<th>Topic 1: like good haha lol one just really love ever yeah</th>
<th>Topic 2: god love good day today will life thank morning great</th>
</tr>
</thead>
<tbody>
<tr>
<td>M⁺: brazil world cup game germany goal win #worldcup argentina messi</td>
<td>M⁺: world #virgo perfect open east west job free lane congestion</td>
</tr>
<tr>
<td>F⁺: cute omg mom dress concert men visit wear probably full</td>
<td>F⁺: enjoy #aries world loving wonderful ready smile sweet friday sun</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 3: lol good love ass like got get yall snapchats lmao</th>
<th>Topic 4: love new amazing thank follow thanks video show liked music</th>
</tr>
</thead>
<tbody>
<tr>
<td>M⁺: bro bruh shit nigga fam nice yeah new nobody dope</td>
<td>M⁺: follow video liked fitness fire babes music pleaseeee track business</td>
</tr>
<tr>
<td>F⁺: boo baby lol hair nice sleep finally yeah job want</td>
<td>F⁺: dress justin series win thx forever twitter sweetie model role</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 5: like just one can well really think pretty yes cool</th>
<th>Topic 6: happy love birthday day thank thanks hope best much miss</th>
</tr>
</thead>
<tbody>
<tr>
<td>M⁺: new thats connect connected partners man album will episode works</td>
<td>M⁺: man bro thanks haha year show dude brother watching soon</td>
</tr>
<tr>
<td>F⁺: watching book need always read saying made seeing sorry post</td>
<td>F⁺: love girl perfect year girls beautiful years world summer lady</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 7: game win good team will best great one better get</th>
<th>Topic 8: thanks great good thank will happy awesome best day can</th>
</tr>
</thead>
<tbody>
<tr>
<td>M⁺: lebron lakers bulls kevin melo trade cavs spurs cleveland carmelo</td>
<td>M⁺: please interview die long hear sorry company seen thx band</td>
</tr>
<tr>
<td>F⁺: brazil cup germany world goal game #worldcup argentina usa lucky</td>
<td>F⁺: free delicious coffee healthy east fabulous money career deals join</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic 9: day today good night tonight get time work just best</th>
<th>Topic 10: love just life like best can people never always will</th>
</tr>
</thead>
<tbody>
<tr>
<td>M⁺: tonight nice great free yeah #nowplaying playing town concert wfr</td>
<td>M⁺: good lol real gotta sleep cool world now well great</td>
</tr>
<tr>
<td>F⁺: new excited lol yay wait little amazing makes free yes</td>
<td>F⁺: love real world sleep baby boyfriend funny friend now miss</td>
</tr>
</tbody>
</table>

Table 11.4: Top 10 words for positive topics for male (M⁺) and female (F⁺) users.
Chapter 12

Conclusions and Future Directions

The emergence of social media services such as Twitter, Google+ and online social networks like Facebook has dramatically influenced peoples’ lives over the last decade. These services allow people to share their lives with others by posting updates and pictures online, follow other peoples’ posts and stay connected across the world. As of 2015 1/7 th of the world’s population is using social media regularly, and this number is constantly growing. Millions of users generate billions of messages daily in more than 100 languages. As a result of people’s engagement in social media communications, tons of personal and up-to-date information has become publicly available.

In this work, we showed how to use this public data with the purpose of learning more about people – not just demographic data, but also their personalities, online behavior, opinions, emotions, likes and interests. Even though the data we use is
CHAPTER 12. CONCLUSIONS AND FUTURE DIRECTIONS

public (according to Twitter’s Terms of Service\(^1\)), we deliberately include moral and ethical procedures in our research workflow. For instance, we report our findings on an abstract and aggregate level, and we do not share or track any personal information when demonstrating or accessing user example posts.

This thesis presents novel approaches and practical techniques for streaming personal analytics in social media focusing on the task of user demographic prediction. We developed methods, features and practical applications for automatically inferring latent user demographics from unstructured, noisy, streaming communications published in social media. Unlike the existing approaches for demographic prediction in social networks that rely on supervised models and lexical features extracted from thousands of messages per user (Rao et al., 2010; Zamal et al., 2012), we proposed approaches for constrained-resource classification, streaming prediction, iterative learning and inference with interactive annotation and weighting techniques. Predictive models developed in this work allow analysis of the relationships between user language, perceived psycho-demographic traits, emotions, opinions and interests in social media on an unprecedented scale – 123k users, 25M messages.

In Chapter 3 we proposed models for constrained-resource prediction to address the limitations of accessing public social media streams, e.g., querying Twitter API rates as described in Appendix C. These models allow us to infer

\(^1\)Twitter’s Terms of Service: [https://twitter.com/tos?lang=en](https://twitter.com/tos?lang=en)
CHAPTER 12. CONCLUSIONS AND FUTURE DIRECTIONS

latent properties of social media users with no or limited content. They rely on posts from user neighbors of different types, e.g., friends, retweets extracted by mining social network structure or joint user-neighbor content. We applied our constrained-resource models to infer three different attributes: gender, age and political preferences. Our experiments demonstrate how much the amount of content per user (e.g., 5 vs. 200 messages), feature types and neighbor types influence classification performance for each attribute. This work is the first to demonstrate that user gender is most predicted by user friends, user age best predicted by followers, and political preferences most predicted by retweets and user mentions. We also explored the trade-offs between exploring vs. exploiting the content of neighbors. We found that the signal is distributed in the neighborhood. Thus, given the limited amount of queries to public APIs, it is more effective to query for more neighbors with less content per neighbor than request data from fewer neighbors with more content per neighbor.

In Chapter 4 we developed novel streaming approaches that model the dynamic nature of social media by allowing iterative predictions of latent user properties over time. Our tweet-based streaming models rely on iterative Bayesian rule updates to dynamically update beliefs about user latent attributes over time. We applied our models to infer user political preferences from user, neighbor and joint user-neighbor streams. Our experiments show that streaming models are more effective than batch models for the personal analytics presented in Chapter 3 (up to 24% improvement in absolute accuracy). We found that inferring from joint
user-neighbor streams improves prediction performance (up to 10% gain in absolute accuracy) and resolves the issue of topical sparsity, which is when users communicate about only one specific topic. We empirically estimated the number of messages per user and the amount of time necessary to obtain certain level of accuracy. For instance, we found that political preference can be often be predicted using roughly 100 tweets, depending on the context of user selection, where this could mean hours, or weeks, based on the author’s tweeting frequency.

In Chapter 5, we used our streaming models from Chapter 4 in several settings that allow training new models in addition to making predictions over time. We experimented with active learning and iterative retraining approaches for inferring user political preferences from user or joint user-neighbor streams of communications. These improvements make is possible to handle the data drift of social media and lead to more accurate and generalizable models. We advanced the proposed approaches with iterative interactive rationale annotation techniques via crowdsourcing. We found that (1) active learning models outperform iterative batch re-training approaches, (2) learning from joint user-neighbor streams is better than learning from a user stream; (3) rationale annotation and filtering techniques significantly improve classification results (up to 27% relative improvement).

In Chapter 6, we showed how to improve models for latent attribute prediction by effectively incorporating human knowledge into the learning process via rationale annotation and weighting techniques. Rationales are highly predic-
CHAPTER 12. CONCLUSIONS AND FUTURE DIRECTIONS

tive n-grams annotators highlight to justify their labeling decisions. We introduced several rationale weighting schemes integrated in different batch models with varied amount of supervision. We found that rationale annotation and weighting techniques not only improve classification results (24% – 28% relative error reduction in a fully supervised and semi-supervised setting, respectively) but also enable the development of models that can be scaled to other languages, e.g., Spanish.

In Chapter 8 we presented a real-word framework for social media predictive analytics. Given a user ID, our system accesses that individual’s 200 most recent tweets and applies static models from Chapter 3 to automatically predict a variety of perceived psycho-demographic attributes per user, six fine-grained emotions and three sentiments expressed in user tweets. Using this system we automatically inferred a variety of perceived attributes for 123k Twitter users and annotated their 25M tweets with emotions and opinions. We then performed an extensive analysis on correlations between perceived user psycho-demographics, emotions and opinions in this dataset.

- First, we found that users with contractive demographics, e.g., male and female users express emotions and opinions differently in social media. More precisely, users predicted to be in a relationship, with children, higher education and higher income tend to express more joy and less sadness than users with respective contrastive demographics. Female users tend to produce more emotional and opinionated tweets than male users.
• Second, we showed that users with perceived contrastive demographics tend to react differently to the emotional tone projected in their respective social environments. Depending on user demographics some users amplify certain emotions expressed by their neighbors whereas other users dampen them. More precisely, we found that users predicted to be with lower income tend to amplify sad tweets from their neighbors, but older users or those with children dampen them; users satisfied with life dampen angry tweets whereas users dissatisfied with life amplify them. Optimists amplify joyful tweets while pessimists dampen them.

• Third, we demonstrated that latent user properties can be predicted using only emotions expressed in user tweets and user-neighbor emotional contrast rather than lexical features extracted from user tweets. This is especially true for the attributes that correlate highly with emotions, e.g., life satisfaction, optimism, stress etc.

In Chapter 9, we qualitatively and quantitatively evaluated user interests to predict their psycho-demographics, and we compared the results with existing models that rely on language. Our findings regarding correlations between user interests and perceived psycho-demographics were interesting and novel. For instance, users perceived to be narcissists tended to be interested in science; those perceived as optimists were interested in health, food and drinks; while pessimists were interested in news.
CHAPTER 12. CONCLUSIONS AND FUTURE DIRECTIONS

In Chapters 10 and 11, we demonstrated how user demographic attributes (aka extra-linguistic features) can be helpful for several downstream applications such as gender-aware opinion analysis and modeling attribute-affect topic variations in social media.

12.1 Summary of Contributions

The major contributions of this thesis are as follows:

Methodologies and practical techniques:

1. This work goes beyond the network-driven models of Zamal et al. (2012) and is the first to compare and study the implications of using different types of relationships in social networks to predict a variety of user demographic attributes.

2. This thesis reports original findings regarding (a) which neighbors are the most predictive of certain demographics attributes and (b) methods to greedily explore user neighborhoods in social networks.

3. This work builds upon the approaches of Burger et al. (2011) and investigates in depth how attribute prediction performance depends on the amount of data available in constrained-resource conditions.

4. This work advances all existing models for batch attribute inference as well
CHAPTER 12. CONCLUSIONS AND FUTURE DIRECTIONS

as the online models proposed by Van Durme (2012b) and presents a novel algorithm for predictive analytics of streaming social media.

5. This thesis outlines original approaches for training models actively on the fly in addition to making predictions of user properties dynamically, going beyond all existing static (batch) approaches.

6. This work presents original methods for incorporating crowdsourced rationale annotations for personal analytics in social media by novel application, evaluation, and extension of previous work on rationales (O. Zaidan & Eisner, 2008; Yessenalina et al., 2010).

7. This thesis is the first to present novel findings and insights about the relationships between user language, demographics, emotions and interests in online social networks on an unprecedented scale, having analyzed thousands of users and millions of tweets.

8. This work demonstrates that incorporating extra-linguistic features such as user demographics and emotions can improve performance for several downstream NLP tasks including sentiment analysis in multiple languages and personalization (e.g., attribute-affect specific language generation in social media).

9. This work is first to address and investigate in detail the issues of (a) san-

\[\text{We report our findings following moral, ethical and privacy-aware procedures recently discussed at the 2015 ICML's Fairness, Accountability, and Transparency in Machine Learning Workshop: http://www.fatml.org/index.html}\]
pling and annotation biases, and (b) model generalization for user attribute prediction, going beyond the existing recent studies on experimenting with “big data analytics” (Cohen & Ruths 2013; Tufekci 2014).

- **Software and Data:**

  1. We, in collaboration with Microsoft Research, developed software (Volkova, Bachrach, Armstrong, & Sharma 2015) for predicting psycho-demographic attributes, emotions and opinions for any public Twitter profile.

  2. We developed a software package that implements online Bayesian update models for streaming analytics.

  3. We released several datasets annotated with latent user properties, psycho-demographic attributes, six basic Ekman’s emotions and 26 Twitter interest categories, as well as lists of annotator rationales in multiple languages. All such data release were in accordance with Twitter’s data sharing policy, and they followed moral, ethical and privacy-aware protocols as discussed in Appendix D.
CHAPTER 12. CONCLUSIONS AND FUTURE DIRECTIONS

12.2 Social Media Analytics Challenges Addressed

The approaches developed in this thesis for latent user attribute prediction addressed several previously unexplored issues for social media predictive analytics, including but not limited to:

- **Dynamic streaming nature of social media:** We developed models for streaming analytics (Chapter 4).

- **Data (concept) drift that causes models to degrade over time:** We developed approaches for iterative learning and inference with interactive rationale annotation (Chapters 5 and 6).

- **Topical sparsity that leads to poor model generalization:** We proposed models that rely on user neighbor content (Chapters 3–5), on user-neighbor emotional contrast (Chapter 8), and on user interests (Chapter 9).

- **User activeness:** Since user activity levels influence both prediction time and speed for streaming analytics as well as static constrained-resource setting, we experimented on multiple datasets throughout this thesis (Chapters 3–11).

- **Data sampling and annotation bias:** To reduce the influence these biases may exert on prediction performance, we experimented on multiple datasets sampled differently and annotated with self-reported vs. perceived attributes (Chapters 3–11).
12.3 Implications

This work may advance the current understanding of social media populations, their online behavior and overall well-being. Not only can it enable researchers from disciplines as diverse as computational psychology, sociology, social science and social media analytics with methodologies, techniques and tools to better understand and study social media users and their behavior online from all over the world in realtime, it can also effectively improve a variety of downstream applications (as we demonstrated in Chapters 10 and 11). These applications could include personalized intelligent user interfaces, recommendation systems, large-scale healthcare analytics or personalized conversational agents which could evolve in ways that reflect and adapt to user-specific demographics and emotions as discussed in detail in Chapter 1.

Moreover, our original findings from Chapters 8–9 have several noteworthy implications for the study of self-disclosure in online settings. First, with regard to self-disclosure, we note that many people make a conscious effort to project a certain image online. Our results indicate that the emotions displayed on social media are predictive of socio-demographic traits. Thus, it might be possible for people to project a desired impression by controlling the emotions they display online. This finding highlights the role that emotions play in self-disclosure even in the online environment. Further, our results indicate that people may disclose more than they intend to, as raw information regarding emotions can be distilled into much deeper and more intimate insights. This result stands in contrast to the progression of social
penetration theory in offline settings, where records of emotions are not stored for long periods of time.

12.4 Future Directions

Many questions remain open for future research, including but not limited to:

- How predictive is other content for social media analytics, e.g., images, videos, links to the external web-pages (You, Bhatia, Sun, & Luo, 2014)?

- How might we combine image and language features for social media predictive analytics, e.g., using text embeddings (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Levy & Goldberg, 2014) and deep learning (Szegedy et al., 2014; Yosinski, Clune, Bengio, & Lipson, 2014)?

- How will models trained on Twitter content perform on content from other social networks, e.g., Facebook, Google+, Instagram, YouTube (Sap et al., 2014)?

- What approaches can be effectively applied to jointly infer user demographics, emotions and interests for a large set of interconnected users?

- Can models trained on social media data be successfully applied to develop virtual personal assistants (also known as conversational agents) that are:

  - personalized: estimating and reflecting latent user properties, e.g., a woman who is sad, an older user who is surprised,
CHAPTER 12. CONCLUSIONS AND FUTURE DIRECTIONS

- **real-time**: interpreting and adapting to language in context and visual environment,
- **adaptive**: interacting with a user based on latent user characteristics with the objective of satisfying user intents,
- **predictive**: anticipating the user.

- Can “big data analytics” potentially be used as an alternative to or a validation tool\(^3\) to time-consuming, state-of-the-art user studies conducted by sociologists and psychologists (Collaboration et al., 2015)?

\(^3\)Over half of psychology studies fail reproducibility test: [http://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248](http://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248)
Appendix A

Mechanical Turk HIT Examples

In Figures A.1–A.7 we present three example HITs (Human Intelligence Tasks posted on Amazon Mechanical Turk) developed to crowdsource gender, age and political preference annotations for the datasets used in Chapter 3, gender and age attributes with rationale labels for the datasets used in Chapter 6, and, finally, demographics and online identity for the datasets used in Chapters 8, 9 and 11.
# Identify Person’s Gender, Age and Political Orientation from a Twitter Account

**Show/Hide Informed Consent Form**

**ATTENTION!** In this HIT we compare your answers with our standard answers. Please be diligent!

<table>
<thead>
<tr>
<th>Instructions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our goal is to determine whether a user is Male or Female, Below 35 or Above 35 years old, Democratic or Republican from his or her tweets.</td>
</tr>
</tbody>
</table>

You are given a link to a Twitter account. Your task is to read as many tweets as you need (you can access all of them by clicking "View full profile" after you open the profile page) to determine person’s:

1. gender: Male or Female;
2. age: Below 35 or Above 35 years old;
3. political affiliation: Democratic or Republican.

ONLY if a user profile or tweets are not available/private or a profile is not associated with a person, you choose N/A and specify the reason in the comment. In all other cases you have to determine all attributes.

<table>
<thead>
<tr>
<th>Examples:</th>
</tr>
</thead>
</table>
| **Tweets:** That’s all I’ve been doing thus far. Lots of places to fly fish. Been kayaking bunch of times. Explored the vineyards and did some tasting yesterday. On our way down the Mountain to finally pick up the Browning Shotgun I bought almost two weeks ago. Gender: Male
Age: Above 35
Political: Republican
**Tweets:** I’m partying dude. Another day in America.... Another school shooting.... Thinking of all those involved at Lone Star College! #WakeUpAmerica
Gender: Male
Age: Below 35
Political: Democratic |

Figure A.1: Screen 1 (Instructions and Examples) in MTurk HIT, designed to crowdsource annotations for gender, age and political preferences of Twitter users.
Figure A.2: Screen 2 (Task) in MTurk HIT, designed to crowdsource annotations for gender, age and political preferences of Twitter users.
APPENDIX A. MECHANICAL TURK HIT EXAMPLES

Figure A.3: Screen 1 (Instructions and Examples) in MTurk HIT, designed to crowdsourced annotations for gender and age of Twitter users, and the annotator rationales (clues).

Instructions:

Our goal is to predict whether an author is male or female and what is his/her age from ARABIC tweets.

You are given a link to person’s Twitter account and you need to:

1. select person's gender based on his/her tweets in Arabic and copy gender clues from his/her tweets;
2. select person's age based on his/her tweets in Arabic and copy age clues from his/her tweets.

When Twitter profile link is not available you have to make judgments about gender/age based on a given tweet (which is always in Arabic).

Examples:

Tweet: That's all I've been doing thus far. Lots of places to fly fish. Been kayaking bunch of times. Explored the vineyards and did some tasting yesterday.

Gender: Male [High Confidence]
Gender clues: fish; kayaking
Age: Middle-Age Adult (35-59) [Middle Confidence]
Age clues: fish; vineyards; kayaking

Tweet: How unfortunate! My husband picked Ami up at Owariasahi-city, but is still trying to find a way to cross the river.

Gender: Female [Certain]
Gender clues: my husband
Age: Young Adult (20-34) [Middle Confidence]
Age clues: husband; picked Ami
APPENDIX A. MECHANICAL TURK HIT EXAMPLES

Figure A.4: Screen 2 (Task) in MTurk HIT, designed to crowdsource annotations for gender and age of Twitter users, and the annotator rationales (clues).

To access person's Twitter profile click here: ${username}

Below is a tweet in ARABIC from a single person. You have to judge from this tweet only if a link to his/her Twitter profile is not available.

${tweet}

SELECT GENDER FROM OPTIONS BELOW:

○ Male [Certain]
○ Male [High Confidence]
○ Male [Medium Confidence]
○ Male [Low Confidence]

○ Female [Certain]
○ Female [High Confidence]
○ Female [Medium Confidence]
○ Female [Low Confidence]

Please copy 5+ words or phrases from tweets that you used as CLUES for the author's GENDER (male/female). Please separate them by semicolons.

SELECT AGE FROM OPTIONS BELOW:

○ Teenager (14-19) [Certain]
○ Teenager [High Confidence]
○ Teenager [Middle Confidence]
○ Teenager [Low Confidence]

○ Young Adult (20-34) [Certain]
○ Young Adult [High Confidence]
○ Young Adult [Middle Confidence]
○ Young Adult [Low Confidence]

○ Middle-Age Adult (35-59) [Certain]
○ Middle-Age Adult [High Confidence]
○ Middle-Age Adult [Middle Confidence]
○ Middle-Age Adult [Low Confidence]

○ Older Adult (60+) [Certain]
○ Older Adult [High Confidence]
○ Older Adult [Middle Confidence]
○ Older Adult [Low Confidence]

Please copy 5+ words or phrases from tweets that you used as CLUES for the author's AGE. Please separate them by semicolons.
APPENDIX A. MECHANICAL TURK HIT EXAMPLES

Introduction

Twitter Demographics and Online Identity

For this task, you will need to judge about peoples' demographic attributes and online identity from their Twitter profiles and tweets.

Please note that for every statement in Online Identity, you should select one of the following options: disagree strongly, disagree, neither agree nor disagree, or agree strongly. Please choose the answer that best reflects your opinion.

Note: You will be rejected if you fail to submit your verification code in Mechanical Turk. Everyone who submits a valid verification code (received at the end of the HIT) and delivers high quality work will be approved.

Start Task

Figure A.5: Screen 1 (Instructions) in MTurk HIT, designed to crowdsource annotations for demographics and online personality of Twitter users.
Demographics

Click to see the user’s profile

Age: 

Gender: [Please select...]

Race: [Please select...]

Religion: [Please select...]

Education: [Please select...]

IQ: [Please select...]

Income (in US Dollars): [Please select...]

Marital status: [Please select...]

Has children: [Please select...]

Political affiliation: [Please select...]

Figure A.6: Screen 2 (Demographics) in MTurk HIT, designed to crowdsource annotations for demographics and online personality of Twitter users.
Online Identity

![Please tell us about the way you identify this person from his/her Twitter profile](image)

Do you think this person...

<table>
<thead>
<tr>
<th></th>
<th>Disagree strongly</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Agree strongly</th>
</tr>
</thead>
<tbody>
<tr>
<td>- presents himself/herself on Twitter 'selectively' in order to give a certain impression or posts information in the profile that is not true e.g., posts only about certain topics, posts only positive thoughts, doesn’t tell about his/her failures, only posts pictures where they look good etc.</td>
<td>○ ○ ○ ○ ○ ○</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- is self-promoting e.g., extensively advertising oneself or one’s activities, posts tweets and photos to show people the different aspects of who he/she is (especially in a very assertive / aggressive way)</td>
<td>○ ○ ○ ○ ○ ○</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- is showing-off or arrogant e.g., tries to impress others with the posts, compares himself with others, trying to get others to envy them, gloating, bragging about new purchases or achievements etc.</td>
<td>○ ○ ○ ○ ○ ○</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This person seems to be...

<table>
<thead>
<tr>
<th></th>
<th>Disagree strongly</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Agree strongly</th>
</tr>
</thead>
<tbody>
<tr>
<td>- a narcissist e.g., only interested in themselves and not others or self-centered</td>
<td>○ ○ ○ ○ ○ ○</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- excited, outgoing and enthusiastic e.g., posts very positive tweets, seeks excitement, sociable</td>
<td>○ ○ ○ ○ ○ ○</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- anxious and moody, stressed out e.g., posts very negative / angry tweets, worries a lot, gloomy</td>
<td>○ ○ ○ ○ ○ ○</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please give your best estimate about this person’s **level of optimism**:

Please select...

Please give your best estimate about this person’s **satisfaction with life**:

Please select...

Next

Figure A.7: Screen 3 (Online Personality) in MTurk HIT, designed to crowdsource annotations for demographics and online personality of Twitter users.
Appendix B

Twitter Graph Statistics

In Figure 1.1 we presented the example Twitter statistics that demonstrates differences in user activeness. There are other resources online which report similar statistics for Twitter e.g., [http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/](http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/) As of August 2015, we found that 43% of Twitter users have not tweeted within the past year, 44% have created an account but never sent a tweet.

We estimated user activeness by sampling 2M users from the 1% Twitter feed. For each user we calculated the number of tweets per day (TPD) since the account was created. Figure B presents the histogram of users with a certain number of daily tweets over the past 6 years. We found that the median number of daily tweets is 10. Users who tweet between 4 - 10 TPD are considered to be the “average” users, below 4 TPD – not active users, and above 10 TPD - highly active users.
Figure B.1: Distribution of daily tweets for users sampled randomly over time from the 1% Twitter feed.
APPENDIX B. TWITTER GRAPH STATISTICS

In Table B we outline the groups of Twitter users clustered by the number of tweets per day and the days since the account was created by manually inspecting user profiles within each cluster.

<table>
<thead>
<tr>
<th>Tweets Per Day</th>
<th>Active &lt; 60 days</th>
<th>Active ≥ 60 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low 1 – 3</td>
<td>New users</td>
<td>Not active users</td>
</tr>
<tr>
<td>Medium 4 – 10</td>
<td>Moderately active new users</td>
<td>Moderately active returning users</td>
</tr>
<tr>
<td>High 10 – 27</td>
<td>Active users</td>
<td>Highly active returning users</td>
</tr>
<tr>
<td></td>
<td>(businesses, services)</td>
<td>(dedicated) users</td>
</tr>
<tr>
<td>Extremely high ≥ 27</td>
<td>Hyper active short-term users</td>
<td>Hyper active users</td>
</tr>
<tr>
<td></td>
<td>(spammers, bots)</td>
<td>(News, Celebrities)</td>
</tr>
</tbody>
</table>

Table B.1: Clustering Twitter users by their tweeting frequency and the time since their accounts were created.

Below we analyze one of our sampled Twitter graphs in more detail – candidate-centric graph. More specifically, we visualize the distributions of followers, friends, retweets, replies, hashtags and user mentions in the graph and calculate its basic characteristics – graph size, graph order and maximum degree.

B.1 Candidate-Centric Graph Characteristics

We construct a candidate-centric graph $G_{cand}$ by looking into following relationships between the users and Democratic or Republican candidates during the 2012 US Presidential election as shown in Figure B.2. We define Democratic or Republican
APPENDIX B. TWITTER GRAPH STATISTICS

users as those that followed both @BarackObama and @JoeBiden, or @MittRomney and @RepPaulRyan. We randomly sample $n = 516$ Democratic and $m = 515$ Republican users. We label users as Democratic if they exclusively follow both Democratic candidates but do not follow both Republican candidates and vice versa. We collectively refer to $D$ and $R$ as our “users of interest”.

![Candidate-centric graph with D and R candidates](image)

Figure B.2: Candidate-centric graph with $D$ and $R$ candidates, ●: users of interests (blue: $D$, red: $R$, ○: users not sampled, →: follower relations, ↔: friend relations.

For each such user we collect a biography, recent tweets and randomly sample user-local neighborhoods – the immediate adjacent neighbors of a user of followers, friends, user mentions, replies, retweets and hashtags. To the best of our knowledge this resultant candidate-centric graph is the largest Twitter collection assembled with respect to labeled political preference (50k vertices, 60k edges) and including sampled users across a diverse set of local neighborhoods.

In Figures [B.3 - B.5] we present the distributions of followers, friends, retweets, user mentions, replies and hashtags estimated over candidate-centric graph $G_{cand}$.

---

1As of Oct 12, 2012, the number of followers for Obama, Biden, Romney and Ryan were 2m, 168k, 1.3m and 267k.
Figure B.3: Distribution of followers and friends in the cand-centric graph.
APPENDIX B. TWITTER GRAPH STATISTICS

- The distributions of followers and friends are similar for Democratic and Republican users, and only 1-2% users have less than 10 followers or friends. Overall, both $D$ and $R$ users of interest have from 200 to 500 followers or friends. However, only 50% Democratic and 57% Republican users of interest have less than 500 friends compared to 80% Democratic and 77% Republican users who have less than 500 followers.

- The distribution of replies show that 39% Democratic and 36% Republican users replied less than 25 times per 200 tweets (1% - 12% reply rate), 50% Democratic and 52% Republican users replied 26 - 100 times per 200 tweets (13% - 50% reply rate), and only 11% Democratic and 12% Republican users replied more than 100 times per 200 tweets (more than 50% reply rate).

- The distribution of user mentions shows that 18% Democratic and 26% Republican users utilize less than 50 user mentions, 39% Democratic and 44% Republican users apply from 50 to 100 user mentions, 43% Democratic and 30% Republican users apply more than 100 user mentions per 200 tweets.

- The distribution of retweets demonstrates that 77% Democratic and 72% Republican users of interest retweet less than 100 times per 200 tweets.

- The distribution of hashtags shows that 28% Democratic and 34% Republican users tweet less than 25 hashtags, 67% Democratic and 61% Republican users tweet 26 - 100 hashtags, and only 5% users tweet > 100 hashtags per 200 tweets.
APPENDIX B. TWITTER GRAPH STATISTICS

We compare the original (as of Oct 12, 2012) and sampled follower and friend $C_{\text{cand}}$ graphs in Figure B.6.

To compare the original and sampled graphs, we report graph statistics for the original and sampled graphs in Table B.2. We consider graph statistics such as:

- graph size $G_{\text{size}} = |E| = |E(G)|$ - the number of edges in the graph;

- graph order $G_{\text{order}} = |U| = |U(G)|$ - the number of vertices in the graph;

- graph maximum degree $G_{\text{degree}} = |\Delta| = |\Delta G|$ - the number of edges incident to the vertex, with loops counted twice.

<table>
<thead>
<tr>
<th>Graph types</th>
<th>$G_{\text{order}}^{\text{cand}}$</th>
<th>$G_{\text{size}}^{\text{cand}}$</th>
<th>$G_{\text{degree}}^{\text{cand}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original follower graph</td>
<td>1,185,105</td>
<td>1,371,557</td>
<td>292,999</td>
</tr>
<tr>
<td>Sampled follower graph</td>
<td>10,499</td>
<td>9,894</td>
<td>10</td>
</tr>
<tr>
<td>Original friend graph</td>
<td>4,622,884</td>
<td>836,771</td>
<td>50,344</td>
</tr>
<tr>
<td>Sampled friend graph</td>
<td>9,206</td>
<td>9,772</td>
<td>10</td>
</tr>
</tbody>
</table>

Table B.2: User-follower and user-friend candidate-centric graph statistics.

Finally, we report additional graph statistics for user-follower $G_f$ and user-friend $G_r$ sampled graphs only such as: diameter $- G_f = 38, G_r = 22$, average degree $- G_f = 0.949, G_r = 2.123$, average path length $- G_f = 13,584, G_r = 9,117$, weakly connected components $- G_f = 653, G_r = 161$. 
Figure B.4: Distribution retweets and user mentions in the cand-centric graph.
Figure B.5: Distribution of replies and hashtags in the candidate-centric graph.
Figure B.6: True vs. sampled follower (top) and friend (bottom) $G_{cand}$ graphs. The overlapping vertices on the outer circle can be further scaled to the balanced stars centered in user vertices with out-degree $k = 10$. 
Appendix C

Querying Twitter API: Rate Limits

In order to query Twitter one should register and create a developer account at [https://dev.twitter.com/](https://dev.twitter.com/) then get a unique access key and token for further authorization. Using the access key and token one can write an application that allows to either access the 1% Twitter firehouse or to query API to get specific:

- user timelines (with up to 3200 tweets; 200 tweet per query) using specific user IDs or usernames;
- tweet objects using specific tweet IDs;
- lists of friend IDs or follower IDs (5000 per query) using specific userIDs.

As a result of each query Twitter API returns a JSON object with (a) user-specific fields\[1\] e.g., user name, user id, the number of friends/followers/statuses, profile pic-

\[1\]User JSON object: [https://dev.twitter.com/overview/api/users](https://dev.twitter.com/overview/api/users)
APPENDIX C. QUERYING TWITTER API: RATE LIMITS

| Title                    | Resource family | Requests / 15-min window (user auth) | Requests / 15-min window (app auth) |
|--------------------------|-----------------|--------------------------------------|____________________________________|
| GET followers/ids        | followers       | 15                                   | 15                                   |
| GET friends/ids          | friends         | 15                                   | 15                                   |
| GET lists/statuses       | lists           | 180                                  | 180                                  |
| GET statuses/user_timeline | statuses        | 180                                  | 300                                  |

Table C.1: Twitter rate limits.

ture, language, profile creation time etc. or (b) tweet-specific fields[e.g., tweet id, tweet, tweet creation time, coordinates, entities (hashtags, user mentions) etc.

Twitter policy restricts to sharing only tweet IDs or user IDs rather than complete tweets or user profiles. Therefore, in order to build datasets used in this thesis, we queried Twitter API to get user and neighbor tweets using their tweet IDs (GET statuses/user_timeline and GET lists/statuses); to get lists of neighbors of different types using user IDs (GET followers/ids and GET friends/ids). In Table C.1, we outline Twitter API rate limits [https://dev.twitter.com/rest/public/rate-limits](https://dev.twitter.com/rest/public/rate-limits).

---

2Tweet JSON object: [https://dev.twitter.com/overview/api/tweets](https://dev.twitter.com/overview/api/tweets)
Appendix D

A Note on Privacy and Ethics

Although research on social media predictive analytics can help to answer a variety of important social science questions (Scharff, 2010; R. E. Wilson, Gosling, & Graham, 2012; Sloan et al., 2013) and has many useful applications as discussed in Chapter 1, the ethics of using social media data is a subject of active debate (Bassett & O’Riordan, 2002; Zimmer, 2010; Darcy & Young, 2012; O’Connor, 2013). For example, a recent study claimed to alter the emotions of Facebook users by embedding content into their profiles without obtaining the users’ explicit and informed consent (Coviello et al., 2014; Kramer, Guillory, & Hancock, 2014). Both the public and the media were highly critical of this study, despite the fact that Facebook users implicitly agree to their personal data being used for research purposes when they accept Facebook’s terms and conditions and sign up or an account.

Interestingly, a small scale study conducted by (NatSec, 2014; Williams et al. 2013).
APPENDIX D. A NOTE ON PRIVACY AND ETHICS

(2013) found that 73% of social media users perceived their posts could be used without their explicit consent, and 82% were “not at all concerned” or “slightly concerned” about researches using their social media profiles.

The Economic and Social Research Council (ESRC) Framework for Research Ethics states that it is not practically possible to seek informed consent from Twitter users in “big data” research. Twitter’s Terms of Service (TToS) require users to provide their consent for Twitter to share with third parties any content they post. We believe that researches conducting experiments on social media data must follow not only TToS but also a set of moral and ethical procedures inscribed in their research workflow (Parker, Dutton, Goldin, & Jeffreys, 2010). For instance, researchers should, report their findings on an abstract and aggregate level, removing user personal information when demonstrating their example posts, as we have in this work.

The 2014 and 2015 workshops on Fairness, Accountability and Transparency in Machine Learning shared practical techniques and experiences relating to questions of privacy and ethics. Other researchers have addressed issues of privacy for automated profiling (Schermer, 2011; Pope & Sydor, 2011), transparent models for personalization (El-Arini, Paquet, Herbrich, Van Gael, & Agüera y Arcas, 2012), data anonymization (Hajian & Domingo-Ferrer, 2012) and fairness for data mining (Mascetti, Ricci, & Ruggieri, 2014).

1http://www.esrc.ac.uk/funding/guidance-for-applicants/research-ethics/
APPENDIX D. A NOTE ON PRIVACY AND ETHICS

While these techniques might be sufficient in the short term, we believe that further regulations are necessary for the protection of all parties.
Appendix E

Visualizing Performance for The Most Predictive Neighborhoods

In Figures E.1 – E.6 we present detailed classification results for gender, age and political preference classification using:

- neighbor or user-neighbor models with different number of neighbors per user and different number of tweets per neighbor as described in Chapter 3 section 3.3;
- binary vs. count-based word unigram features;
- different neighborhood types e.g., friends, retweets etc.

These figures demonstrate how accuracy changes for each combination of features, neighbors, the number of neighbors per user and the number of tweets per neighbor.
APPENDIX E. VISUALIZING PERFORMANCE FOR THE MOST PREDICTIVE NEIGHBORHOODS

Figure E.1: Detailed age classification results obtained from our neighborhood model trained using binary (on the left) vs. count-based features (on the right) for friend, follower, and retweet neighborhoods.
APPENDIX E. VISUALIZING PERFORMANCE FOR THE MOST PREDICTIVE NEIGHBORHOODS

Figure E.2: Detailed age classification results obtained from our user-neighbor model trained using binary for friend and follower neighbors (on the left) vs. count-based features (on the right), and count-based features for retweet and user mention neighborhoods.
APPENDIX E. VISUALIZING PERFORMANCE FOR THE MOST PREDICTIVE NEIGHBORHOODS

Figure E.3: **Detailed gender classification results obtained from our neighbor model** trained using binary (on the left) vs. count-based features (on the right) for friend, follower, retweet, and user mention neighbors.
APPENDIX E. VISUALIZING PERFORMANCE FOR THE MOST PREDICTIVE NEIGHBORHOODS

Figure E.4: Detailed gender classification results obtained from our user-neighbor model trained using binary (on the left) vs. count-based features (on the right) for follower, retweet, and user mention neighbors.
APPENDIX E. VISUALIZING PERFORMANCE FOR THE MOST PREDICTIVE NEIGHBORHOODS

Figure E.5: Detailed political classification results obtained from our neighbor model trained using binary (on the left) vs. count-based features (on the right) for friend, retweet, and user mention neighbors.
APPENDIX E. VISUALIZING PERFORMANCE FOR THE MOST PREDICTIVE NEIGHBORHOODS

Figure E.6: Complete political preference classification results obtained from our user-neighbor model trained using binary (on the left) vs. count-based features (on the right) for friends, retweets, and user mentions.
Appendix F

Annotator Rationales for Gender, Age and Political Preference

*Annotator rationales* – targeted annotator feedback regarding why and how they chose a particular annotation collected via crowdsourcing. We collected rationale annotations from a set of user profiles with known to us (but not known to the annotators) gender, age and political preference labels and used these “gold-standard” labels as a quality control check during crowdsourcing. In Chapters 5 and 6 we demonstrated the benefits of incorporating domain knowledge into predictive models using annotator rationales. We showed that rationale filtering during iterative retraining and active learning significantly improves classification of results.

We present the word clouds of annotator rationales for gender, age and political preference prediction in Figures F.1 – F.3. The size of the rationale corresponds
APPENDIX F. ANNOTATOR RATIONALES FOR GENDER, AGE AND
POLITICAL PREFERENCE

to the frequency of being highlighted by multiple annotators. The complete lists
of all annotator rationales with the annotator confidence level [-2 . . .2] (-2 is the
least and 2 is the most confident) in English and Spanish are available at http://
www.cs.jhu.edu/~svitlana/rationales.html

(a) Male rationales

(b) Female rationales

Figure F.1: Gender rationales (word ngrams of size ≤ 3 and frequency ≥ 2). The size
of each rationale reflects the frequency of being highlighted.
APPENDIX F. ANNOTATOR RATIONALES FOR GENDER, AGE AND POLITICAL PREFERENCE

Figure F.2: Age rationales (word ngrams of size $\leq 3$ and frequency $\geq 2$). The size of each rationale reflects the frequency of being highlighted.

(a) Teen rationales (14 - 19 y.o)

(b) Young rationales (20 - 34 y.o.)
APPENDIX F. ANNOTATOR RATIONALES FOR GENDER, AGE AND POLITICAL PREFERENCE

Figure F.3: The 150 most frequently highlighted political preference rationales for each party.
Appendix G

Emotion Hashtags

Here we present the list of emotion hashtag synonyms we used to annotate tweets with six Ekman’s emotions – joy, anger, sadness, fear, disgust, surprise and fear via distant supervision.

- **ANGER**: #rage, #vexation, #exasperation, #displeasure, #crossness, #irritation, #irritability, #indignation, #pique, #annoyance, #fury, #wrath, #ire, #outrage, #irascibility, #illtemper, #slowburn, #aggravation, #choler, #acrimony, #animosity, #antagonism, #enmity, #hatred, #impatience, #resentment, #temper, #violence, #chagrin, #conniption, #dander, #disapprobation, #distemper, #gall, #huff, #infuriation, #miff, #peevedness, #petulance, #rankling, #soreness, #stew, #storm, #tantrum, #tiff, #umbrage, #blowup, #aggravated, #annoyed, #furried, #illtempered, #impatient, #irritated, #outraged, #raged, #violent
APPENDIX G. EMOTION HASHTAGS

• JOY: #delight, #pleasure, #joyfulness, #jubilation, #triumph, #exultation, #rejoicing, #happiness, #gladness, #glee, #exhilaration, #exuberance, #elation, #euphoria, #bliss, #ecstasy, #rapture, #enjoyment, #felicity, #jouissance, #jocundity, #amusement, #charm, #cheer, #comfort, #pride, #satisfaction, #wonder, #alleviation, #animation, #delectation, #diversion, #exulting, #festivity, #frolic, #fruition, #gaiety, #gem, #gratification, #hilarity, #indulgence, #jewel, #jubilation, #liveliness, #luxury, #merriment, #mirth, #prize, #ravishment, #refreshment, #revelry, #solace, #regale, #bang, #belonging, #bonheur, #buoyancy, #carefreeness, #cheerfulness, #amused, #cheered, #proud, #satisfied, #delighted, #happy, #glad

• DISGUST: #revulsion, #repugnance, #aversion, #distaste, #nausea, #abhorrence, #loathing, #detestation, #odium, #horror, #contempt, #outrage, #antipathy, #dislike, #hatred, #abomination, #hatefulness, #objection, #revolt, #satiation, #satiety, #sickness, #surfeit, #nauseation, #nauseousness, #abashment, #chagrin, #confusion, #conscience, #discomfiture, #embarrassment, #self-consciousness, #self-disgust, #shame, #shamefacedness, #outraged, #shamed, #sick

• FEAR: #terror, #fright, #fearfulness, #horror, #alarm, #panic, #agitation, #trepidation, #dread, #consternation, #dismay, #distress, #anxiety, #worry, #angst, #unease, #uneasiness, #apprehension, #apprehensiveness, #nervousness, #nerves, #perturbation, #foreboding, #shivers, #willies, #heebie-jeebies, #jitteriness, #twitch-
APPENDIX G. EMOTION HASHTAGS

iness, #phobia, #aversion, #antipathy, #bugbear, #nightmare, #neurosis, #concern, #despair, #doubt, #jitters, #scare, #suspicion, #abhorrence, #awe, #cowardice, #creeps, #discomposure, #disquietude, #faintheartedness, #funk, #misingiving, #presentiment, #qualm, #reverence, #revulsion, #timidity, #trembling, #tremor, #chickenheartedness, #coldfeet, #coldsweat, #recreancy, #hysteria, #ststagefright, #chill, #diffidence, #intimidation, #hesitance, #susnsense, #unconcern, #presage, #shadow, #shyness, #agitated, #anxious, #nervous, #scared, #shivering, #suspended, #worried

• SADNESS: #anguish, #attrition, #blahs, #bleakness, #bluedevils, #blue-devils, #bluefunk, #brokenheart, #bummer, #cheerlessness, #compunction, #dejection, #demoralization, #depressed, #depression, #desolation, #despair, #despaired, #de-spondency, #disconsolateness, #dismals, #disspiritedness, #distress, #dolefulness, #dolor, #downcastness, #downer, #downheartedness, #dysphoria, #forlornness, #funk, #gloom, #gloomed, #gloominess, #grief, #grieving, #guilt, #heartache, #heartbreak, #heavy heart, #heavyheartedness, #helplessness, #hopelessness, #joylessness, #letdown, #listlessness, #lost-sorrow, #melancholy, #misery, #moodiness, #moody, #mopes, #mournfulness, #mourning, #oppression, #plaintiveness, #poignancy, #regret, #repentance, #self-pity, #sorrow, #sorrowfulness, #the blues, #the dumps, #tribulation, #unhappiness, #unhappy, #weepiness, #weight, #woe, #world-weariness, #wretchedness
APPENDIX G. EMOTION HASHTAGS

- **SURPRISE**: #abruptness, #amazement, #astonishment, #astoundment, #awe, #bewilderment, #bolt from the blue, #bombshell, #consternation, #curiosity, #curve-ball, #disbelief, #disillusion, #epiphany, #eureka, #eye-opener, #fortune, #god-send, #incredulity, #jolt, #kick, #marvel, #miracle, #miscalculation, #phenomenon, #portent, #precipitance, #precipitation, #precipitousness, #prodigy, #rarity, #revelation, #rude awakening, #shock, #shocked, #shocker, #stupefaction, #suddenly, #suddenness, #surprised, #thunderbolt, #unexpected, #unforeseen, #wakeupcall, #whammy, #wonder, #wonderment, #astonished, #amazed
Bibliography


Bibliography

*LREC* (p. 2200-2204).


Bibliography


Bibliography


Bibliography


for German. In Proceedings of WASSA (p. 7-13).


Bibliography


Bibliography


Bibliography

Advances in Information Retrieval, 5478, 202-215.


Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using
Bibliography

*Distant Supervision* (Tech. Rep.).


Bibliography

*National Academy of Sciences, 101*(suppl 1), 5228–5235.


Bibliography


Bibliography


McNeal, G. (2014). Facebook manipulated user news feeds to create emotional
Bibliography

responses. *Forbes*.


Bibliography


328


Bibliography

spam by any stretch of the imagination. In *Proceedings of ACL* (pp. 309–319).


Paul, & Girju, R. (2009). Cross-cultural analysis of blogs and forums with mixed-
Bibliography

Bibliography

with application to Twitter. In *Proceedings of NAACL*.


Bibliography


Bibliography


Bibliography


Thelwall, M., Buckley, K., & Paltoglou, G. (2012, January). Sentiment strength detec-
Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography

*ICWSM.*


Vita

Svitlana Volkova grew up in Ukraine. In 2004 and 2006, she received a bachelor’s and a master’s degrees in Intellectual Systems for Decision Making from Petro Mohyla Black Sea State University. In 2008, Svitlana was awarded a Fulbright Graduate Student Scholarship to study in the USA. She was admitted to Kansas State University in 2008 and received her master’s degree in Computer Science in 2010.

In 2010, Svitlana joined the Johns Hopkins University Computer Science Department as a Ph.D. student at the Center for Language and Speech Processing. Svitlana’s research interests lie in natural language processing, machine learning and social media analytics. She is a recipient of the Google Anita Borg Scholarship. During her Ph.D., Svitlana interned at the Microsoft Research Natural Language Processing group in 2011, 2012 and at the Microsoft Research Machine Learning Group in 2014. Svitlana is now a research scientist at Pacific Northwest National Laboratory.