

COMPUTATIONAL APPROACHES TO UNDERSTANDING
STYLISTIC VARIATION IN ONLINE WRITING

A Dissertation
Presented to
The Academic Faculty

By

Umashanthi Pavalanathan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology

December 2018

COMPUTATIONAL APPROACHES TO UNDERSTANDING
STYLISTIC VARIATION IN ONLINE WRITING

Approved by:

Dr. Jacob Eisenstein, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology

Dr. Eric Gilbert
School of Interactive Computing
Georgia Institute of Technology

Dr. Michael Gamon
Knowledge Technologies Group
Microsoft Research

Dr. Scott Kiesling
Department of Linguistics
University of Pittsburgh

Date Approved: August 21, 2018

To My Parents.

ACKNOWLEDGEMENTS

The research presented in this dissertation would not have been possible without the help, support, and contributions of many people during the last five years. I would like to take this opportunity to express my deepest gratitude to everyone involved. I would like to start by thanking my advisor Jacob Eisenstein for giving me the opportunity to work in his research group and all the support he has given throughout my time at Georgia Tech. Jacob guided and oriented my PhD research with his vision, inspired me with his boundless energy and expertise in several fields, and taught me to think big. I am greatly indebted to him for all his advise and support. Thank you for taking a chance on me, for believing in my work, and for being available always—I could not have asked for a better advisor.

I have been very fortunate in having had a fantastic thesis committee. Many thanks to Munmun De Choudhury, Eric Gilbert, Michael Gamon, and Scott Kiesling. Munmun has been very supportive, resourceful, and inspiring since our collaboration during the early days of my PhD. Eric has always been inspiring and I am very fortunate to have you on my thesis committee and to collaborate on some projects. Micheal has been very supportive since I first emailed him to be on my committee and provided helpful feedback on my thesis work. Special thanks to Micheal for being very flexible in adjusting your sabbatical plans to join my defense. Thanks to Scott for all the support during our collaboration past few years, for your inspiring work in Sociolinguistics, and for serving on my committee. Thanks to Kathleen McKeown for providing feedback on my thesis work.

In addition to my thesis committee, I am thankful for several other faculty and staff at Georgia Tech. My special thanks to Annie Anton and Amy Bruckman for being supportive in several ways during the past five years. Thanks to Jane Chisholm and Kurt Belgum for helping me improving my academic writing skills. Thanks to Cynthia

Bryant, Jessica Celestine, Carolyn Daley-Foster, Renee Jamieson, and Cynthia Jordan for all the help with administrative matters.

This thesis would not have been possible without the help of many amazing people I worked with during my PhD years. My thesis work is the direct result of collaborations with Xiaochuang Han, Jim Fitzpatrick, Scott Kiesling, and my advisor Jacob Eisenstein. Thanks to Adam Glynn for allowing me to take your causal inference class at Emory University and for your time brainstorming ideas, which has been instrumental for most of my thesis work. Thanks to Tyler Schnoebelen for the feedback and discussions on multiple projects. I am also thankful to my mentors and collaborators during PhD internships. Thanks to Lada Adamic, Kai-Wei Chang, Jennifer Chayes, Courtney Corley, Bistra Dilkina, Amaç Herdağdelen, Hoifung Poon, and Ellen Zegura for your mentoring and for shaping my research skills in several ways.

I am indebted to all my teachers at Chundikuli Girls' College and University of Moratuwa. My sincere thanks to Chandana Gamage, Shahani Markus, Vishaka Nanayakkara, Beth Plale, Louiqa Raschid, Samitha Samaranayake, and Sanjiva Weerawarana, all of whom were instrumental in several ways to make graduate school possible. Thanks to my undergraduate friends Keheliya Gallaba and Pivithuru Wijegunawardana for making me think of graduate school. I am very glad all of us get to experience graduate school.

Thanks to all members of the computational linguistics lab for your continuous help, support, and for making life in the lab enjoyable. To Yangfeng Ji, Yi Yang, Ana Smith, Ian Stewart, Sandeep Soni, Yuval Pinter, and Sarah Wiegrefe, thank you for being such awesome labmates. Thanks to Dong Nguyen for making your time at Georgia Tech so memorable, not only for our research discussions, but also for the time we spent exploring Atlanta. I am also very thankful for all the members of the social computing research groups at Georgia Tech. My special thanks to Catherine

Grevet, Chaya Hiruncharoenvate, and Tanushree Mitra for all your help, support, and encouragement since I first visited Georgia Tech for PhD recruitment weekend. Your support was invaluable, especially when I needed the most during the early days of graduate school.

It would be impossible to mention all of my other friends in Atlanta, Bloomington, and elsewhere who supported me throughout the years, but I thank Pasanthi, Thakshila, Jinath, Krisha, Priyastha, and Shuaishuai, in particular for being part of the ups and downs during the PhD years. I mention Tanushree and Shauvik separately for their support throughout the years and for being there during the challenging times. Specials thanks to Alicia, Kenechi, and Nan for being wonderful roommates during my stay in Atlanta.

Finally, I would like to thank my parents, my sister, and my brother, without whom none of this would have been possible. Thank you for all your love, support, and countless sacrifices through the years.

TABLE OF CONTENTS

Acknowledgments	v
List of Tables	xiv
List of Figures	xviii
Chapter 1: Introduction	1
1.1 Linguistic Style Variation	3
1.2 Computational Approaches to Understanding Social Meaning of Stylistic Variation	6
1.3 Comparison to Existing Approaches to Linguistic Style Variation . . .	8
1.4 Implications for Natural Language Processing, Social Media Analysis, and Sociolinguistics	10
1.5 Thesis Statement	11
1.6 Contributions	11
Chapter 2: Background and Related Work	13
2.1 Linguistic Style Innovations: Non-standard Orthography	14
2.1.1 Emoticons	15
2.1.2 Emojis	15
2.2 Inter-person Variation: Variation and Social Variables	18

2.3	Intra-person Variation: Variation and Social Interactions	19
2.3.1	Linguistic Style-Shifting	19
2.3.2	Linguistic Style-shifting in Computer Mediated Communication	20
2.3.3	Linguistic Style and Interactional Meaning	23
2.4	Community Norms: Variation and Multi-Communities	24
2.4.1	Multi-community Variation	24
2.4.2	Online Community Norms	25
2.4.3	Writing Style Norms	26
2.4.4	Enforcing Norms	27
2.4.5	Effects of Norm Enforcement	28
2.5	Causal Inference for Social Media Analysis	29
Chapter 3: Confounds and Consequences of Using Social Media Data to Study Online Writing¹		31
3.1	Background	31
3.2	Data	32
3.3	Representativeness of Geotagged Twitter Data	33
3.4	Impact on Linguistic Generalizations	37
3.5	Impact on Text-based Geolocation	41
3.6	Conclusions	43
Chapter 4: Linguistic Style Innovations and Intra-Person Variation in Social Media²		45
4.1	Background	46
4.2	The Social Environment of Twitter	47

4.3	Linguistic Style Innovations	48
4.3.1	Geography-specific lexical innovations	49
4.3.2	<i>Tweetspeak</i> variables	50
4.4	Data	50
4.4.1	Building a balanced corpus	51
4.4.2	Geolocating tweet recipients	52
4.5	Analysis and Results	52
4.5.1	Model-I	53
4.5.2	Model-II	54
4.6	Conclusions	55
 Chapter 5: Linguistic Innovations and the Competition for Paralinguistic Functions³		58
5.1	Background	59
5.2	Dataset	60
5.3	Study Design	62
5.3.1	Confounds	63
5.3.2	Matching	64
5.3.3	Estimation of Treatment Effects	64
5.4	Causal Inference Analysis	66
5.4.1	Analysis-I: Effects of Emoji Adoption on Emoticon Usage	66
5.4.2	Analysis-II: Effects of Emoji Adoption in Standard Word Usage	67
5.5	Conclusions	69

Chapter 6: Multi-Community Style Variation: A Multidimensional Lexicon for Interpersonal Stancetaking⁴	72
6.1 Background	73
6.2 Data	75
6.3 Stance Lexicon	76
6.3.1 Seed lexicon	76
6.3.2 Lexicon expansion	76
6.4 Linguistic Dimensions of Stancetaking	77
6.4.1 Extracting Stance Dimensions	78
6.4.2 Results: Stance Dimensions	79
6.5 Intrinsic Validation	80
6.5.1 Word Intrusion Task	81
6.5.2 Pre-registered Hypotheses	82
6.6 Extrinsic Evaluation	84
6.6.1 Predicting Cross-posting in Subreddits	84
6.6.2 Stancetaking Dimensions and Social Phenomena	86
6.7 Conclusions	88
Chapter 7: Multi-Community Style Variation: Linguistic Style-Shifting	90
7.1 Background	91
7.2 Dataset	92
7.3 Methods	93
7.3.1 Measuring Linguistic Style	93
7.3.2 Style-shifting Models	94

7.3.3	Model Implementation	96
7.3.4	Model Evaluation	96
7.4	Results	97
7.4.1	RQ1: Style-shifting	98
7.4.2	RQ2: Style-shifting and Subreddit Attributes	98
7.4.3	RQ3: Style-shifting and Style Persistence	99
7.5	Discussion and Future Work	99
Chapter 8: Effects of Norm Enforcement on Online Writing Style⁵ .		101
8.1	Background	102
8.1.1	Norm Enforcement in Wikipedia	103
8.1.2	Motivation for Participation and Editor Roles in Wikipedia . .	104
8.1.3	Research Questions	105
8.1.4	Terminology	105
8.2	Datasets	106
8.2.1	Wikipedia NPOV Corpus	107
8.2.2	Identifying NPOV Tagging and Removal	108
8.2.3	Identifying Treatment Editors	109
8.2.4	Dataset for Article-Level Analysis	110
8.2.5	Dataset for Editor-Level Analysis	111
8.3	Methods	111
8.3.1	Identifying Biased Language	112
8.3.2	Measuring the Effect of Treatment on Biased Language Usage	113

8.4	RQ1: Article-Level Effects of NPOV Tagging	115
8.5	RQ2: Editor-Level Effects of NPOV Correction	117
8.5.1	Editor Writing Style: RQ2a	119
8.5.2	Editor Engagement: RQ2b	121
8.6	Discussion	125
8.6.1	RQ1: Article-level Effects of NPOV Tagging	125
8.6.2	RQ2: Editor-level Effects of NPOV Correction	126
8.6.3	Possible Reasons for the Differences in the Effectiveness of Norm Enforcement	127
8.6.4	Implications for Wikipedia and Online Language Moderation .	128
8.6.5	Limitations and Future Work	129
8.7	Conclusions	131
Chapter 9: Conclusion and Future Work		133
Appendix A: Linguistic Variables		138
A.1	Example tweets with linguistic variables	138
A.2	Lexical variable frequencies for each MSA	139
Appendix B: Emoticon Tokens		140
Appendix C: Robustness Check for Article-Level effects of NPOV Tagging		141
References		164

LIST OF TABLES

3.1	L1 distance between county-level population and Twitter users and messages	36
3.2	Demographic statistics for each dataset	37
3.3	Most characteristic words for demographic subsets of each city, as compared with the overall average word distribution	41
4.1	List of geographical linguistic variable from each ten most populous metropolitan statistical areas (MSAs) in the United States.	50
4.2	List of Tweetspeak lexical variables.	51
4.3	Predictors used in each model.	53
4.4	Results for Model-I predictors and geographical lexical variables. Statistical significance is indicated with asterisks, *** : $p < 0.01$, ** : $p < 0.05$, * : $p > 0.05$. ‘Weight’ is the logistic transformation of the logistic regression coefficient, yielding a value between 0 and 1, with 0.5 indicating indifference. ‘Empirical %’ indicates the percentage of messages with each predictor which include a non-standard variable, and ‘N’ is the total number of messages with each predictor.	54
4.5	Results for Model-I predictors and Tweetspeak lexical variables.	55
4.6	Model-II predictors and geographical lexical variables.	56
4.7	Model-II predictors and geographical lexical variables.	56
5.1	Standard word list statistics	62
5.2	Causal inference results	70

6.1	Dataset size	75
6.2	Stance lexicon: seed and expanded terms.	77
6.3	For each of the six dimensions extracted by our method, we show the five words with the highest loadings.	80
6.4	Results for pre-registered hypothesis that stance dimensions will not split synonym pairs.	83
6.5	Results for preregistered hypothesis that stance dimensions will align with stance feature groups of Biber and Finegan [1].	84
6.6	Examples of subreddit pairs that have large and small amount of overlap of contributing members.	85
6.7	Accuracy for prediction of subreddit cross-participation.	86
7.1	Model evaluation results. We used the multidimensional stance lexicon and LIWC to characterize linguistic style. <i>Diff(acc.)</i> : difference in expected predictive accuracy of the two models, <i>Std. Error</i> : standard error of the difference. When comparing two models, <i>A</i> vs. <i>B</i> , a negative value for <i>Diff(acc.)</i> indicates that model <i>A</i> performs better than model <i>B</i>	97
7.2	List of subreddits that enable both high and low style-shifting, ranked based on the α_r parameters of \mathbf{M}_2 . Low α_r indicates more subreddit influence (i.e., more style-shifting) while high α_r indicates less style-shifting.	98
7.3	Spearman ranked correlation between subreddit style shifting parameter α_r and measures of subreddit moderation.	99
8.1	List of Lexicons Used to Characterize Bias Language.	114
8.2	Results of RQ1 Analysis. Article lexicon coverage is computed for the textual content of the article results from each revision. Coefficient β_3 indicates the change in slope after treatment. Statistical significance after correcting for multiple comparisons using Benjamini and Hochberg [2] adjustment for false discovery rate are shown; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Percentage change is computed as the change in post-treatment lexicon coverage after 40 revisions.	117

8.3	Results of RQ2 Analysis after controlling for editor experience and talk page discussion during treatment. Editor lexicon coverage is computed for their textual contribution in each revision. Coefficient β_2 indicates the change in level after treatment; \mathbf{d} indicates whether there was any talk page discussion during treatment (reference: <i>discussed</i> = 0); \mathbf{e} indicates whether the editor is considered experienced or not (reference: <i>experienced</i> = 0); \mathbf{x}_t is the indicator variable for post-treatment; $\mathbf{d} : \mathbf{x}_t$ and $\mathbf{e} : \mathbf{x}_t$ are the interaction terms. Statistical significance after correcting for multiple comparisons using Benjamini and Hochberg [2] adjustment for false discovery rate are shown; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Percentage change is computed as the level change following the treatment. Note that none of the coefficients for the $\mathbf{e} : \mathbf{x}_t$ interaction term are significant at $p < 0.05$	123
8.4	Results of RQ2b Analysis: predictor coefficients of the count model of the zero-inflated negative binomial regression. The dependent variable, editor engagement is computed as number of edits during each of the three-day period. We considered 60 days prior to the treatment and 60 days after the treatment. <i>Post-treatment</i> is the indicator for whether the observation is post (=1) or pre (=0) treatment. <i>Timebin</i> is the three-day window; we consider 20 timebin windows pre- and post-treatment, which includes 120 days in total. <i>Discussion</i> is the indicator whether there was talk page discussion between correcting and corrected editors during treatment (<i>Discussion</i> = 1) or not. <i>Experienced</i> is the indicator for whether the editor is experienced (>250 edits, <i>Experienced</i> = 1) or not. Interaction terms include: Post-treatment : Discussion(=1) and Post-treatment : Experienced(=1). Statistical significance after correcting for multiple comparisons using Benjamini and Hochberg [2] adjustment for false discovery rate are shown; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$	124
A.1	Examples of each of the linguistic variables in bold, with glosses in parentheses.	138
B.1	Top-twenty extracted emoticon tokens and their cumulative frequencies.	140

C.1 Results of RQ1 Analysis robustness check on the subset of data after removing articles which are tagged for more than three years. Article lexicon coverage is computed for the textual content of the article results from each revision. Coefficient β_3 indicates the change in slope after treatment. Statistical significance after correcting for multiple comparisons using Benjamini and Hochberg [2] adjustment for false discovery rate are shown; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Percentage change is computed as the change in post-treatment lexicon coverage after 40 revisions. 142

LIST OF FIGURES

1.1	Axes of online writing variation considered in this dissertation.	8
3.1	Proportion of census population, Twitter messages, and Twitter user accounts, by county. New York is shown on the left, Atlanta on the right.	35
3.2	User counts by number of Twitter messages	37
3.3	Difference in age probability distributions between GPS-MSA-BALANCED and LOC-MSA-BALANCED.	38
3.4	Aggregate statistics for geographically-specific non-standard words and entity names across imputed demographic groups, from the GPS-MSA-BALANCED sample.	40
3.5	Classification accuracies	43
4.1	Affordances offered by Twitter for influencing the scope of the audience.	49
5.1	Examples of emoji characters used in Twitter (created using http://www.iemoji.com).	60
5.2	Selection of authors into treatment and control groups.	63
5.3	Matching procedure.	64
5.4	Analysis-I: Emoji usage of treatment and control groups.	66
5.5	Analysis-I results: Emoticon usage of treatment and control groups.	67
5.6	Analysis-II: Emoji usage of treatment and control groups.	68
5.7	Analysis-II results: Standard token usage of treatment and control groups.	68

6.1	Mapping of subreddits in the second and third stance dimensions, highlighting especially popular subreddits.	79
6.2	Kernel density distributions for stance dimensions 2 and 5, plotted with respect to annotations of politeness and formality.	88
8.1	An NPOV tag displayed at the top of an article.	103
8.2	Flowchart showing the data and method set up for each research question.	107
8.3	Article level lexicon coverage with respect to the treatment of NPOV tagging. We computed the average lexicon coverage rate of treatment articles' 40 revisions before and after the treatment revision. Lexicons for which we observe a statistically significant drop in slope ($p < 0.05$) are marked with an asterisk (*).	118
8.4	Editor level lexicon coverage with respect to the treatment of NPOV correction. We computed the average lexicon coverage rate of treatment users' textual contributions 40 revisions before and after the treatment revision. Lexicons for which we observe a statistically significant level drop ($p < 0.05$) are marked with an asterisk (*).	122
8.5	Visualization of RQ2b regression coefficients showing the pre- and post-treatment engagement levels for the control factors <i>experienced</i> and <i>discussion</i>	125
A.1	Frequencies of each linguistic variable from each metropolitan statistical area (MSA), with the local frequency in gray and the national frequency in black.	139
C.1	The distribution of time-intervals between the addition of the NPOV tag and the data collection time.	141

ABSTRACT

Language use in online interactions varies from community to community, from individual to individual, and even for individuals in different contexts. While prior work has identified these differences, far less is understood about why these differences have arisen in online writing. My dissertation focuses on this *why* question. The reasons for linguistic diversity in online writing could be multifold. As more and more interpersonal social interactions are conducted through technology-mediated channels, there is an increasing need to express multiple social meanings in varied social situations through linguistic means. In the absence of non-verbal cues, the technology-mediated channels provide several affordances to conduct interpersonal interactions. How do factors that are unique to online writing, such as the need to convey varied social meanings and the affordances in technology-mediated channels, shape online writing? My dissertation investigates this interplay through a series of large-scale computational studies of linguistic style variation in online writing.

Using unsupervised methods and causal statistical analysis, I have investigated the social meaning of varied non-standard language usage in social media and the effects of new technological affordances in online social platforms on individuals' writing style. To quantitatively study community-level stylistic variation at scale, I have developed a multi-dimensional style lexicon using unsupervised techniques and used it to study style-shifting in online multi-communities. Further, I have investigated how writing style norm enforcement in online platforms affects stylistic variation in online writing. My dissertation will advance our understanding of how individuals utilize the affordances in online social platforms and shift style to achieve varied social goals in online interpersonal interactions. Understanding the social dimensions of linguistic style variation in online writing has important consequences for the design of language technology and social computing systems, and beyond.

CHAPTER 1

INTRODUCTION

“All words have the *taste* of a profession, a genre, a tendency, a party, a particular work, a particular person, a generation, an age group, the day and hour. Each word tastes of the context and contexts in which it has lived its socially charged life.” [3]

Online writing¹ is often stylistically distinct from other written genres, such as academic papers, books, and print newspapers [4, 5], but it also displays an immense internal stylistic diversity [6, 7]. Posts from social platforms such as Twitter, Facebook, and Reddit, use a variety of stylistic constructs including capitalization (e.g., *I’m VERY happy*), repetitions (e.g., *cooollll!!!!*), abbreviations (e.g., *ikr*), transcripts of spoken variants (e.g., *jawns*), emoticons (e.g., *:-D*), and emojis (e.g., 😂) [7]. Differences in the usage of these stylistic variables in social media have been shown to align with macro-level social categories such as geographic location [8, 9], age [10, 11, 12], gender [13, 14, 15, 16], and race [17]. While these differences in online writing are incontrovertible, far less is understood about *why* or *how* this stylistic variation arises. The purpose of this dissertation is to shed light on this question using computational methods on large social media corpora, coupled with sociolinguistic theories.

In face-to-face interactions, we make adjustments in our speech when we are talking to different people, about different topics, and in different places. Some of this adjustment is influenced by context [18]. For example, the language we use in a courtroom is very different from the language we use at a gathering of friends, and we speak differently to our boss from the way we do to our peers, even when the topic is

¹By online writing, I refer to the writing in online social platforms such as Twitter, Facebook, and Reddit.

the same. Similar to speech, online writing also happens in a social context that can vary a lot: different addressees or audiences, topics of discussion, social goals of the writing, social group or community norms, etc. So, it is reasonable to believe that the social context plays an important role in linguistic style variation in online writing. But how does the social context influence online writing style? Audience or addressee is a social context, and sociolinguistic investigations found stylistic variation in speech depending on the nature of the audience: whether the audience are direct addressees or overhearers [19]. But, how can we characterize audience in social media? Community or group norm is another social context, and ethnographic studies found stylistic variation depending on group norms: high school students who identify themselves as nerds (as opposed to cool) use formal register in casual speech and avoid popular slang [20, 21]. How can we identify or characterize norms in online communities and study the influence of community norms in online writing variation?

In addition to the social context, the technical affordances in the social platforms in which online interactions take place could also affect the writing style. For example, emojis are introduced to online platforms as a new form of writing [22]. How do these affordances influence the writing style of individuals? An important aspect in which online writing differs from other forms of writing is the possibility of the content being moderated [23]. How does such moderation or norm enforcement impact the writing style?

The questions about how context—such as social situation, technical affordances, and norm enforcement—affects linguistic style variation in online writing are interrelated. This dissertation investigates how these different aspects shape online writing through a series of large scale computational studies using techniques that are more general than what has been used in the past. In the remainder of this chapter, I first provide a brief background on linguistic style variation, and then describe four dimensions along which I study the social meaning of linguistic style variation in

online writing. Then I provide a brief overview of the work presented in the remaining chapters of this dissertation. Following that, I provide a brief comparison to existing approaches to studying stylistic variation and discuss implications of this thesis work. Finally, I present my thesis statement and conclude with a list of contributions of this dissertation.

1.1 Linguistic Style Variation

Stylistic variation refers to altering vocabulary, syntactic structure, or discourse structure based on the context and situation. For example, an individual may speak in a technical register in a professional situation, while they might use a more dialectal feature in a casual social situation [24].

- (1) The lunch served in the cafeteria today was not very appetizing.
- (2) The lunch served in the cafeteria today was nasty.

While the examples (1) and (2) convey the same objective information, the difference in the choice of vocabulary ‘not very appetizing’ and ‘nasty’ indicates a difference in social situation: formal versus informal.

- (3) We don’t expect any help from the government.
- (4) We don’t expect no help from the government.

The two examples (3) and (4) above differ in that (4) includes a negative marker on the object noun phrase (no help) as well as on the verb phrase (don’t expect), whereas (3) avoids the double negative by replacing the noun negator with any. Despite this difference, the two examples convey the same semantic meaning. However, these two examples do convey different social meaning as a direct consequence of the grammatical difference. With its standard forms, (3) is often considered as characteristic of middle-class speech, while (4) is often considered as characteristic of working-class speech [25].

These differences carry sociolinguistic significance and are recognized by the speakers of the language [24].

In early investigations, the study of sociolinguistic variation is often characterized along three principal components: *linguistic* or internal constraints, *social* or inter-speaker constraints, and *stylistic* or intra-speaker constraints [26]. The study of linguistic variation involves constraints on variable speech output, sound change, vowel shifts, and structural relations among regional dialects [26]. The study of social variation involves understanding the relation between variation and social categories such as gender, age, socioeconomic class, ethnicity, and geographic origin. Eckert and Rickford [26] define stylistic variation as any intra-speaker variation that is not directly attributable to performance factors or to factors within the linguistic system, and they propose that the next phase of stylistic studies will have to focus on the highly permeable boundaries among linguistic, social, and stylistic constraints. Motivated by this, I consider the following dimensions to study stylistic variation in online writing: new forms of linguistic variables or linguistic innovations in online writing (parallel to linguistic constraints in speech), inter-person variation, and intra-person variation. Later work in sociolinguistic variation also considered the dimension of speech community, which involves the notion of shared norms or practices [27, 28, 29]. Similar to speech communities, online social platforms have various virtual communities such as Reddit and 4chan. Therefore, I also consider the dimension of community-variation. To summarize, in this dissertation, I study the patterns and social meaning of stylistic variation in online writing along the following four dimensions: (1) stylistic innovations, (2) inter-person variation, (3) intra-person variation, and (4) community (or communicative) norms. Next, I briefly explain each of these dimensions.

Linguistic Innovations: As more and more interpersonal interactions are conducted through technology-mediated channels, social media language is becoming

lexically diverse [4, 5] due to the needs to convey social cues and due to the changing affordances in the socio-technical channels [30, 31]. These factors lead to the emergence of new linguistic variables in online writing [7]. Some of these variables are transcriptions of existing spoken language variables (e.g., *hella* meaning ‘very’ or ‘a lot of’) [5], and others are linguistic innovations created as a result of the need to convey social cues in the new medium (e.g., emoticons, abbreviations such as *lol*, expressive lengthening such as *coooooo!!!!*) [30]. Another form that emerged due to the technological affordances in online platforms is the pictographs called emojis [22]. These lexical innovations are used to convey varied social meanings [32], and my work focuses on the patterns of variation in the usage of lexical variables by individuals in various contexts.

Inter-Person Variation: Individuals have different characteristics in terms of their age, gender, geographic origins, socio-economic status, and cultural preferences. As the rich body of literature in sociolinguistics suggests, these characteristics of individuals are reflected in their language [33]. Such individual traits could also lead to various stylistic choices in online writing, and my work focuses on inter-person variation in adapting linguistic styles.

Intra-Person Variation: Online written interactions take place in various social contexts. For example, an individual may use Twitter to interact with close friends, professional colleagues, and also with a virtual community of Twitter users following an event using a common hashtag [34]. These various interactional contexts have different social goals which require changes in self-representation using linguistic means [19]. My work focuses on intra-person variation to convey varied social meanings in online writing.

Community Norms: Online platforms provide virtual spaces for groups of individuals to get together and discuss about a common topic or activity of interest [35, 36]. When groups of individuals with various traits come together and form communities, as in many online communities, a set of practices or norms emerge in the course of their interactions [37, 38, 39]. These norms are not fixed, but rather they evolve simultaneously with the community’s membership and by the interactions in which the members engage [37, 40]. When individuals participate in various communities, they construct different forms of identities. Eckert argues that a key to this entire process of individual identity construction is the stylistic practices in different communities [41]. My work focuses on linguistic style norms in online communities and investigates how individuals adapt to those linguistic norms.

1.2 Computational Approaches to Understanding Social Meaning of Stylistic Variation

The work presented in Chapters 4-8 focuses on multiple of the dimensions described in the previous section (Figure 1.1). This section provides an outline of each of those chapters. Before presenting my work in linguistic style variation in online writing, in Chapter 3, I address some of the issues related to the appropriateness of using social media data to study sociolinguistic phenomena. Specifically, I present findings from our work [42] on identifying the biases and limitations of using social media data to make generalized conclusions.

Linguistic Style Innovations and Intra-Person Variation (Chapter 4): In this chapter, I investigate intra-person variation in using two types of innovations on Twitter—geography specific variables, and variables that emerged in online writing and are non-standard in other written genres—through the lens of the sociolinguistic concept of audience design. Findings from this work show that Twitter users are

attuned to both the nature of their audience and the social meaning of lexical variation, and they customize their self-presentation accordingly.

Linguistic Innovations and the Competition for Paralinguistic Functions (Chapter 5): In this chapter, I investigate the causal effects of the adoption of a new socio-technical linguistic affordance, emojis, on individuals' writing style on Twitter. Findings from this work suggest that the new linguistic form of emojis competes with prior orthographic means of non-standard writing in conveying varied social meaning.

Multi-Community Style Variation (Chapter 6): In this chapter, I quantitatively operationalize the sociolinguistic concept of interpersonal stancetaking to study intra-person, inter-person, and inter-community variation in the online multi-community platform Reddit. I validate the multidimensional stancetaking lexicon that I developed in this work using intrinsic and extrinsic evaluations, and I demonstrate the utility of the stance lexicon in studying interpersonal social phenomena such as politeness and formality.

Linguistic Style-shifting in Online Multi-Communities (Chapter 7): In this chapter, I investigate the patterns of linguistic style variation in online multi-community environments using the multidimensional stancetaking lexicon to measure linguistic style of individual members and communities on Reddit. Preliminary findings from this work suggest that the users who participate in multiple communities change their writing style depending on in which community they participate.

Effects of Norm Enforcement on Writing Style (Chapter 8): In this chapter, I study the causal effects of moderation on writing style using Wikipedia's neutral point of view (NPOV) tagging as a case study. Findings from this work suggest that the NPOV writing style norm enforcement helps Wikipedia articles to converge to a desired neutral language style. However, for individual editors, NPOV corrections and talk page discussions yield no significant change in their usage non-neutral words, including Wikipedia's own list of "words to watch."

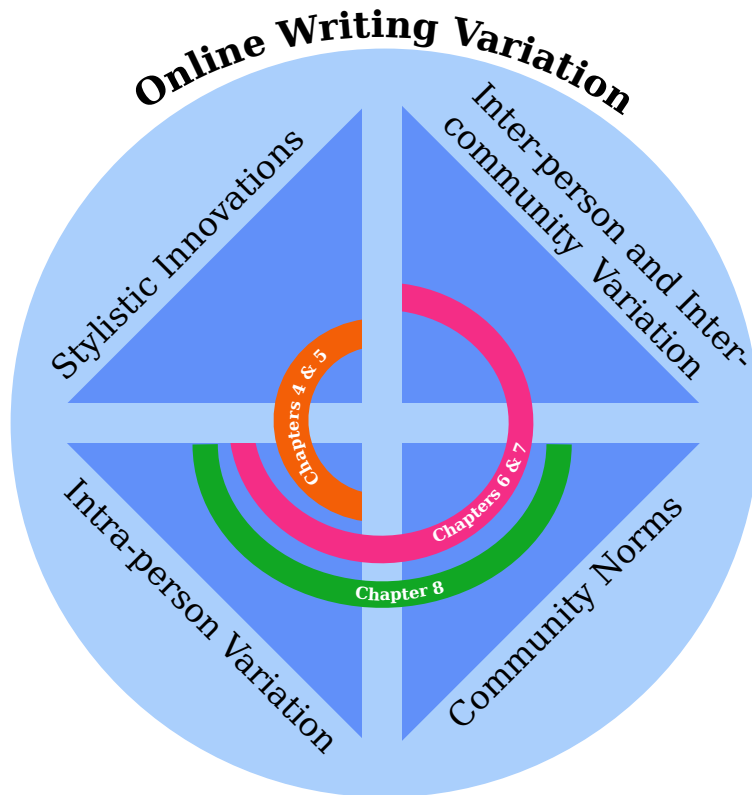


Figure 1.1: Axes of online writing variation considered in this dissertation.

1.3 Comparison to Existing Approaches to Linguistic Style Variation

Data-driven methods to extract and construct linguistic measures. An important aspect in which the work presented in this dissertation differs from prior work in stylistic variation, both in sociolinguistics and computational linguistics, is the selection of linguistic measures to study variation and social phenomena. Studies in sociolinguistics and computer supported communication (CMC) primarily use a small number of hand-chosen linguistic variables to investigate stylistic variation [e.g., 43, 44, 45]. Linguistic variables used in computational studies mainly consist of lists of function words [46] or part-of-speech (POS) tags [47, 48]. It is less intuitive how much of *linguistic style* aspects are captured in such variables. In contrast, I use a data-driven approach to automatically extract large number of linguistic variables:

I first use corpus frequency based automated approaches that are specialized for linguistic variable extraction [e.g., 17, 49], and then I use qualitative filtering on the extracted list of variables to eliminate irrelevant terms (e.g., work presented in Chapter 4). Further, prior work focused on predicting social phenomena based on varied language (e.g., politeness [50], formality [51]) uses crowdsourcing to obtain gold annotations and build large datasets for supervised predictions. These annotation efforts draw on the annotators' own intuitions about the meaning of the sociolinguistic constructs. In contrast, I use unsupervised techniques to extract dimensions based on the underlying linguistic functions performed through groups of variables (e.g., work presented in Chapter 6).

Causal inference for the study of linguistic style variation. Prior work on linguistic predictors of online social phenomena, such as power imbalances [52, 53, 54], politeness [50], and formality [51], identifies and tests a wide range of sociolinguistic correlations. However, correlations alone cannot present the complete picture of the social meaning of associated linguistic variables [44, 55]. A major challenge in studying the social meaning of online stylistic variation is isolating the causal role of varied interactional context—an issue that has received relatively little consideration in prior computational studies of sociolinguistic phenomena, but is crucial when using observational data from online platforms. In my work, I use carefully constructed datasets and causal inference techniques to eliminate possible confounds, and I study the causal role of social context in conveying varied social meaning through multiple linguistic styles (e.g., work presented in Chapter 4, Chapter 5, and Chapter 8).

1.4 Implications for Natural Language Processing, Social Media Analysis, and Sociolinguistics

Understanding the social dimensions of linguistic style differences has implications for the design of language technology and social computing systems, and beyond. While the computational linguistics community has made great progress in understanding the objective meaning in language [56], understanding language in social context still remains as a challenge in automated processing of natural language. Traditional natural language processing (NLP) tools consider language as static and these tools fail to account for how language varies across social context and how language depends on individuals' attitudes [57, 58, 59]. Hence these tools are brittle to variation in online writing, particularly to non-standard language and informal social context [5]. Understanding the social dimensions of language variation will give insights into how we can adapt these systems and make them more robust to language variation (e.g., [60]).

Existing tools to automatically extract insights from social data to answer questions, for example in social sciences and public health, primarily consider the topical aspect of textual content (e.g., health, body, and work sub-categories in the psychological processes category of LIWC (Linguistic Inquiry Word Count; [61])). While the topical aspect conveys important information, the stylistic aspect of language also conveys additional information. Popular tools, such as LIWC, count words in predefined categories such as anxiety, family, and pronoun. These categories are identified first, based on psychological constructs and syntactic groups, and the lexicons for each category were created manually. Such predefined categories may not capture the diversity of linguistic styles in online writing. Even if these lexicons can be expanded to include non-standard variants, the pre-defined nature of the categories limits its usage to capture varying linguistic styles that emerge in online writing.

As a first step towards building robust language processing and social systems, the work presented in this dissertation uncovers patterns of linguistic style variation in various social contexts. From a sociolinguistic perspective, linguistic variation is fundamental to language change, and understanding the stylistic variation in online writing is essential to track language change in progress.

1.5 Thesis Statement

In this dissertation, I investigate the interplay between evolving social media language, the need to convey varied social meanings in online interpersonal interactions, and the affordances in technology-mediated channels, through a series of large-scale computational studies of linguistic style variation in online writing. **My work shows that by using data-driven computational algorithms and causal statistical analysis of observational data we can extract insights about how social context influences online writing.** My work provides a holistic perspective on the connections between online linguistic style innovations, intra-person variation, inter-person variation, and community linguistic norms.

1.6 Contributions

This dissertation makes the following specific contributions in terms of our understanding of linguistic style variation in online writing:

- **An understanding of the effects of audience on linguistic style variation in social media.** By focusing on the relevance of audience size to intra-person variation in using non-standard forms in Twitter, I show that individuals use multiple stylistic innovations to convey varied social meaning. Specifically, I show that the users of Twitter are attuned to both the nature of their audience and the social meaning of lexical variation, and that they

customize their self-presentation accordingly.

- **An understanding of the effects of new technical affordances on online writing style.** Through a causal statistical analysis on the adoption of a new form of stylistic innovation in online writing, emojis, I show that the new forms of stylistic innovations compete with other orthographic resources in fulfilling social media users' need to convey varied social meanings.
- **Construction of a data-driven multidimensional lexicon to quantitatively measure linguistic style.** To systematically study the inter-person, intra-person, and inter-community linguistic style variation in online multi-community environments, I quantitatively operationalize the sociolinguistic concept of interpersonal stancetaking by developing a multidimensional stance lexicon. I validate the stance dimensions intrinsically by showing the linguistic coherence and interpretability of the extracted stance dimensions and extrinsically by showing the utility of the stance dimensions in predicting social phenomena.
- **An understanding of community linguistic norms and linguistic style-shifting in online multi-communities.** Using the multidimensional stance lexicon to characterize the linguistic style of sub-communities and individual members in Reddit and using statistical models, I provide preliminary evidence that individuals change their writing style depending on the community they participate in.
- **An understanding of the effects of moderation on online writing style.** By analyzing the causal effect of writing style norm enforcement in Wikipedia, I show that linguistic moderation in Wikipedia helps the convergence of writing style norms at the platform-level (i.e., articles), but not at individual member-level (i.e., editors).

CHAPTER 2

BACKGROUND AND RELATED WORK

Style in language is related to how things are said as opposed to what is said. In sociolinguistics, style is a set of linguistic variants with specific social meanings, including indications of personal attributes, group membership, or social situations. Variation is integral to the concept of stylistic practice—without variation there is no notion of varied social meaning [26]. Stylistic variation can occur phonologically, lexically, syntactically, and semantically. The concept of style in the context of sociolinguistics was first introduced in the work of William Labov in 1960’s, in which Labov studied New York City department store employees and found that variable pronunciations of the /r/ sound reflect socio-economic status of the speakers [43]. Other approaches studying the social meaning of linguistic style variation include indexicality [62, 63], language ideology [64, 65], and stancetaking [66]. Eckert and Rickford [26] provide a detailed survey about style and sociolinguistic variation. My thesis work is informed by this rich body of sociolinguistic literature.

Next, I survey prior related work under each of the four dimensions that I focus in this dissertation: (1) stylistic innovations, (2) inter-person variation, (3) intra-person variation, and (4) community norms. I end this chapter with a discussion of prior work using causal inference techniques for social media analysis. Part of the content of this chapter is based on published work related to this dissertation, presented in Chapters 4-6.

2.1 Linguistic Style Innovations: Non-standard Orthography

As the result of the need to convey social meaning in interpersonal interactions conducted through technology-mediated channels, online writing has become lexically diverse compared to other written genres [4, 5]. The majority of these lexical innovations are non-standard in other written genres, yet frequent in online writing. For example, a new set of orthographic variations such as abbreviations (e.g., *lol*), expressive lengthening (e.g., *cooolllll!!!*), and emoticons (e.g., *:-O*) are introduced to online writing [4] and these varieties do not follow the conventions of well-established standard orthographies [67].

One of the roles of such non-standard orthography is to express paralinguistic information, in the absence of nonverbal cues present in speech. Examples include the usage of repeated exclamation marks at the end of an utterance (e.g., *yes!!!!*) to emphasize the point of the author, asterisks surrounding words (e.g., the **real** question) to indicate that those words contained within them are to be heard with a different tone than the rest of the utterance, and a series of periods to indicate pause [68, 30]. Along this line, in a recent computational study Kalman and Gergle [69] investigate how non-standard character repetition is used to enrich communication using a large dataset of e-mail messages.

Another variety of non-standard orthography in online writing is the transcriptions of existing spoken language variables which have geographic (e.g., *hella* meaning ‘very’ or ‘a lot of’ originated from San Francisco [70]) or group relevance (e.g., *hooyah*, a spirited cry meaning anything positive, used by U.S. Navy to build morale [71, 72]). Such spoken language varieties index social and regional identities [73, 74]. Although prior work has extracted large-scale patterns of variation in non-standard language use in online writing in terms of macro-level properties of the authors, such as age [10], gender [13], and geographic location [9], far less is understood why these differences

have arisen in online writing. To fill this gap, in Chapter 4, I study the social meaning of non-standard language variation through the lens of the sociolinguistic concept of audience design. Emoticons and emojis are two special types of non-standard orthography in online writing. The next two subsections provide more details of these variants.


2.1.1 Emoticons

Emoticons are not transcriptions or alternations of existing words, but they are created by creatively combining multiple ASCII characters in some meaningful ways. There is a long history of research on the role of emoticons in textual communication. While some researchers view emoticons as signals of emotional expression [75, 76, 4], later research has shown that the usage of emoticons in CMC goes beyond the expression of emotions, and that the emotional interpretation of emoticons depends on properties of the communicative context, the author, and the reader [77, 78, 79]. Dresner and Herring [30] identify three broad linguistic functions of emoticons: (1) as emotion indicators, mapped directly onto facial expressions (e.g., happy or sad), (2) as non-emotional meaning, mapped conventionally onto facial expressions (e.g., joking), and (3) as an indication of the speaker’s communicative intention (e.g., ending an utterance with a smiley to mitigate the commitment of the author’s claims or opinions).

2.1.2 Emojis

Emojis are “picture characters”, including not only faces, but also pictographs representing objects, actions, concepts, and ideas. Emojis are recent additions to non-standard orthography and have similarities to emoticons. Similar to emoticons, emojis are not replications of existing words and they are visually similar to emoticons. In contrast to emoticons, which are created from ASCII character sequences, emojis are represented by unicode characters, and are continuously increasing in number with

the introduction of new characters in each new unicode version.¹ In mid-2015, new emoji characters were introduced to represent people with different skin tones and hair colors;² this diversity may help to show attributes of authors such as identity.

Emojis originated in Japanese mobile phones in the late 1990s. They have recently become popular worldwide in text messaging and social media, thanks in part to the adoption of smartphones supporting input and rendering of emoji characters. The increasing popularity of emojis has led to the selection of one of the pictographs  widely known as the “face with tears of joy” emoji, as the Oxford Dictionaries Word of the Year 2015.³ Twitter introduced emojis to its Web interface in early 2014, following emoji support on Twitter’s apps for Android and iOS.⁴ Emojis are also becoming widely popular in marketing, and recently Twitter started to support custom emojis for brands.⁵

With the increased popularity of emojis in CMC, researchers have started to explore the role of emojis in textual communication. Stark and Crawford [22] provide details of the origin of emojis and examine emojis as historical, social, and cultural objects. Kelly [80] interviewed a culturally diverse set of 20 participants about how they use emojis in mediated textual communication with close personal ties. As with emoticons, emojis are sometimes viewed as a vehicle for expressing sentiment and affect: for example, Novak et al. [81] developed a sentiment lexicon for emojis. However, as with emoticons, the communicative role of emojis is considerably broader and more nuanced than a direct expression of emotion. For example, the interviews performed by Kelly [80] revealed that emojis are used for a range of purposes beyond affect, such as maintaining a conversational connection, permitting a playful interaction, and creating a shared and secret uniqueness within a particular relationship. Emojis are also found to be

¹http://www.unicode.org/reports/tr51/index.html#Selection_Factors

²<http://www.unicode.org/reports/tr51/index.html#Diversity>

³<http://blog.oxforddictionaries.com/2015/11/word-of-the-year-2015-emoji/>

⁴<https://twitter.com/support/status/451399393850052610>.

⁵<http://www.nytimes.com/2016/03/07/business/media/picture-this-marketers-let-emojis-do-the-talking.html>

performing complex conversational functions such as situational meaning, adjusting tone, making a message more engaging to the recipient, conversation management, and relationship management [82]. Furthermore, the interpretation of emojis may be complicated by differences across viewing platforms [83].

Barbieri et al. [84] studied emoji usage and meaning over different languages using distributional semantics and found that the overall semantics of the subset of emojis they studied is preserved across American English, British English, Spanish, and Italian, but some emojis are interpreted differently across languages, which they attribute to socio-geographical differences. In an analysis of the global distribution of emojis, emoji usage is found to be correlated with living conditions of various parts of the world [85]. Recent work has also explored the compositionality of multi-emoji expressions [86], regularities of emojis usage across different topics of tweets [87], and emoji as a marketing rhetoric by influencers [88].

A question that has not been explored before and has relevance to social meaning of linguistic style in online writing is whether the new technological affordance of emojis—which seem to have similarities to emoticons and other non-standard lexical items in conveying paralinguistic cues in writing, but are more visually appealing and orthographically different in the sense that they are pre-defined symbols created by technologists⁶—are going to coexist or compete with earlier forms of non-standard orthographies in conveying social meaning in online writing. In Chapter 5, I systematically study this question using a causal inference approach, considering the adoption of emojis by individuals as the treatment and changes in emoji adopters’ writing styles as the effect of the treatment.

⁶Emojis are created and adopted through a standardization process that is run by the Unicode Consortium, a non-profit organization whose membership is made up mostly from large software and technology companies.

2.2 Inter-person Variation: Variation and Social Variables

In sociolinguist Penelope Eckert’s view, the chronological development of social meaning in variation in sociolinguistic studies has taken place in three waves [89]. The “first wave” of language variation studies in sociolinguistics focused on broad correlations between linguistic variables and macro-scale social categories such as age, gender, ethnicity, socio-economic status, and geographic origin [89]. Specific features and subtle linguistic variants are found to be communicating important information about individuals. For example, in Labov’s 1966 study, the differences in the use of /th/-stopping is found to be correlated with socio-economic class [90]. First wave studies also found gender stratification in variation: women’s speech is found to be consistently more standard than men’s speech across socioeconomic hierarchy [91, 92]. In the first wave, variables were taken to mark socioeconomic status, and stylistic and gender dynamics were considered to be the effects of these categories on speaker’s orientation to their assigned strata in these categories. While the “second wave” of studies used ethnographic methods to explore the local categories and configurations, in both the first and second waves, variation was seen as marking social categories [89].

In parallel with the first wave of variation studies in sociolinguistics, a large body of work in computational sociolinguistics [59] have focused on using computational techniques to study the relationship between language variation and macro-scale social variables. These studies primarily use large amounts of data gathered from online social platforms such as blogs, Twitter, and online forums. Formulated as a problem of predicting author attributes from text [14], this computational work demonstrate that it is possible to achieve high accuracy in predicting author age [10, 11, 12], gender [10, 14, 15, 16], race [17], and geography [8, 9]. Post hoc analysis reveals that the most informative predictors include proper names [93], spoken language dialect words [94, 95], transcriptions of phonological variation [5], as well as “netspeak”

phenomena, such as emoticons and abbreviations [96]

2.3 Intra-person Variation: Variation and Social Interactions

The previous section reviewed prior research about the connections between linguistic variation and macro-scale social categories, which has been the focus of the first wave of variation studies in sociolinguistics. The third wave of variation studies moved from a view of variation as a reflection of macro-scale social categories to variation as a reflection of the interactional context [89]. The social interactional context—such as the addressee or audience, topic of discussion, and social goals of the individuals—shape the linguistic style choices of individuals. This context-dependent, intra-person variation is known as linguistic style-shifting [97]. In this section I first provide a review of style-shifting in the sociolinguistic literature, and then I describe studies of style-shifting in computer mediated communication. I conclude this section with a review of research on linguistic style and interactional meaning.

2.3.1 Linguistic Style-Shifting

Linguistic style-shifting is the differential deployment of linguistic variables depending on the interactional context [97]. The sociolinguistic interplay between macro-scale social categories and interactional contexts has been the central concern of variation studies since the 1960s. William Labov’s 1966 study on the pronunciation of the /r/ sound in the speech of New York City population [43] offered a classical example of style-shifting. In this study, the differences in the pronunciation of the /r/ sound are shown to be modulated by both the socioeconomic status and the interactional situation. A key feature of this study is the interaction between fixed properties of the speaker (in this case, New York origin and socioeconomic status) and fluid properties of the interaction (in this case, attention to speech, which is experimentally modulated by the interviewer). Investigations on the linguistic consequences of the interaction

between fixed and fluid social properties could provide an explanation of the social meaning of linguistic variation, which cannot be explained only by the “first-order” associations between language and macro-level social categories [44, 55].

The degree of attention paid to speech is one possible explanation for style-shifting as proposed by Labov. Subsequent work in sociolinguistics has introduced other perspectives as the explanation for style-shifting [26], including *communication accommodation theory* and *audience design*. In communication accommodation theory, speakers adjust their linguistic style to converge or diverge, depending on their desire to reduce or increase social differences [98]. The theory of audience design is related to accommodation, but the target audience need not necessarily be the addressee [19].

2.3.2 Linguistic Style-shifting in Computer Mediated Communication

Studies on style-shifting. The literature on computer-mediated communication (CMC) has addressed the issue of style-shifting, mostly through qualitative or small-scale experiments. Through a manual analysis of a small number of dialogues in the #mannheim Internet Relay Chat (IRC) channel, Androutsopoulos and Ziegler [99] found evidence of regional language variation that matches the North-South gradation of German dialects. Paolillo [45] examined linguistic variation on Internet Relay Chat (IRC) channel (#india) to test the hypothesis that standard linguistic variants tend to be associated with weak social ties, while vernacular variants are associated with strong network ties [100]. He used factor analysis to identify social characteristics of 94 members of the #india community, and then examined the use of five linguistic variables in conversations across members of different factor groups. The results were mixed, with some variables being used more often among individuals with strong ties and other being used more often among individuals with weak ties. This may be in part because the variables included phonetic spellings like *r* (“are”) and *u* (“you”), which Paolillo notes were already widely accepted in computer-mediated communication in

the 1990s.

Studies on code-switching. A phenomena that is related to style-shifting, but different from style-shifting is code-switching. While style-shifting may occur across varieties of regional or social dialect or between different dialects, code-switching implies two distinct languages coexisting throughout a linguistic community providing speakers a much wider linguistic repertoire [59]. In his study of the #india IRC channel, Paolillo [45] reports codeswitching between Hindi and English, and recent studies have replicated this finding in Facebook and Twitter. For example, Androutsopoulos [101] investigate how a group of multilingual youth on Facebook strategically construct their audience using different language choices, finding that language choice is employed to maximize or partition the audience when starting new posts, or to align or disalign when responding to posts. Johnson [102] studied a convenience sample of 25 Welsh/English bilingual users of Twitter, finding that they tend to use Welsh when writing to individuals who are also bilinguals, but write tweets in English when the message is not directed to any specific user. Nguyen et al. [103] report similar results on a much larger scale, showing that Dutch Twitter users were more likely to employ the minority languages of Frisian and Limburgish in conversations with other users of these minority languages.

Audience-regulated style-shifting. To understand the relevance of social context to variation in online writing, the study presented in Chapter 4 explores the relevance of audience to stylistic variation in social media. Specifically, we focus on the context of addressee or audience of the interaction and use the theory of audience design to understand the differences in the usage of non-standard language on Twitter. Relevant to this work, in an ethnographic investigation, Marwick and boyd [34] conducted a series of interviews to understand how Twitter users navigate their “imagined audience”, and found that people consciously present themselves differently for different audiences

on Twitter, by modulating their language, cultural references, and style. However, Bernstein et al. [104] found that users of Facebook often have little idea of the size of the audience for their messages. Clearly, perceptions of audience vary across social media platforms, so the unique properties of each platform must be taken into account. When using the theory of audience design, the question of “which audience” is particularly salient in the context of publicly-readable social media: in principle any of the millions of users of Twitter could witness the conversation, though users may be aware of the specific preferences and interests of their own follower networks [34] and can indirectly manipulate the composition of the audience through affordances offered by the platform (see § 4.2).

In a large scale study of Twitter, Danescu-Niculescu-Mizil et al. [46] find evidence of linguistic accommodation in a large-scale corpus, but their linguistic variables mainly consist of lists of closed-class words such as articles and pronouns. They do not consider style-shifting in non-standard lexical varieties, as our work discussed in Chapter 4. Shoemark et al. [105] studied the usage of distinctively Scottish words in tweets related to the Scottish Independence Referendum, and found that while distinctively Scottish terms are used at a higher rate by users of pro-independence hashtags, in general people are less likely to use distinctively Scottish words in tweets with referendum-related hashtags than in their general Twitter activity. This suggests style-shifting relative to audience. Replicating Shoemark et al. [105]’s study of the usage of local variants in Scottish Referendum related tweets, Stewart et al. [106] studied the use of Catalan variants (as opposed to majority Spanish variants) in the Twitter discourse on 2017 Catalonia Referendum, and found that Catalan variants are used less in @-replies than in hashtag tweets. This contradicts the findings of Shoemark et al. [105] and the authors suggest a possible explanation that in the discourse related to the referendum, the @-replies could be targeted at well-known individuals such as politicians and that indirectly targets at a larger audience. In

another study Shoemark et al. [107] examined how Twitter users shift their use of Scottish variants depending on the topic and audience by looking at tweets from two groups of users with different overall rates of Scottish and found that both topic and audience influence the choice of Scottish variants.

Style-shifting in multi-communities. Prior work on style-shifting in CMC has primarily focused on a single social platform such a specific IRC channel and Twitter, and different social contexts such as audience and tie strength. What happens when users participate in multiple online communities in a single platform? In Chapter 7, I focus on linguistic style-shifting in the multi-community environment of Reddit. Specifically I explore whether members use the same linguistic style across multiple communities or whether they shift style to adapt to the linguistic style norms of each of the community in which they participate.

2.3.3 Linguistic Style and Interactional Meaning

Following recent sociolinguistic research on situational and stylistic variation, and the interactional meaning that such variation can convey [26], computational linguists have started to focus on computational efforts to quantify phenomena such as subjectivity [108], sentiment [109], politeness [50], formality [51], and power dynamics [54]. While these interpersonal phenomena are expressed through various linguistic means, often these phenomena do not occur in isolation. For example, a polite statement in some specific contexts is expected to be formal and expressing positive sentiment. Therefore it is essential to investigate how these range of interpersonal phenomena are manifested through linguistic means. The sociolinguistic concept of interpersonal stancetaking [110, 111, 112] provides a conceptual framework that accounts for a range of interpersonal phenomena, subsuming formality, politeness, and subjectivity. This framework has been applied almost exclusively through qualitative methods,

using close readings of individual texts or dialogs to uncover how language is used to position individuals with respect to their interlocutors and readers. Our work discussed in Chapter 6 provides the first quantitative operationalization of interpersonal stancetaking.

Prior computational efforts to quantify interactional meaning, such as the work on politeness by Danescu-Niculescu-Mizil et al. [50] and formality by Pavlick and Tetreault [51], make use of crowdsourced annotations to build large datasets of, for example, polite and impolite text. These annotation efforts draw on the annotators' intuitions about the meaning of these sociolinguistic constructs. In Chapter 6, I present a different approach, using unsupervised techniques, to quantitatively operationalize multiple dimensions of linguistic variation using the concept of stancetaking.

2.4 Community Norms: Variation and Multi-Communities

While the previous sections considered variation within or between individuals, individuals' style choices are also found to be influenced by the groups or communities they are part of [89]. In this section, I summarize research in multi-community variation, online community norms, norm enforcement in communities, and the effects of community norm enforcement.

2.4.1 Multi-community Variation

Social media platforms such as Reddit, Stack Exchange, and Wikia can be considered multi-community environments, in that they host multiple subcommunities with distinct social and linguistic properties. Such subcommunities can be contrasted in terms of topics [113, 114] and social networks [115]. Our work discussed in Chapter 6, considers community-wide differences in norms for interpersonal interaction. In the same vein, Tan and Lee [48] characterize stylistic differences across subreddits by focusing on very common words and parts-of-speech; Tran and Ostendorf [47] use

language models and topic models to measure similarity across threads within a subreddit. One distinction of our approach discussed in Chapter 6 is that the use of multidimensional analysis gives us interpretable dimensions of variation. This makes it possible to identify the specific interpersonal features that vary across communities. The study presented in Chapter 7 focuses on what happens when members participate in multiple online communities: whether they consistently use the same linguistic style across different communities or whether they adapt to the writing style of the specific community to which they post.

2.4.2 Online Community Norms

Norms are habitual behaviors that characterize a social group and differentiate it from other social groups [116]. While norms are explicitly defined through detailed guidelines (e.g., Reddit) and FAQs (e.g., Usenet) in some communities, in other communities norms are not formally codified, but emerge socially through the interactions of the members [38]. Online communities exhibit norms [39] related to member behavior such as content appropriateness [117], writing style [118], and adherence to community policies [119]. Community norms evolve as its members negotiate norms [37] and as newcomers arrive and old members become less active [40]. While community norms are studied in terms of deviant behavior such as abusive language [120], trolling [121], and vandalism [122], norms also encourage participation [123] and help communities achieve their goals [124, 125]. As a collaborative content production community, Wikipedia has several norms [126] related to neutral point of view [127], supportive communication [128], vandalism [129], and member participation [124]. In this dissertation, I focus on Wikipedia's norms related to NPOV writing style and present a study in Chapter 8 which investigates the effectiveness of NPOV norm enforcement.

2.4.3 Writing Style Norms

Communicative competence is the ability to use language appropriately, in accord with social situation and associated norms [130]. This notion of communicative competence can be used to explain linguistic norms of communities—the language used by community members should not only be correct, but should also be appropriate. A related concept is *legitimate language*, which is defined in any community as the language produced by a subset of speakers/writers with symbolic authority, and is often codified into explicit standards [131]. In the context of collaborative online writing in Wikipedia, legitimate language can be considered to be the language produced by the high-status members of the community, and is governed by community policies.

Among online communities, Wikipedia is well known for its focus on the quality of the content produced by volunteer contributors [124]. Wikipedia provides detailed guidelines of the preferred writing styles through a manual of style [118]. A large body of research focused on Wikipedia’s writing style norms including the prediction of article quality [132], biased content [133], quality flaws [134, 135], and vandalism detection [136]. Focusing specifically on bias in Wikipedia, Recasens et al. [137] built a dataset of Wikipedia phrases that are corrected for NPOV, and then trained a classifier using linguistic features from hand-crafted lexicons of factive verbs, hedges, and subjective intensifiers. We build on this prior work by using these lexicons to characterize biased language, but note that while Recasens et al. [137] aimed to identify the linguistic aspects of norms, the focus of the study presented in Chapter 8 is to understand the effects of NPOV tagging and correction. Specifically, we are interested in the change in the rate of biased language use when an article has an active NPOV tag, and when an editor is corrected for NPOV.

2.4.4 Enforcing Norms

Deviant behavior is an ongoing challenge to online communities, requiring persistent regulatory effort [138]. In addition to providing detailed guidelines, online communities also take active moderation actions against the violations of community guidelines [139], including both technical and social approaches [140]. Technical moderation includes banning posts with any keywords from predefined word lists (e.g., Yik Yak [141]) and banning posts based on the IP addresses from which posts originate (e.g., Hacker News [142], Yelp [143]). Social moderation includes both distributed approaches [144] where the community members vote on content (e.g., Yik Yak [141], Slashdot [119]) and centralized approaches where a small number of users called moderators maintain community practices by removing posts they find inappropriate (e.g., Reddit [140]).

In Wikipedia, a rich set of policies and guidelines articulate strategies for seeking consensus, principles of encyclopedic content, and appropriate user behavior. The policy environment in Wikipedia encodes and explains norms, but the policies are not imposed top down; rather policies are created and managed by the Wikipedia community itself [145]. Norms are enforced by policy citations [145], where template messages [146] are added to articles and discussion talk pages [147]. The policy pages are subject to the same editing processes of Wikipedia articles and contributors are given a participatory role when creating or editing policy pages. While several quality control mechanisms to enforce objective language in Wikipedia articles have been in place for a long time, the impact of these mechanisms has not been quantitatively measured. This gap in the literature motivates our broad research question of the study presented in Chapter 8: whether NPOV tagging and corrections help Wikipedia articles and editors to converge to desired writing styles norms.

2.4.5 Effects of Norm Enforcement

Most prior work on the effects of norm enforcement has focused on moderation techniques related to the design of online communities. Users report significantly higher intent to participate in a community that is moderated, compared to an unmoderated community [148]. A study on Slashdot — a popular website with fair moderation [119] — found that users come to a consensus about the community’s moderation policies and such moderation practices enable large scale civil participation [144]. Reddit’s decision to ban hate communities resulted in a reduction of abusive language [149]. An analysis of different moderation styles in online health support communities found that positive and rewarding moderation styles are more effective than negative and punishing styles [150]. In online discussion forums related to education, peer moderation is found to be more encouraging active participation than moderation by superiors [151].

The effects of moderation on the quality of collaboratively created content and long term behavior of individual community members have not been investigated much. In the context of Wikipedia, Halfaker et al. [152] found that the action of “reverting” edits has the effect of reducing motivation and quantity of work, particularly for new editors. However, they also found that reverts result in higher quality contributions. Halfaker et al. [153] found that enforcing quality control mechanisms affects the retention of high-quality newcomers. Personalized warning messages, as opposed to pre-defined template messages, are found to be effective in retaining newcomers [154]. Motivated by this prior literature, which studies the effectiveness of norm enforcement actions both at the platform-level (i.e., articles) and individual member-level (i.e., editors), the study presented in this dissertation in Chapter 8 seeks to answer the research question about the effectiveness of NPOV norm enforcement both at the article level and at the editor level.

2.5 Causal Inference for Social Media Analysis

The use of the statistical framework of causal inference is very common in fields such as epidemiology and political science, but until recently it has been rare in social media research [155, 156]. While social media is a rich resource to study naturally occurring social phenomena, there exists some challenges when using observational data. In general, observational studies of causal phenomena are susceptible to confounds because subjects are not randomly assigned to treatment and control groups as in randomized experiments [157]. For example, the Twitter client from which users post tweets (e.g., Web browser, smart phone app) is a potential confound on writing style, because different Twitter clients have varied text-input capabilities which could affect Twitter users' writing style [158] (§ 5.3.1). Similarly, the experience level of Wikipedia editors is a potential confound on how they adhere to community norms [152] (§ 8.5.1).

Studies using large scale observational data have used techniques from the causal inference literature, such as matching [159] and stratification [160], to reduce the effects of potential confounds. For example, Dos Reis and Culotta [161] used matching to create treatment and control groups of users to understand the effects of exercise on mental health; De Choudhury et al. [162] used matched samples to examine dietary choices and nutritional challenges in food deserts using Instagram posts; Cheng et al. [121] used matching on users to study antisocial behavior in online forums; Chandrasekharan et al. [149] used Mahalanobis Distance Matching [163] and difference-in-differences strategy [164] to study the effects of subreddit banning on hate speech usage on Reddit. Another causal inference technique that has been used in longitudinal studies of social media data is interrupted time series analysis [165]. This technique has been used to study hate speech [149] and conspiratorial discussions [166] on Reddit.

As mentioned in § 1.3, one of the methodological aspects in which my dissertation

work differs from prior studies on stylistic variation is the focus on causal analysis as opposed to correlation effects. In the studies presented in Chapter 5 and Chapter 8, we apply causal inference approaches to the analysis of linguistic style.

CHAPTER 3

CONFOUNDS AND CONSEQUENCES OF USING SOCIAL MEDIA DATA TO STUDY ONLINE WRITING¹

What are the biases inherent to social media data and how can these biases affect the research in stylistic variation using data from social media and other online platforms? In this chapter, I will examine the geographic biases in Twitter and the impact of these biases on the study of online writing. First, I will address the geographic biases at population level, by comparing the observed population distributions in Twitter data with expected distributions that are based on ground-truth demographic data. Specifically, I will explain how we create three datasets of tweets using different sampling methods and measure the geographic and demographic biases in each of these samples. Then, I will discuss how these biases impact linguistic investigations, such as our studies on online writing. Although datasets from most online social platforms are prone to these biases, in this work we focus on geographic linguistic biases in Twitter data, which has more relevance to this dissertation.

3.1 Background

Social media content has often been used to study several social phenomena, ranging from finding political opinions [167] to public health surveillance [168]. Related to this dissertation, data from social media platforms such as Twitter is frequently used to study geographical linguistic phenomena such as linguistic style and regional dialectal differences [9, 94, 169, 170], and text-based location inference [8, 171, 172]. Compared to traditional interview and survey techniques used in sociolinguistics [173] and other

¹Content of this chapter is based on Umashanthi Pavalanathan and Jacob Eisenstein. “Confounds and Consequences in Geotagged Twitter Data.”, EMNLP 2015.

social sciences [174], social media provides larger volumes of datasets that can be used for both fine-grained as well as global-scale analyses. However, social media data is not a representative sample of the true population and this leads to geographic and demographic biases in population level analyses [175, 176, 177, 178].

This chapter examines the effects of the geographic and demographic biases on the geo-linguistic inferences that can be drawn from Twitter. Using tweets from the top ten largest metropolitan areas in the United States and three different data sampling techniques, we measure the impact of geographical and demographic biases in both language use and in the accuracy of text-based location inference.

3.2 Data

We used data gathered from Twitter’s streaming API from February 2014 to January 2015. We removed retweets, which are repetitions of previously posted messages and performed the following filtering steps to eliminate spam and automated accounts [179]: removed tweets with URLs, user accounts with more than 1,000 followers or followees, user accounts with more than 5,000 messages posted during the time period of the dataset, and top 10% of accounts based on the number of messages in our dataset. As we focus on English language tweets, we removed users who have written more than 10% of their tweets in any language other than English.

To quantify the geographic biases, we consider the top ten largest Metropolitan Statistical Areas (MSAs) in the United States (Table 3.1). The U.S. Census Bureau defines the MSAs as geographical regions of high population with density organized around a single urban core; MSAs are not legal administrative divisions. The urban core of an MSA may be surrounded by substantially less urban areas. For example, the Atlanta MSA is centered on Fulton County (1750 people per square mile), but extends to Haralson County (100 people per square mile), on the border of Alabama. A per-county analysis of this data therefore enables us to assess the degree to which

Twitter’s skew towards urban areas biases geo-linguistic analysis.

3.3 Representativeness of Geotagged Twitter Data

Extracting Locations of Tweets We built a dataset of GPS-tagged (\mathcal{D}_G) tweets by extracting the GPS latitude and longitude coordinates and then we did reverse geocoding to identify the corresponding counties. Only 1.24% of the tweets had GPS tags, and the users who geotag their tweets could have unique characteristics. Therefore we built another dataset (\mathcal{D}_L) by combining the tweets which had a location field in the user’s profile and mapping the locations to MSAs. An analysis of the intersection of \mathcal{D}_G and \mathcal{D}_L found good agreement between both data sources which implies that the location information in user profiles are accurate majority of the times.

Data Subsampling We resampled \mathcal{D}_G and \mathcal{D}_L to create three balanced datasets: **GPS-MSA-BALANCED:** We randomly sampled 25,000 tweets per MSA from \mathcal{D}_G as the message-balanced sample, and all the tweets from 2,500 users per MSA from the same set as the user-balanced sample. Balancing across MSAs ensures that the largest MSAs do not dominate the linguistic analysis.

GPS-COUNTY-BALANCED: We resampled \mathcal{D}_G based on county-level population, and again obtained message-balanced and user-balanced samples. These samples are more geographically representative of the overall population distribution across each MSA.

LOC-MSA-BALANCED: We randomly sampled 25,000 tweets per MSA from the \mathcal{D}_L as the message-balanced sample, and all the tweets from 2,500 users per MSA from the same set as the user-balanced sample. Exact geographical coordinates are not available in the \mathcal{D}_L and therefore it is not possible to obtain county-level geolocations for this set.

Age and Gender Identification To estimate the distribution of age and gender in each sample, we queried statistics from the Social Security Administration, which records the number of individuals born each year with each given name. Using this information, we obtained the probability distribution of age values for each given name. We then matched the names against the first token in the name field of each user’s profile, enabling us to induce approximate distributions over ages and genders. Unlike Facebook and Google+, Twitter does not have a “real name” policy, so users are free to give names that are fake, humorous, etc. We eliminate user accounts whose names are not sufficiently common in the social security database. While some individuals will choose names not typically associated with their gender, we assume that this will happen with roughly equal probability in both directions. So, with these caveats in mind, we induce the age distribution for the GPS-MSA-BALANCED sample and the LOC-MSA-BALANCED sample as,

$$p(a \mid \text{name} = n) = \frac{\text{count}(\text{name} = n, \text{age} = a)}{\sum_{a'} \text{count}(\text{name} = n, \text{age} = a')} \quad (3.1)$$

$$p_{\mathcal{D}}(a) \propto \sum_{i \in \mathcal{D}} p(a \mid \text{name} = n_i). \quad (3.2)$$

We induce distributions over author gender in much the same way [175]. This method does not incorporate prior information about the ages of Twitter users, and thus assigns too much probability to the extremely young and old, who are unlikely to use the service. While it would be easy to design such a prior — for example, assigning zero prior probability to users under the age of five or above the age of 95 — we see no principled basis for determining these cutoffs. We therefore focus on the *differences* between the estimated $p_{\mathcal{D}}(a)$ for each sample \mathcal{D} .

Results: Geographical Biases in the GPS Sample We first assess the differences between the true population distributions over counties, and the per-tweet and per-user distributions. Because counties vary widely in their degree of urbanization and other demographic characteristics, this measure is a proxy for the representativeness of GPS-based Twitter samples (county information is not available for the LOC-MSA-BALANCED sample). Population distributions for New York and Atlanta are shown in Figure 3.1. In Atlanta, Fulton County is the most populous and most urban, and is overrepresented in both geotagged tweets and user accounts; most of the remaining counties are correspondingly underrepresented. This coheres with the urban bias noted earlier by Hecht and Stephens [176]. In New York, Kings County (Brooklyn) is the most populous, but is underrepresented in both the number of geotagged tweets and user accounts, at the expense of New York County (Manhattan). Manhattan is the commercial and entertainment center of the New York MSA, so residents of outlying counties may be tweeting from their jobs or social activities.

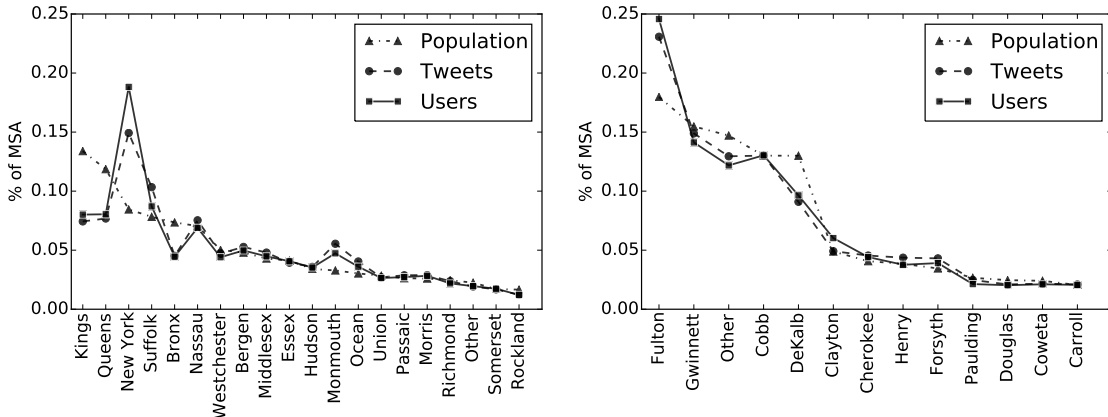


Figure 3.1: Proportion of census population, Twitter messages, and Twitter user accounts, by county. New York is shown on the left, Atlanta on the right.

To quantify the representativeness of each data sample, we use the L1 distance $\|\mathbf{x} - \mathbf{y}\|_1 = \sum_c |p_c - t_c|$, where p_c is the proportion of the MSA population residing in county c and t_c is the proportion of tweets (Table 3.1). County boundaries are determined by states, and their density varies: for example, the Los Angeles MSA

Table 3.1: L1 distance between county-level population and Twitter users and messages

MSA	Num. Counties	L1 Dist. Population vs. Users	L1 Dist. Population vs. Tweets
New York	23	0.2891	0.2825
Los Angeles	2	0.0203	0.0223
Chicago	14	0.0482	0.0535
Dallas	12	0.1437	0.1176
Houston	10	0.0394	0.0472
Philadelphia	11	0.1426	0.1202
Washington DC	22	0.2089	0.2750
Miami	3	0.0428	0.0362
Atlanta	28	0.1448	0.1730
Boston	7	0.1878	0.2303

covers only two counties, while the smaller Atlanta MSA is spread over 28 counties. Table 3.1 shows that while New York is the most extreme example, most MSAs feature an asymmetry between county population and Twitter adoption.

Results: Usage Biases Next, we turn to differences between the GPS-based and profile-based techniques for obtaining ground truth data. As shown in Figure 3.2, the LOC-MSA-BALANCED sample contains more low-volume users than either the GPS-MSA-BALANCED or GPS-COUNTY-BALANCED samples. We can therefore conclude that the county-level geographical bias in the GPS-based data does not impact usage rate, but that the difference between GPS-based and profile-based sampling does; the linguistic consequences of this difference will be explored in the following sections.

Results: Demographics Biases Table 3.2 shows the expected age and gender for each dataset, with bootstrap confidence intervals. Users in the LOC-MSA-BALANCED dataset are on average two years older than in the GPS-MSA-BALANCED and GPS-COUNTY-BALANCED datasets, which are statistically indistinguishable. Focusing on the difference between GPS-MSA-BALANCED and LOC-MSA-BALANCED, we plot the difference in age probabilities in Figure 3.3, showing that GPS-MSA-BALANCED in-

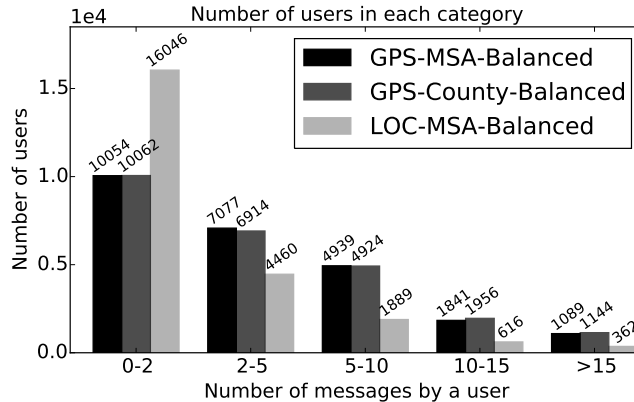


Figure 3.2: User counts by number of Twitter messages

cludes many more teens and people in their early twenties, while LOC-MSA-BALANCED includes more people at middle age and older. Young people are especially likely to use social media on cellphones [180], where location tagging would be more relevant than when Twitter is accessed via a personal computer. Social media users in the age brackets 18-29 and 30-49 are also more likely to tag their locations in social media posts than social media users in the age brackets 50-64 and 65+ [181], with women and men tagging at roughly equal rates. Table 3.2 shows that the GPS-MSA-BALANCED and GPS-COUNTY-BALANCED samples contain significantly more women than LOC-MSA-BALANCED, though all three samples are close to 50%.

Table 3.2: Demographic statistics for each dataset

Sample	Expected Age	95% CI	% Female	95% CI
GPS-MSA-BALANCED	36.17	[36.07 – 36.27]	51.5	[51.3 – 51.8]
GPS-COUNTY-BALANCED	36.25	[36.16 – 36.30]	51.3	[51.1 – 51.6]
LOC-MSA-BALANCED	38.35	[38.25 – 38.44]	49.3	[49.1 – 49.6]

3.4 Impact on Linguistic Generalizations

Twitter data has been used to study geographical linguistic variation. Previous sections of this chapter identified several demographic differences when using different data

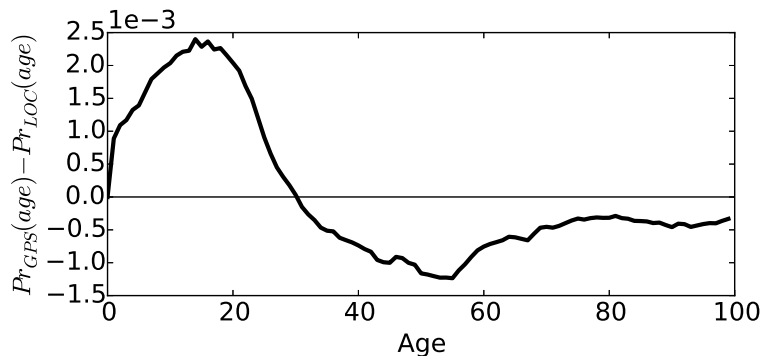


Figure 3.3: Difference in age probability distributions between GPS-MSA-BALANCED and LOC-MSA-BALANCED.

samples. How do these differences impact the linguistic generalizations that we draw from Twitter data? To answer this question, we measure the linguistic differences between GPS-MSA-BALANCED and LOC-MSA-BALANCED.² Because GPS-MSA-BALANCED and LOC-MSA-BALANCED differ in the composition of age and gender, we also directly measure the impact of these demographic factors on geographical linguistic variables.

Discovering geographical linguistic variables Lexical variation in text corpora can be identified using different statistics as surveyed by Monroe et al. [49]. A regularized log-odds ratio strikes a good balance between distinctiveness and robustness [170]. We use such an approach implemented in SAGE [17]³ and identify the 25 most salient lexical items for each metro area in both GPS-MSA-BALANCED and LOC-MSA-BALANCED. Two main types of geographical lexical variables identified in previous work are: (1) non-standard words and spellings (e.g., *hella*, *yinz*) [170], which are primarily stylistic; (2) entity names (e.g., places, local merchants, etc.) [182], which are primarily topic-based. These two groups need to be differentiated since this would decide whether geo-linguistic differences are primarily topic-based or stylistic.

²Since the GPS-MSA-BALANCED and GPS-COUNTY-BALANCED methods have nearly identical patterns of usage and demographics, we focus on the linguistic difference between GPS-MSA-BALANCED and LOC-MSA-BALANCED.

³<https://github.com/jacobeisenstein/jos-gender-2014>

Therefore we annotate each group of 25 lexical variables as NONSTANDARD-WORD, ENTITY-NAME, or OTHER. Annotation for ambiguous cases is based on the majority sense in randomly-selected examples. Overall, we identify 24 NONSTANDARD-WORDS and 185 ENTITY-NAMES.

Inferring author demographics We sharpen the demographic inference method discussed in § 3.3 by considering the text of the tweets and build a simple latent variable model in which both the name and the word counts are drawn from distributions associated with the latent age and gender [183]. We use expectation-maximization to perform inference in this model, binning the latent age variable into four groups: 0-17, 18-29, 30-39, above 40. A detailed description of this inference can be found in the published version of this work, Pavalanathan and Eisenstein [42].

Results: Linguistic differences by dataset The keywords identified in GPS-MSA-BALANCED dataset feature more geographically-specific non-standard words, which occur at a rate of 3.9×10^{-4} in GPS-MSA-BALANCED, versus 2.6×10^{-4} in LOC-MSA-BALANCED; this difference is statistically significant ($p < .05, t = 3.2$).⁴ For entity names, the difference between datasets was not significant, with a rate of 4.0×10^{-3} for GPS-MSA-BALANCED, and 3.7×10^{-3} for LOC-MSA-BALANCED. Note that these rates include only the non-standard words and entity names detected by SAGE as among the top 25 most distinctive for one of the ten largest cities in the US; of course there are many other relevant terms that are below this threshold.

Results: Demographics Aggregate linguistic statistics for demographic groups are shown in Figure 3.4. Men use significantly more geographically-specific entity names than women ($p \ll .01, t = 8.0$), but gender differences for geographically-specific

⁴We employ a paired t-test, comparing the difference in frequency for each word across the two datasets. Since we cannot test the complete set of entity names or non-standard words, this quantifies whether the observed difference is robust across the subset of the vocabulary that we have selected.

non-standard words are not significant ($p \approx .2$).⁵ Younger people use significantly more geographically-specific non-standard words than older people (ages 0–29 versus 30+, $p \ll .01, t = 7.8$), and older people mention significantly more geographically-specific entity names ($p \ll .01, t = 5.1$). Of particular interest is the intersection of age and gender: the use of geographically-specific non-standard words decreases with age much more profoundly for men than for women; conversely, the frequency of mentioning geographically-specific entity names increases dramatically with age for men, but to a much lesser extent for women. The observation that high-level patterns of geographically-oriented language are more age-dependent for men than for women suggests an intriguing site for future research on the intersectional construction of linguistic identity.

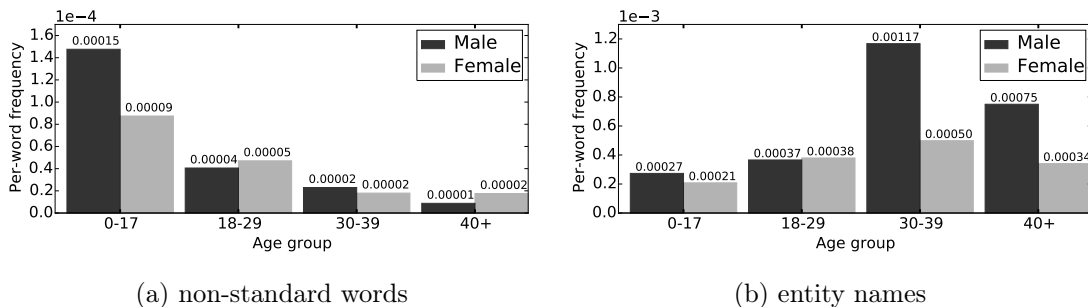


Figure 3.4: Aggregate statistics for geographically-specific non-standard words and entity names across imputed demographic groups, from the GPS-MSA-BALANCED sample.

For a more detailed view, we apply SAGE to identify the most salient lexical items for each MSA, subgrouped by age and gender. Table 3.3 shows word lists for New York (the largest MSA) and Dallas (the 5th-largest MSA), using the GPS-MSA-BALANCED sample. Non-standard words tend to be used by the youngest authors: *ilysm* (‘I love you so much’), *ight* (‘alright’), *oomf* (‘one of my followers’). Older authors write more about local entities (*manhattan*, *nyc*, *houston*), with men focusing on sports-related entities (*harden*, *watt*, *astros*, *mets*, *texans*), and women above the age

⁵But see Bamman et al. [16] for a much more detailed discussion of gender and standardness.

of 40 emphasizing religiously-oriented terms (*proverb, islam, rejoice, psalm*).

Table 3.3: Most characteristic words for demographic subsets of each city, as compared with the overall average word distribution

Age	Sex	New York	Dallas
0-17	F	<i>niall, ilysm, hemmings, stalk, ily</i>	<i>fanuary, idol, lmbo, lowkey, jonas</i>
	M	<i>ight, technique, kisses, lesbian, dicks</i>	<i>homies, daniels, oomf, teenager, brah</i>
18-29	F	<i>roses, castle, hmmm, chem, sinking</i>	<i>socially, coma, hubby, bra, swimming</i>
	M	<i>drunken, manhattan, spoiler, guardians, gonna</i>	<i>harden, watt, astros, rockets, mavs</i>
30-39	F	<i>suite, nyc, colleagues, york, portugal</i>	<i>astros, sophia, recommendations, houston, prepping</i>
	M	<i>mets, effectively, cruz, founder, knicks</i>	<i>texans, rockets, embarrassment, tcu, mississippi</i>
40+	F	<i>cultural, affected, encouraged, proverb, unhappy</i>	<i>determine, islam, rejoice, psalm, responsibility</i>
	M	<i>reuters, investors, shares, lawsuit, the-aters</i>	<i>mph, wazers, houston, tx, harris</i>

3.5 Impact on Text-based Geolocation

Prediction of user geolocation is one of the major applications of geotagged social media data [9, 8, 93, 171, 172]. Beyond commercial applications, geotagged social media data is also useful to study geographical phenomena in social media. Previous research has obtained impressive accuracies for text-based geolocation: for example, Hong et al. [171] report a median error of 120 km, which is roughly the distance from Los Angeles to San Diego, in a prediction space over the entire continental United States. The training and test datasets used for computing these accuracies are acquired through the same procedures and therefore the biases in the acquisition procedures could bias the resulting accuracy estimations. This could lead to overly optimistic or pessimistic estimations for some groups of users. In this section, we explore where these text-based geolocation methods are most and least accurate.

Methods We used user-balanced samples from the ten largest metropolitan areas in the United States, and we formulate text-based geolocation as a ten-way classification problem, similar to Han et al. [172].⁶ We apply ten-fold cross validation, and tune the regularization parameter on a development fold, using the vocabulary of the sample as features.

Results: Accuracy vs. Method of Data Acquisition Accuracies of author-attribute prediction tasks depend on the amount of data available [184]. Since GPS-MSA-BALANCED and LOC-MSA-BALANCED have very different usage rates (Figure 3.2), perceived differences in accuracy may be purely attributable to the amount of data available per user, rather than to users in one group being inherently harder to classify than another. For this reason, we bin users by the number of messages in our sample of their timeline, and report results separately for each bin. All errorbars represent 95% confidence intervals. As seen in Figure 3.5a, there is little difference in accuracy across sampling techniques: the location-based sample is slightly easier to geolocate at each usage bin, but the difference is not statistically significant. However, due to the higher average usage rate in GPS-MSA-BALANCED (see Figure 3.2), the overall accuracy for a sample of users will appear to be higher on this data.

Results: Demographics We also measure the classification accuracy by gender and age, using the GPS-MSA-BALANCED sample. As shown in 3.5b, text-based geolocation is consistently more accurate for male authors, across almost the entire spectrum of usage rates. As shown in 3.5c, older users also tend to be easier to

⁶Many previous papers have attempted to identify the precise latitude and longitude coordinates of individual authors, but obtaining high accuracy on this task involves much more complex methods, such as latent variable models [9, 171], or multilevel grid structures [8, 182]. Tuning such models can be challenging, and the resulting accuracies might be affected by initial conditions or hyperparameters. We therefore focus on classification, employing the familiar and well-understood method of logistic regression.

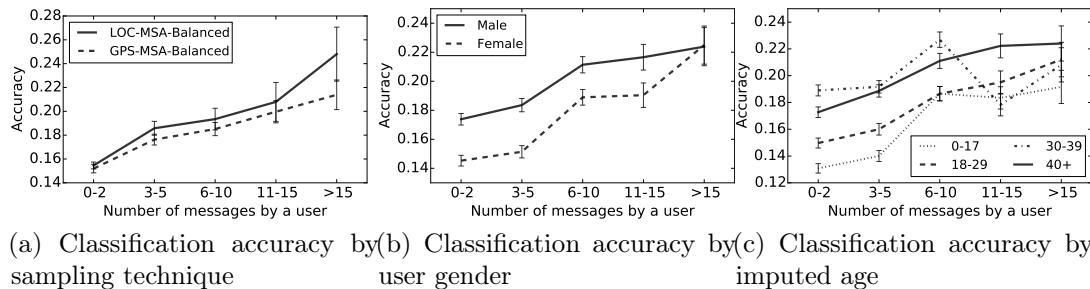


Figure 3.5: Classification accuracies

geolocate: at each usage level, the highest accuracy goes to one of the two older groups, and the difference is significant in almost every case. As discussed in § 3.4, older male users tend to mention many entities, particularly sports-related terms; these terms are apparently more predictive than the non-standard spellings and slang favored by younger authors.

3.6 Conclusions

In this chapter, I presented our work which examines how the biases inherent in social media data affect the research in online writing. To assess the biases, we first created three datasets of tweets using different sampling methods and then measured the population level biases in each of these datasets. Then we examined how these biases impact the linguistic inferences we make from datasets from social media such as Twitter. Through these investigations, we have identified demographic and linguistic biases that emerge from Twitter corpora created using different data acquisition techniques. Primary findings include:

- In comparison with tweets with self-reported locations, GPS-tagged tweets are written more often by young people and by women.
- There are corresponding linguistic differences between datasets created using different data acquisition techniques, with GPS-tagged tweets including more geographically-specific non-standard words.

- Young people use significantly more geographically-specific non-standard words. Men tend to mention more geographically-specific entities than women, but these differences are significant only for individuals at the age of 30 or older.
- Users who GPS-tag their tweets tend to write more, making them easier to geolocate. Evaluating text-based geolocation on GPS-tagged tweets probably overestimates its accuracy.
- Text-based geolocation is significantly more accurate for men and for older people.

These findings suggest that social media data collected using different methods may not be representative of the actual user population and these biases should be considered when making generalizations of research findings using the skewed datasets. Overall, findings from this work inform the data sampling methods used in the sociolinguistic investigations described in the following chapters.

CHAPTER 4
LINGUISTIC STYLE INNOVATIONS AND INTRA-PERSON
VARIATION IN SOCIAL MEDIA¹

We observe an enormous amount of stylistic diversity in social media data. New non-standard lexical forms such as emoticons (e.g., *:P*), emojis (e.g., 😂), abbreviations (e.g., *ikr*), and expressive lengthening (e.g., *coooooo!!!!!!*) are introduced to online writing. A rich body of previous work has identified differences in using such non-standard language based on social attributes such as age and gender. But we also observe variation at the individual level. Individuals use different amounts of these lexical forms in different messages. What is the social meaning or social goal for using varying amounts of non-standard language? In this chapter, I will explore this question using the sociolinguistic construct called *audience design*, which suggests relationships between audience size and stylistic variation. To study this, I will first explain how we can characterize different audience sizes using the socio-technical features or affordances present in the Twitter platform. Then I will detail how we use a data-driven approach to automatically extract a large number of linguistic variables for our study. Following this, I will describe how we construct our study data in order to eliminate potential confounds, such as the ones identified in the previous chapter. After describing the dataset construction, I will explain how we perform the analysis using logistic regression models. I will conclude this chapter with the findings from the analysis. Overall, this chapter focuses on the following two dimensions I described in Chapter 1: *stylistic innovations* and *intra-person variation*.

¹Content of this chapter is based on Umashanthi Pavalanathan and Jacob Eisenstein. “Audience-modulated Variation in Online Social Media.”, American Speech, May 2015.

4.1 Background

As discussed in Chapter 1 and Chapter 2, stylistic variation in online writing has been attributed to macro-scale social variables such as author age [10], geography [9], race [17], and gender [13]. These macro-scale variables capture the inter-person variation. However, individuals' writing style is not uniform. Moreover, individuals' stylistic choices depend on the social context of the written utterances. This intra-person variation is known as linguistic style-shifting in the sociolinguistic literature [43, 26].

The sociolinguistic literature provides a rich theoretical foundation about the social meaning of style-shifting. Labov [43] presented one of the early studies in style-shifting and focused on attention to speech as an explanation for style-shifting. Other explanations of style-shifting include the desire to reduce or increase social differences [Communication Accommodation Theory; 98], addressing a target audience [Audience Design; 19], situational variation [185, 110], and identity construction [97]. The CMC literature provides several studies of style-shifting, but these studies are based on either small-scale empirical investigations [99] or qualitative interviews [34]. The CMC literature also provides studies on code-switching [45, 103] but these studies focus on code-switching between different languages, and not dialects in a single language as is the focus in this work.

Motivated by the rich theoretical literature on the relevance of audience to stylistic variation, such as the models of accommodation [98], audience design [19], and stancetaking [110], in this work, we focus on the notion of audience modulated stylistic variation in online writing. Specifically, the goal of this study is to understand the linguistic style innovations on Twitter and the social meaning of intra-person variation in using these stylistic innovations. Ethnographic research of Marwick and boyd [34] suggests that users of Twitter have definite ideas about who their audience is, and that they customize their self-presentation accordingly. Furthermore, contemporary social

media platforms such as Twitter and Facebook offer increasingly nuanced capabilities for its users to manipulate the composition of their audience, enabling them to reach both within and beyond the social networks defined by explicitly-stated friendship ties (§ 4.1).

We use a carefully constructed dataset of tweets from thousands of users and more than 200 lexical variables. These variables are organized into two sets: the first consists of terms that distinguish major American metropolitan areas from each other, and is obtained using an automatic technique based on regularized log-odds ratio. The second set of variables consists of the most frequently-used non-standard terms among Twitter users in the United States. We test the following two hypotheses related to audience size and non-standard linguistic innovations:

- **Hypothesis I:** Non-standard linguistic innovations are used more often in messages to limited audience.
- **Hypothesis II:** Non-standard linguistic innovations are used more often in messages targeted at local individuals.

4.2 The Social Environment of Twitter

Twitter provides a set of affordances that allow its users to control the likely composition of the audience of their messages, as shown in Figure 4.1. Also, previous work has found that social media users have specific understanding of their audience as they make self-presentation decisions about language and cultural references [34]. Therefore we rely on the following audience navigation affordances of Twitter to build our dataset and analysis for this study:

- **@-mention:** Tweets can be directed to individual users by including their username with an “@” symbol (e.g., @DeepakChopra in Figure 4.1). @-mention can be either at the beginning of the tweet or in the body of the tweet. In both

cases, the mentioned users are notified by default. If the tweet begins with an @-mention, then it will be visible to only the users who follow both the user who tweeted and the user who was mentioned.

- **Hashtags:** Keywords or phrases prefixed by a ‘#’ symbol serve as a tagging convention on Twitter and it enables users to search for tweets of interest outside their following network. Hashtags can be considered as a mechanism to create a virtual community of users who are interested in a specific topic or event and popular hashtags have a higher likelihood of reaching an audience outside a user’s follower network [186].
- **Broadcast:** Following Kwak et al. [187], we consider messages that do not contain an @-mention or a hashtag as broadcasts, since these tweets can potentially be viewed by all followers of user who posted the tweet.

The comparison between these affordances, as shown in Figure 4.1, enables us to quantify the different types of intended audience and measure the effect of linguistic style. Specifically, we design predictors that capture these different types of audience-navigation affordances and test their effect on the use of two types of linguistic style innovations on Twitter. As one category of our linguistic style innovations has geographical relevance, within @-mentions we differentiate pairs of users who are geolocated in the same metropolitan area.

4.3 Linguistic Style Innovations

To study the audience-driven intra-person variation, we use two sets of non-standard lexical innovations, which are easier to automatically quantify at scale than phonetic or morphosyntactic variables: (1) non-standard variants that have association with specific geographic locations in the United States, including regional dialects, and (2) non-standard style innovations that emerged on Twitter.

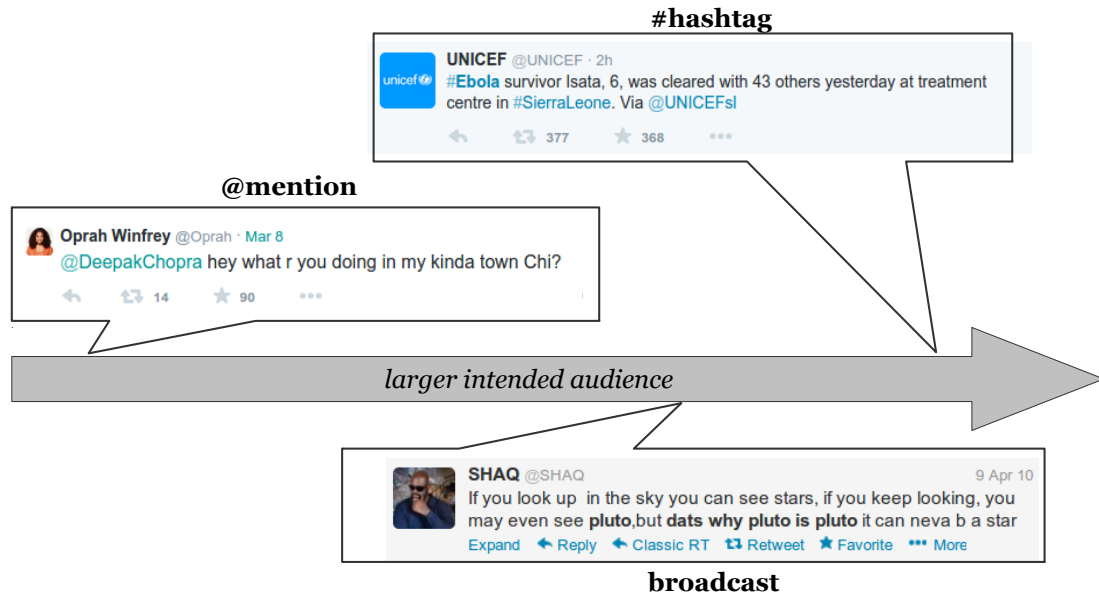


Figure 4.1: Affordances offered by Twitter for influencing the scope of the audience.

4.3.1 Geography-specific lexical innovations

As mentioned in Chapter 2, geography-specific lexical variation in Twitter is documented in several papers. Using approaches in prior work, we first geolocate each tweet in our corpus to a Metropolitan Statistical Area (MSA), which is a geographical category defined by the U.S. Census Bureau as a high-density region organized around a single urban core. Using these geolocation annotation, we extracted a list of terms that are specific each of the top ten most populous MSAs in the United States. Similar to our approach in Chapter 3, we applied a regularized log-odds based non-parametric approach from Eisenstein et al. [17], which is shown to be well-suited for similar tasks [170]. Using this approach, we obtained the top 30 words for each of the ten MSA and then manually removed entity names, standard English words, and non-English words. This resulted in a list of 120 lexical variables, shown in Table 4.1. Example usage of these lexical variables in tweets (Table A.1) and MSA frequencies (Figure A.1) are shown in Appendix A.

Table 4.1: List of geographical linguistic variable from each ten most populous metropolitan statistical areas (MSAs) in the United States.

MSA	Linguistic Variables
New York	lml, deadass, od, odee, werd, cud, nuttin, nicee, sed, lata, buggin, wrd, noe, w , layin, okk, lols, lolrt, crazyy, sour, wid
Los Angeles	fasho, ahah, cuh, koo, cuhz, fkn, ahahah, ;o
Chicago	mfs, goofy, nbs, lbvs, bogus, 2ma, lbs, mf, ikr, lmmfao, hoop, crackin
Dallas	ion, nun, oomf, tf, (;, finna, dang, fa, (:, <<, >>, <-, .!
Houston	trippin, y'all
Houston	mayne, fwm, jammin, shid, jamming, tripping, azz, bck, ma'am , bae, whoop, ole, sho, fck, lowkey, lawd, fa, trippin
Philadelphia	ard, jawn, cdfu, bul, wya, 1omf, jawns, ctfu, ctfuu, hbu, rd, foh, sike, hype, nut, bull
Washington DC	lt, lrt, llss, bait, fakin, stamp, ji, brova, siced, hu, wholetime, guh
Miami	bol, jit, bih, vibe
Atlanta	preciate, fye, frfr, slick, shid, fr, ain, ikr, followback, flex, gotcha
Boston	legit, deff, gunna

4.3.2 *Tweetspeak* variables

In addition to the geographically-specific lexical innovations, tweets also contain other popular variants which are non-standard. We create a list of Twitter specific non-standard lexical innovations, which we refer as “Tweetspeak”, based on the 1,000 most frequent terms in our tweet sample. From this list of most frequent terms we removed words that are in standard dictionaries, words that refer to entities, punctuations, special symbols, numbers, hashtags, and non-English words. Uncertain cases were resolved by manual inspection of a set of random tweet examples. In total the Tweetspeak list contains 94 non-standard lexical innovations as shown in Table 4.2.

4.4 Data

We used a dataset of 114M geo-tagged tweets from 2.77M different user accounts obtained from the public Gardenhose/Decahose [96] from June 2009 to May 2012.

Table 4.2: List of Tweetspeak lexical variables.

lol	:)	dat	lmaoo	ohh	dm	luv
im	jus	dnt	lmfao	wats	<3	pics
lmao	ppl	idk	hmm	ahh	didnt	ii
ya	lil	aww	w/	nah	naw	comin
haha	wat	wtf	fuckin	umm	fam	tryin
nigga	yall	thats	talkin	outta	congrats	b4
da	yu	imma	tv	bday	yess	wassup
dont	omg	bro	:(def	noo	;))
aint	goin	tryna	yay	lookin	oo	nothin
cuz	knw	hahaha	til	tha	watchin	abt
soo	ima	af	gon	shyt	a lot	
mm	lmfaoo	txt	sayin	s/o	finna	
feelin	dat	gettin	dang	chillin	app	
smh	niggas	doin	pic	tht	bc	

This dataset included only the tweets geo-located within the United States. We did several preprocessing steps to remove retweets (repetitions of previously posted messages), tweets with URL (automated or marketing oriented). Before the analysis, we performed two textual preprocessing steps: tokenizing all the tweets using the Ttokenize program² and downcasing.

4.4.1 Building a balanced corpus

As we have seen in Chapter 3, the volume of tweets varies for each metropolitan area; further, the language of Twitter users differs based on extralinguistic variables. Different Twitter users can also have different habitual use of Twitter: more interaction or more broadcast messages. Therefore, we need to consider such potential confounds when constructing the corpus for our analyses. We used resampling to create a balanced corpus ensuring that each metropolitan area and each author contributes the same number of “positive” and “negative” messages, where positive messages contain any of the non-standard lexical innovations, and the negative messages contain none.

²The software is available at <https://code.google.com/p/ark-tweet-nlp/>

We first identified the list of all tweets that contain each non-standard variable and randomly sampled a single message to our set of positive messages. We then sampled a message from the same author, which does not contain any of the non-standard variables in our lexicon. We repeated this process until there are a total of 1,000 tweets for each of the linguistic variable. In this way, our balanced corpus contains 224,000 tweets for the geographical variables, and 188,000 tweets for the Tweetspeak variables.

4.4.2 Geolocating tweet recipients

GPS location of the tweet sender is available from the metadata of each tweet in our dataset and we geo locate each sender to a Metropolitan Statistical Area (MSA). However, to investigate if the use of non-standard lexical variables is influenced by the location of the users mentioned in the tweet, we need to obtain the geo location of the users mentioned in the tweets. Because this information is not available from the tweet metadata and it is not feasible to query the Twitter API to obtain this information, we opted for an alternative approach. We linked the username to the location of the tweets that mention them [188]. If a username is mentioned by at least three different individuals within a given metropolitan area, and is never mentioned by anyone from outside that MSA, then we can guess with reasonable confidence that this MSA is the correct location for that username. We treated usernames which do not meet this criterion as unknown. Although this threshold could be overly strict, we opted for a high-precision heuristic.

4.5 Analysis and Results

We treat testing the two hypotheses mentioned in § 4.1 as binary prediction problems—whether a non-standard linguistic innovation is used in a tweet or not. Following the standard practices in variationist sociolinguistics [189], we use two logistic regression

Table 4.3: Predictors used in each model.

Predictor	Model-I	Model-II	Description
<i>Local audience features</i>			
@-INIT-SAME-METRO		x	message begins with the username of an individual from the same MSA
@-INTERNAL-SAME-METRO		x	message contains a username of an individual from the same MSA, but not at the beginning
<i>Limited audience features</i>			
@-INIT	x	x	message begins with a username
@-INTERNAL	x	x	message contains a username, but not at the beginning
<i>Wider audience features</i>			
#-INIT	x	x	message begins with a hashtag
#-INTERNAL	x	x	message contains a hashtag, but not at the beginning
<i>Controls</i>			
NUM-WORDS- <i>i</i>	x	x	message contains exactly <i>i</i> tokens

models. In the next two subsections, I present the details of the models and the results. Details of the predictors for each model are shown in Table 4.3. Both of these models test the frequency of non-standard linguistic innovations comparing to the baseline of broadcast messages (no use of @-mentions or hashtags). Therefore a positive coefficient for a message type predictor indicates greater tendency towards using a non-standard variant in that type of message compared broadcast messages, and a negative coefficient for a predictor indicates an inhibition towards using a non-standard variant compared to broadcast messages.

4.5.1 Model-I

Dependent variable: Whether a tweet has any of the non-standard linguistic innovations?

Predictors: @-INIT, @-INTERNAL, #-INIT, #-INTERNAL

Controls: Length of the message in number of words

Results: Table 4.4 and Table 4.5 show the results of the logistic regression analysis

Table 4.4: Results for Model-I predictors and geographical lexical variables. Statistical significance is indicated with asterisks, *** : $p < 0.01$, ** : $p < 0.05$, * : $p > 0.05$. ‘Weight’ is the logistic transformation of the logistic regression coefficient, yielding a value between 0 and 1, with 0.5 indicating indifference. ‘Empirical %’ indicates the percentage of messages with each predictor which include a non-standard variable, and ‘N’ is the total number of messages with each predictor.

Feature	Weight	Coefficient	95% C.I.	Empirical %	N
<i>Limited audience</i>					
@-INIT	0.5701	0.2821 ***	[0.264, 0.300]	51.85	96,954
@-INTERNAL	0.5827	0.3340 ***	[0.299, 0.369]	56.41	15,494
<i>Wider audience</i>					
#-INIT	0.4004	-0.4037 ***	[-0.453, -0.355]	35.86	7,980
#-INTERNAL	0.4891	-0.0437 ***	[-0.076, -0.011]	50.40	16,937
<i>Range</i>	18				
<i>Total</i>				50.00	224,000

for Model-I predictors for the geographically-specific lexical variables and Tweetspeak variables respectively. Both models show broadly similar trends for the predictors: (1) both @-INIT and @-INTERNAL predictors show strong positive association with the use of non-standard variables, which indicates that the non-standard linguistic innovations are used more often in messages directed to smaller audience; (2) both #-INIT and #-INTERNAL predictors show strong negative association with the use of non-standard variables, which indicates larger audience size inhibits the use of the non-standard linguistic innovations.

4.5.2 Model-II

Dependent variable: Whether a tweet has any of the non-standard linguistic innovations?

Predictors: @-INIT-SAME-METRO, @-INTERNAL-SAME-METRO, @-INIT, @-INTERNAL, #-INIT, #-INTERNAL

Controls: Length of the message in number of words

Results: Table 4.6 and Table 4.7 show the results of the logistic regression analysis

Table 4.5: Results for Model-I predictors and Tweetspeak lexical variables.

Feature	Weight	Coefficient	95% C.I.	Empirical %	N
<i>Limited audience</i>					
@-INIT	0.5997	0.4042***	[0.384, 0.425]	53.12	78,047
@-INTERNAL	0.5826	0.3333***	[0.294, 0.373]	54.12	12,076
<i>Wider audience</i>					
#-INIT	0.4079	-0.3728***	[-0.423, -0.323]	34.72	8,062
#-INTERNAL	0.4814	-0.0743***	[-0.108, -0.041]	49.10	16,472
<i>Range</i>	<i>19</i>				
<i>Total</i>				<i>50.00</i>	<i>188,000</i>

for Model-II predictors for the geographically-specific lexical variables and Tweetspeak variables respectively. Both models show broadly similar trends for the predictors. The small, but statistically significant positive coefficients for the new predictors @-INIT-SAME-METRO and @-INTERNAL-SAME-METRO indicate that the non-standard linguistic innovations are especially likely to be used in messages that are directed to individuals from the same metropolitan area as the individual who sent the message. While the coefficient for @-INTERNAL-SAME-METRO is slightly higher than the coefficient for @-INIT-SAME-METRO, the overlapping confidence intervals indicate that the difference is not statistically significant—although both are significantly greater than zero.

4.6 Conclusions

In this chapter I examined the social meaning of varied usage of non-standard language in tweets using the sociolinguistic construct of audience design. First I explained how we characterized the size of different audience on Twitter using the features of the platforms. Then I described the extraction of two sets of linguistic variables: geography specific non-standard forms and Twitter-specific terms using an automated data-driven approach. After describing the dataset construction, I detailed two

Table 4.6: Model-II predictors and geographical lexical variables.

Feature	Weight	Coefficient	95% C.I.	Empirical %	N
<i>Local audience</i>					
@-INIT-SAME-METRO	0.5225	0.0900***	[0.055, 0.126]	53.23	14,976
@-INTERNAL-SAME-METRO	0.5272	0.1089**	[0.016, 0.202]	58.59	2,248
<i>Limited audience</i>					
@-INIT	0.5667	0.2685***	[0.249, 0.288]	51.85	96,954
@-INTERNAL	0.5789	0.3182***	[0.281, 0.355]	56.41	15,494
<i>Wider audience</i>					
#-INIT	0.4006	-0.4031***	[-0.452, -0.354]	35.86	7,980
#-INTERNAL	0.4894	-0.0424***	[-0.075, -0.010]	50.40	16,937
<i>Range</i>	<i>18</i>				
<i>Total</i>				<i>50.00</i>	<i>224,000</i>

Table 4.7: Model-II predictors and geographical lexical variables.

Feature	Weight	Coefficient	95% C.I.	Empirical %	N
<i>Local audience</i>					
@-INIT-SAME-METRO	0.5247	0.0990***	[0.050, 0.148]	53.64	7,349
@-INTERNAL-SAME-METRO	0.5523	0.2100***	[0.075, 0.345]	58.09	995
<i>Limited audience</i>					
@-INIT	0.5976	0.3954***	[0.375, 0.416]	53.12	78,047
@-INTERNAL	0.5783	0.3160***	[0.275, 0.357]	54.12	12,076
<i>Wider audience</i>					
#-INIT	0.4080	-0.3721***	[-0.422,-0.322]	34.72	8,062
#-INTERNAL	0.4816	-0.0735***	[-0.107,-0.040]	49.10	164,72
<i>Range</i>	<i>19</i>				
<i>Total</i>				<i>50.00</i>	<i>188,000</i>

analyses we performed to test two hypotheses related to the audience size and stylistic variation.

Results from both analyses indicate a clear and consistent relationship between the different audience-selection affordances of Twitter and the frequency of the non-standard variants used in messages targeted at different intended audience. Non-standard variants are more likely to be used in messages targeted at smaller and more geographically local messages; but when the intended audience size becomes larger and less geographically local, then non-standard variants are less likely to be used. These findings suggest that Twitter users utilize the socio-technical affordances of the platform to control their intended audience and self-present them differently using different linguistic innovations that are geographically-differentiated or emerged on Twitter to convey varied social meaning.

In this chapter we have looked at the social meaning of intra-person variation in the usage non-standard forms that are unique to online social platforms. In the next chapter, I focus on a special form of non-standard orthography that was recently introduced to social media, and investigate how it affects individual linguistic style.

CHAPTER 5

LINGUISTIC INNOVATIONS AND THE COMPETITION FOR PARALINGUISTIC FUNCTIONS¹

In the previous chapter I explained how we examined the social meaning of varying usage of non-standard language in online writing. We considered only the ASCII-based linguistic innovations such as emoticons and non-standard spellings. The socio-technical platforms in which we conduct interpersonal interactions also provide new technological and linguistic affordances. Recently, a new form writing or orthography, called *emojis*, was introduced to online writing. Emojis are a pre-defined set of picture characters based on the Unicode character encoding systems. Emojis look similar to emoticons, but they are more colorful and increasing in number as new icons are introduced frequently. This raises the question of whether emojis are performing the same linguistic role of emoticons, that is to express non-verbal cues in online writing. An even bigger question is whether the pre-defined symbols of emojis are going to replace ASCII-based linguistic creativity such as emoticons and make online writing more standard, but less creative?

In this chapter, I describe our examination of the impact of this new technological affordance of emojis on online writing style using a causal inference approach. The correlations between emoticon and emoji usage over time will not adequately answer the question of whether the introduction of emojis causes a decline in other forms of ASCII-based non-standard orthography. Therefore we use causal statistical analysis to measure the effect of emoji adoption on individuals' writing style. In this chapter, I will first explain the procedure for extracting emojis, emoticons, and a lexicon of

¹Content of this chapter is based on Umashanthi Pavalanathan and Jacob Eisenstein. "More Emojis, Less :) The Competition for Paralinguistic Functions in Microblog Writing.", First Monday, November 2016.

standard words. Then I will explain our study design of an approximated randomized experiment using observational data from Twitter, the procedure for selecting the treatment and control users, and the procedure for matching each of the treatment user with a control user in order to eliminate the potential confounds of observational data. Following that, I will detail the difference-in-difference method of causal statistical inference which we used to measure the treatment effect of emoji adoption on writing style. Finally, I will describe the causal experiments and summarize findings. Similar to the previous chapter, the study presented in this chapter also focuses on the dimensions of *stylistic innovations* and *intra-person variation*.

5.1 Background

Social media and other online platforms provide a new set of affordances for interpersonal communication. Affordances of online platforms, including technological, social, and linguistic means, shape online interactions and online writing in several aspects. For example, the work discussed in Chapter 4 looked at how technological and social affordances of Twitter help users to navigate their audience and self-present them accordingly using different stylistic innovations in writing. Recently, a new type of linguistic affordance was introduced to many online platforms, including Twitter. It is the pictograph characters known as “emojis”. Figure 5.1 shows a list of emojis. As emojis look visually similar to the previously well-known non-standard variant of emoticons, the introduction of emojis raises several questions about the changes in orthographic writing styles in online platforms: Do emojis play the same linguistic role as emoticons and other non-standard terms? If so, will emojis replace previous forms of non-standard terms?

Although emojis look visually similar to emoticons, they differ from emoticons and other non-standard forms in one important aspect: emoticons and other non-standard variants like abbreviations, expressive lengthenings, and phonetic spellings are creative



Figure 5.1: Examples of emoji characters used in Twitter (created using <http://www.iemoji.com>).

combination of multiple ASCII characters; but emojis are a predefined set of symbols created by the unicode consortium. These bottom-up and top-down orthographic differences make the competition between emojis and previous non-standard forms linguistically important.

We formulate the question about the effects of emoji adoption on online writing styles as a causal statistical inference question, considering the introduction of emojis as the *treatment* and changes in individuals' writing style as the *effect* of the treatment. Specifically, we evaluate the following two hypotheses:

Hypothesis I: Adopting emojis causes writers to use fewer emoticons.

Hypothesis I: Adopting emojis causes writers to use fewer ASCII-based non-standard orthographies overall.

The first hypothesis would suggest that emojis and emoticons are competing for a shared linguistic function and the second hypothesis would suggest that emojis are displacing all forms of non-standard orthography. We can consider the first hypothesis as a special case of the second hypothesis. As simple correlation analyses would not capture the causal claims, we use a matching approach to causal inference to evaluate both hypotheses.

5.2 Dataset

For our analyses we used a corpus of tweets from February 2014 to August 2015, collected continuously querying the Twitter streaming API.

Selecting the study population: As our analyses focus on individuals’ writing style, we used several criteria to select the study population: we removed users who have more than 5,000 followers or followees and more than 200 tweets on average each month. We further removed retweets, messages that were not originally posted by a user.

To quantify the writing style of the study population, in addition to emojis, we use two other sets of lexicons: *emoticon tokens* and *standard tokens*. We define the *lexicon usage rate* as the ratio of the number of tokens that appear in a lexicon to the total number of tokens.

Extracting emoji tokens: We extracted the emoji tokens by converting the tweets into unicode representation and then using regular expressions to extract the unicode characters in the ranges of the “Emoji & Pictographs” category of unicode symbols. Through this process, we extracted 1,235 unique emoji characters from a random sample of tweets.

Extracting emoticon tokens: We followed a data-driven approach to extract emoticons, as there is no comprehensive list of Twitter emoticons and new emoticons get introduced over time. We used a set of heuristics to write regular expressions (e.g., two or more non-alpha numeric characters) and extracted an initial list of candidate tokens and then through manual inspection of a random examples of usage for each token, we created a final list of 44 and 52 unique emoticons from tweets of March 2014 and March 2015.² Extracted emoticon tokens are shown in Appendix B (Table B.1).

Extracting standard tokens: Because our second hypothesis tests the changes in non-standard language usage and language becoming more standard, we create a list of words that are widely used in online writing and considered to be standard. As standard dictionaries such as the Unix dictionary do not contain enough coverage for entity names and new words, we used a dump of the English Wikipedia articles and

²Reasons for the selection of this time period is discussed later in the Study Design section.

Table 5.1: Standard word list statistics

Source	No. of unique terms
Redhat Unix Dictionary	479,829
Ubuntu Unix Dictionary	99,171
English Wikipedia Corpus	266,695 ³
Combined List	621,968

augmented the dictionary words with the words that appear in more than 50 different Wikipedia articles. In total, the standard token list contains 621,968 unique tokens. Details about the sources of tokens are shown in Table 5.1.

5.3 Study Design

We evaluate both of our hypotheses using causal statistical inference. Specifically we test the hypotheses using the following causal inference questions: (1) whether the rise in emojis causes a decline in emoticon usage rate, and (2) whether the rise in emojis causes an increase in standard token usage rate.

In a completely randomized experiment, we would select a group of individuals who have never heard of emojis and randomly assign them to treatment and control conditions, and then do some intervention to the treatment group to use emojis. The effect of the treatment would be the difference in the lexicon usage rates between the treatment and control groups after the treatment. As it is not feasible to conduct such idealized experiment, we approximate this randomized experiment using observational data from Twitter.

Treatment and Control groups: As Twitter introduced emojis to its web interface in April 2014, following emoji support on Twitter’s apps for Android and iOS, we considered the month of March 2014 as the *pre-treatment* period and the month of March 2015 as the *post-treatment* period. As shown in Figure 5.2, we selected users

³Terms which appeared in more than 50 different articles.

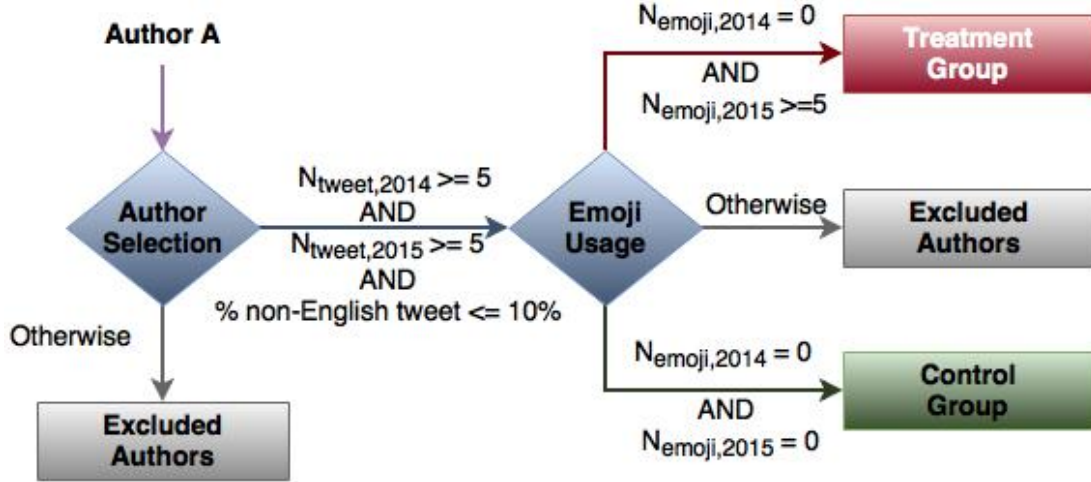


Figure 5.2: Selection of authors into treatment and control groups.

who had not used any emojis in March 2014, but used at least five emojis in March 2015 into the treatment group, and we selected users who had not used emojis in both March 2014 and March 2015 into the control group.

5.3.1 Confounds

As we approximate a randomized experiment using observational data, users are not randomly assigned to the treatment and control groups. Therefore, if there are any confounds that affect both the treatment and outcome, then the treatment effect estimation will be flawed. Although it is not feasible to control for all the confounds that affect individual writing styles, we identified two confounds that could impact the treatment effect estimation: *lexicon usage* and *source of the tweet*.

Lexicon usage: Writing style of Twitter users might differ based on their use of standard and non-standard words. As these writing style preferences could influence the adoption of emojis, we control for this confound by matching on the pre-treatment lexicon usage rates of the treatment and control users.

Source of the tweet: Different Twitter clients (e.g., Twitter client for iPhone vs. Twitter extension for desktop Firefox browser) have varied text input affordances and auto-correction capabilities which impact the creation of tweet messages. These

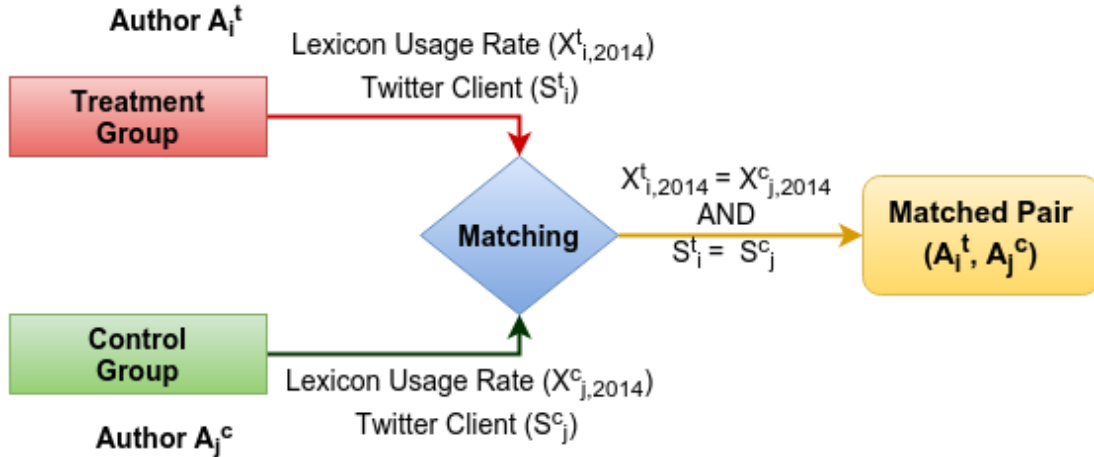


Figure 5.3: Matching procedure.

different characteristics of the clients might impact the treatment, which is the adoption of emojis, and the outcome, which is the change in writing style. Therefore, we also match on the source of the tweet.

5.3.2 Matching

We use the technique of matching [159, 190], to ensure that both the treatment and control groups have similar pre-treatment characteristics. To test Hypothesis-I, we matched the users in the treatment and control group based on the pre-treatment emoticon usage rate and their majority source of tweets (yielded 3,112 treatment-control pairs). To test Hypothesis-II, we matched the users in the treatment and control groups based on their pre-treatment standard token usage rate and their majority source of tweets (yielded 3,115 treatment-control pairs). Figure 5.3 shows a schematic diagram of the matching process.

5.3.3 Estimation of Treatment Effects

We used the *difference-in-difference* method [191] for treatment effect estimation and evaluated our hypotheses. As the overall lexicon usage rates change over time, we need to account for the control group’s outcome as well. The difference-in-difference

approach accounts for this by calculating the effect of a treatment by comparing the average change in the outcome variable after the treatment for both the treatment and control groups. The following quantities describe the treatment effect estimation process:

Lexicon usage rates

- $X_{i,pre}^t$ pre-treatment lexicon usage rates for author i , who is in the treatment group t
- $X_{j,pre}^c$ pre-treatment lexicon usage rates for author j , who is in the control group c
- $X_{i,post}^t$ post-treatment lexicon usage rates for author i , who is in the treatment group t
- $X_{j,post}^c$ post-treatment lexicon usage rates for author j , who is in the control group c

Differences

- Y_i^t the difference between the post-treatment and pre-treatment lexicon usage rates for author i , who is in the treatment group t

$$Y_i^t = X_{i,post}^t - X_{i,pre}^t$$

- Y_j^c the difference between the post-treatment and pre-treatment lexicon usage rates for author j , who is in the control group c

$$Y_j^c = X_{j,post}^c - X_{j,pre}^c$$

- \bar{Y}^t the average difference between post-treatment and pre-treatment lexicon usage rates for the treatment group which has n_t authors

$$\bar{Y}^t = \frac{1}{n_t} \sum_{i \in T} Y_i^t$$

- \bar{Y}^c the average difference between post-treatment and pre-treatment lexicon usage rates for the control group which has n_c authors

$$\bar{Y}^c = \frac{1}{n_c} \sum_{j \in C} Y_j^c$$

We can then define the average treatment effect as,

$$ATE = \bar{Y}^t - \bar{Y}^c.$$

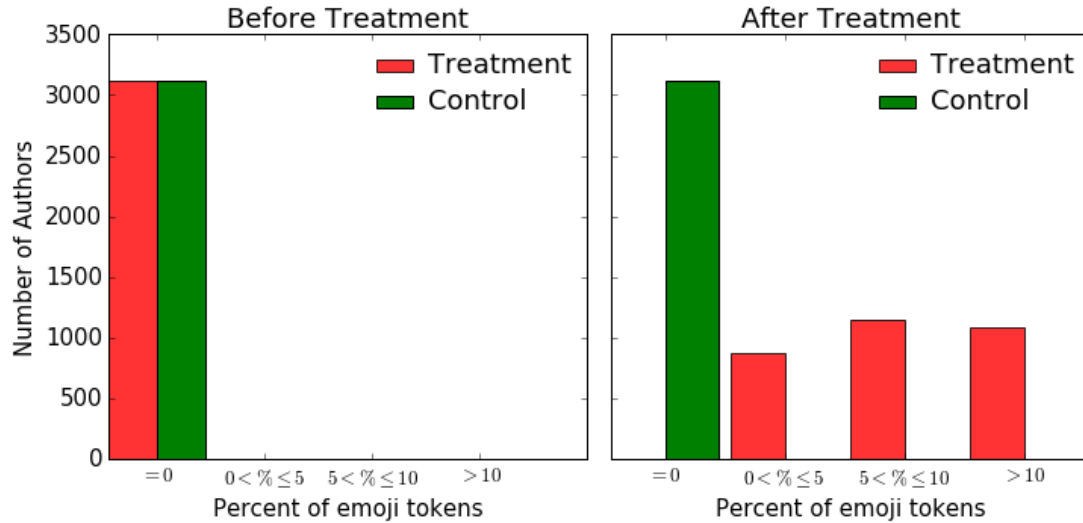


Figure 5.4: Analysis-I: Emoji usage of treatment and control groups.

Our **null hypothesis** in both the analyses is that there is no treatment effect and any observed differences are due to chance variation in a finite sample.

5.4 Causal Inference Analysis

5.4.1 Analysis-I: Effects of Emoji Adoption on Emoticon Usage

We test the hypothesis that using emojis causes a decline in individuals' emoticon usage in the first analysis. Figure 5.4 and Figure 5.5 show the distribution of the emoji and emoticon usage rates for both the treatment and control groups before and after the treatment. Based on our study design, both the treatment and control groups have zero emoji usage during pre-treatment period and the treatment group has a non-zero emoji usage rate during the post-treatment period while the control group remains the same. Based on our matching procedure, both the treatment and control groups have similar emoticon usage rate (0.40%) before the treatment (Figure 5.5 left).

There is a decrease in emoticon usage for the treatment group after the treatment. Specifically, after the treatment, the treatment group has an average emoticon token

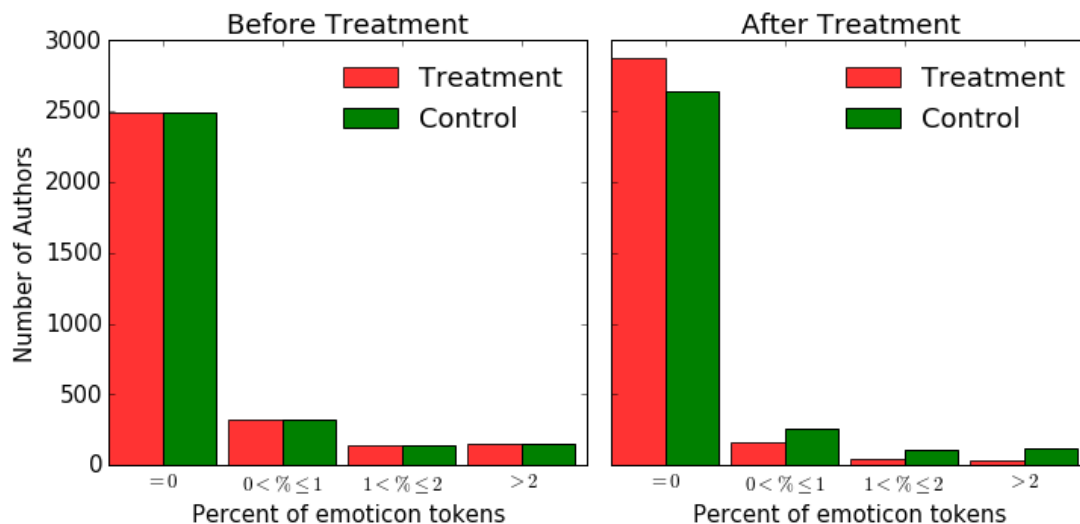


Figure 5.5: Analysis-I results: Emoticon usage of treatment and control groups.

usage rate of 0.12%, while the control group has an average rate of 0.31%. The average treatment effect is a 0.19% decrease in emoticon usage per token, which indicates that after the adoption of emojis, the treatment group is 2.5 times less likely to use emoticons than the control group. These differences are statistically significant for a paired t-test ($t = -9.612$ at $p < 10^{-21}$). Summary of the results are shown in Table 5.2.

5.4.2 Analysis-II: Effects of Emoji Adoption in Standard Word Usage

We test the hypothesis that emoji adoption leads to language becoming more standard in the second analysis. Figure 5.6 and Figure 5.7 show the distribution of the emoji and standard token usage rates for both the treatment and control groups before and after the treatment. Based on our study design, both the treatment and control groups have zero emoji usage during pre-treatment period and the treatment group has a non-zero emoji usage rate during the post-treatment period while the control group remains the same. Based on our matching procedure, both the treatment and control groups have similar standard token usage rate (85.2%) before the treatment (Figure 5.7 left).

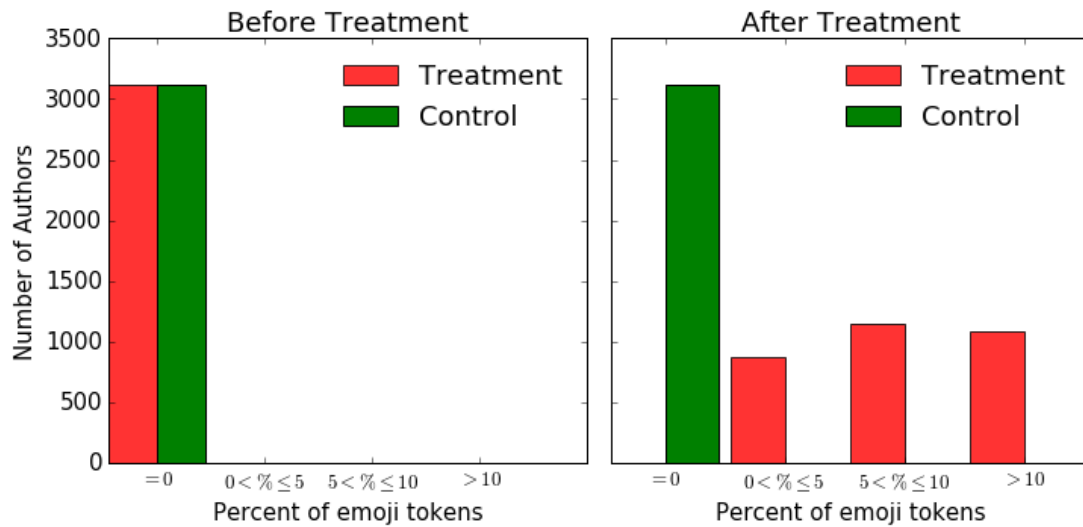


Figure 5.6: Analysis-II: Emoji usage of treatment and control groups.

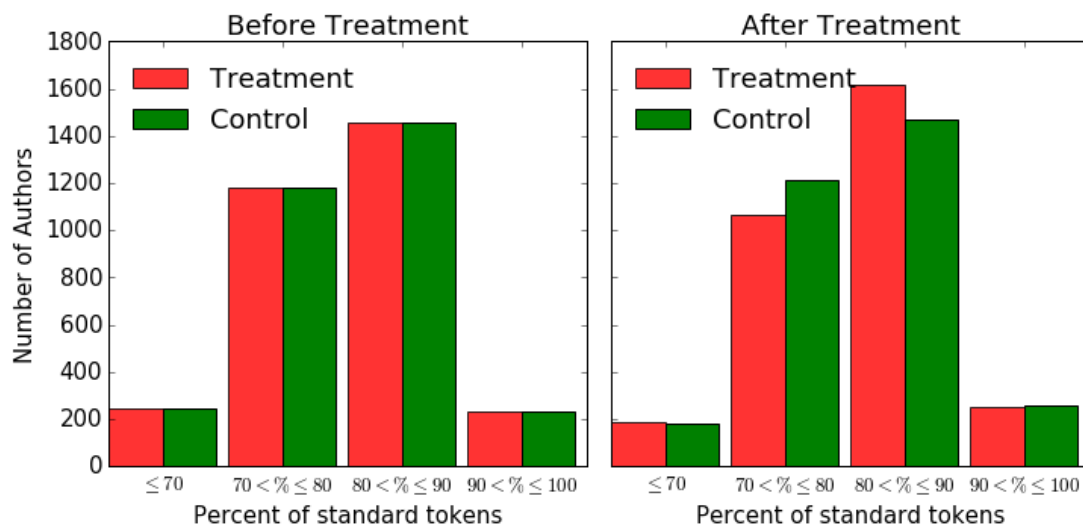


Figure 5.7: Analysis-II results: Standard token usage of treatment and control groups.

There is an increase in standard token usage for the treatment group after the treatment. Specifically, after the treatment, the treatment group has an average standard token usage rate of 86.2%, while the control group has an average rate of 85.7%. The average treatment affect is a 0.45% increase in standard terms usage per token, which indicates that after the adoption of emojis, the treatment group’s language become more standard compared to the control group. These differences are statistically significant for a paired t-test ($t = -2.755$ at $p < 10^{-3}$). Summary of the results are shown in Table 5.2.

Post-hoc placebo analysis: To examine whether the selection of our pre-treatment and post-treatment periods are sensitive to our findings, we performed a placebo-type analysis and computed the average differences in lexicon usage considering the periods adjacent to the pre-treatment and post-treatment periods. Results of the analysis showed no statistically significant of the average differences, which suggests that the treatment effects we measure are indeed due to the adoption of emojis.

5.5 Conclusions

In this chapter, I have described our investigation of how new linguistic and technological affordances shape intra-person stylistic variation. Specifically, I explained how we systematically studied the impact of the adoption of emojis, a new form of writing introduced to social media recently. I explained how we extracted emojis, emoticons, and a lexicon of standard words to quantify the writing style on Twitter. Then, I explained our study design of an approximated randomized experiment including the selection of treatment and control users using the technique of matching. Following that, I described the procedure we used to measure treatment effect of emoji adoption and tested two hypotheses related to the writing style in online language.

The finding that Twitter users who adopt emojis tend to reduce their usage of

Table 5.2: Causal inference results

	Pre-treatment Period		Post-treatment Period		t-stat	p-value	Average Treatment Effect
	Treatment group	Control group	Treatment group	Control group			
Analysis-I							
Average % of emoticon token rate	0.40%	0.40%	0.12%	0.31%	-9.612	$1.01e^{-21}$	0.19% ↓
Analysis-II							
Average % of standard token rate	85.20%	85.17%	86.15%	85.66%	2.755	$5.89e^{-3}$	0.45% ↑

emoticons, in comparison to the matched users who do not use emojis, supports our first hypothesis: the increasing frequency in emojis leads to a corresponding decline in emoticons. Of course, since Twitter has a restriction on the number of characters, in some sense all features compete for space, but not necessarily for the same linguistic functions. Nonetheless, the overwhelming majority of Twitter messages are not near the character limit [5], indicating that this is unlikely to be the main reason for the decrease in emoticon characters. Given evidence that emoticons and emojis both perform a paralinguistic function [30, 80], it seems more likely that the replacement of emoticons by emojis reflects the ongoing success of emojis in a competition for this role.

Results from our second analysis show that the users who adopt emojis tend to use standard tokens at an increased rate after emoji adoption. This suggests that emojis are in competition with the entire range of non-standard orthographies, including expressive lengthening (e.g., *coooooo!!!!*), non-standard words (e.g., *gud*), and abbreviations (e.g., *lol*). If these trends continue, emojis may be a sociotechnical solution to the “problem” of non-standard orthography, representing a rare case in which an institutionally-driven change from above succeeds in displacing bottom up linguistic variation.

This chapter and the previous chapter focused on the dimensions of stylistic innovations and intra-person variation. In the next chapter I move beyond this and focus on community-level linguistic style variation.

CHAPTER 6

MULTI-COMMUNITY STYLE VARIATION: A MULTIDIMENSIONAL LEXICON FOR INTERPERSONAL STANCETAKING¹

In the last two chapters I presented two studies looking at stylistic innovations and intra-person variation. In the next three chapters, I move beyond this and look at community level linguistic norms and how community linguistic styles affect intra-person variation. In this chapter I focus on studying the interactional style aspects in the multi-community environment of Reddit. Reddit is a large online community which is a collection of several sub forums called subreddits. Each subreddit is dedicated to discuss various topics through discussions. The conversational nature of Reddit makes it ideal for studying interactional style.

Prior computational work on interactional styles primarily focused on single social impressions such as politeness and formality. However, interactional styles span a range of interpersonal phenomena. In order to study how different linguistic resources are arrayed to create such social impressions, in this chapter I will explain how we quantitatively operationalize the sociolinguistic concept of interpersonal stancetaking. Stancetaking reflect the relationship of speakers or writers to the audience, topic and the talk itself. Although previous work has used crowd sourced annotations to characterize phenomena such as politeness and formality, such an approach is not appropriate for stancetaking because there is no pre-defined set of stances. In this chapter, I will explain how we use a data-driven unsupervised method to extract the latent dimensions of stancetaking by grouping markers of stance based on their

¹Content of this chapter is based on Umashanthi Pavalanathan, Jim Fitzpatrick, Scott F. Kiesling, and Jacob Eisenstein. “A multidimensional Lexicon for Interpersonal Stancetaking.”, ACL 2017.

underlying linguistic functions. After presenting the output of the model, I will describe how we evaluate the model using intrinsic and extrinsic evaluations. The work presented in this chapter focuses on the dimensions of *community norms*, *intra-person*, and *inter-community variation*.

6.1 Background

Online communities are a space for groups of individuals to come together around mutual engagement in some activities of shared interest. These communities can be considered as *communities of practice* where shared norms are formed and evolve as the community members interact with each other. Individuals construct different forms of identities through participating in a variety of such communities of practice and a key to this entire process of individual identity construction is the stylistic practices in different communities [41]. How can we study these varied linguistic style practices in large-scale by mining large corpora of online interactions? In this chapter, I provide a framework to study inter-person, intra-person, and inter-community variation, by quantitatively operationalizing the sociolinguistic concept of *interpersonal stancetaking*.

Interpersonal stancetaking represents an attempt to unify concepts such as sentiment, politeness, formality, and subjectivity under a single theoretical framework [110, 111].² The key idea, as articulated by Kiesling [112], is that stancetaking captures the speaker’s relationship to (a) the topic of discussion, (b) the interlocutor or audience, and (c) the talk (or writing) itself. Various configurations of these three legs of the “stance triangle” can account for a range of phenomena. For example, epistemic stance relates to the speaker’s certainty about what is being expressed, while affective stance indicates the speaker’s emotional position with respect to the content [194].

The stancetaking framework has been applied almost exclusively through qualitative

²Stancetaking is distinct from the notion of *stance* which corresponds to a position in a debate [192]. Similarly, Freeman et al. [193] correlate phonetic features with the *strength* of such argumentative stances.

methods, using close readings of individual texts or dialogs to uncover how language is used to position individuals with respect to their interlocutors and readers [195, 196, 197, 198]. But despite its strong theoretical foundation, we are aware of no prior efforts to operationalize stancetaking at scale. Since annotators may not have strong intuitions about stance — for example, in the way that they do about formality [51] and politeness [50] — we cannot rely on the annotation methodologies employed in prior work. We take a different approach, performing a multidimensional analysis of the distribution of likely stance markers.

We attempt the first large-scale operationalization of stancetaking through computational methods. Our approach is based on a theoretically-guided application of unsupervised learning, in the form of factor analysis. Stancetaking is characterized in large part by an array of linguistic features ranging from discourse markers such as *actually* to backchannels such as *yep* [112]. We first compile a lexicon of stance markers, combining prior lexicons from Biber and Finegan [1] and the Switchboard Dialogue Act Corpus [199]. We then extend this lexicon to the social media domain using word embeddings. Finally, we apply a multi-dimensional analysis of co-occurrence patterns to identify a small set of *stance dimensions*.

Our operationalization of stancetaking is based on the induction of lexicons of stance markers. The lexicon-based methodology is related to earlier work from social psychology, such as the General Inquirer [200] and LIWC [61]. In LIWC, the basic categories were identified first, based on psychological constructs (e.g., positive emotion, cognitive processes, drive to power) and syntactic groupings of words and phrases (e.g., pronouns, prepositions, quantifiers). The lexicon designers then manually constructed lexicons for each category, augmenting their intuitions by using distributional statistics to suggest words that may have been missed [201]. In contrast, we follow the approach of Biber [202], using a multidimensional analysis to identify latent groupings of markers based on co-occurrence statistics.

To measure the internal coherence (construct validity) of the extracted stance dimensions, we use a word intrusion task [203] and a set of pre-registered hypotheses [204]. To measure the utility of the stance dimensions, we perform a series of extrinsic evaluations. A predictive evaluation shows that the membership of online communities is determined in part by the interactional stances that predominate in those communities. Furthermore, the induced stance dimensions are shown to align with annotations of politeness and formality.

6.2 Data

Reddit, one of the internet’s largest social media platforms, is a collection of subreddits organized around various topics of interest. As of January 2017, there were more than one million subreddits and nearly 250 million users, discussing topics ranging from politics (*r/politics*) to horror stories (*r/nosleep*).³ Although Reddit was originally designed for sharing hyperlinks, it also provides the ability to post original textual content, submit comments, and to vote on content quality [205]. Reddit’s conversation-like threads are therefore well suited for the study of interpersonal social and linguistic phenomena.

We used an archive of 530 million comments posted on Reddit in 2014, retrieved from the public archive of Reddit comments.⁴ This dataset consists of each post’s textual content, along with metadata that identifies the subreddit, thread, author, and post creation time. More statistics about the full dataset are shown in Table 6.1.

Table 6.1: Dataset size

Subreddits	126,789
Authors	6,401,699
Threads	52,888,024
Comments	531,804,658

³<http://redditmetrics.com/>

⁴https://archive.org/details/2015_reddit_comments_corpus

6.3 Stance Lexicon

Stancetaking can be characterized in part by an array of linguistic features such as hedges (e.g., *might, kind of*), discourse markers (e.g., *actually, I mean*), and backchannels (e.g., *yep, um*). To quantitatively operationalize stancetaking, we first build a lexicon of these markers.

6.3.1 Seed lexicon

We began with the list of 448 stance markers from Biber and Finegan [1], which includes several lexical categories such as certainty adverbs (e.g., *actually, of course, in fact*), affect markers (e.g., *amazing, thankful, sadly*), and hedges (e.g., *kind of, maybe, something like*) among other adverbial, adjectival, verbal, and modal markers of stance. As our domain of focus is online interactions, and as online language differ significantly from the printed text [5] considered when building the lexicon in Biber and Finegan [1], we augment the initial list of stance markers with 74 dialog act markers from the Switchboard Dialog Act Corpus (SWDA) [199]. The SWDA corpus is based on two-sided telephone conversations, which involve dialogic discourse as in online discussions. The final seed lexicon consists of 517 unique markers, from these two sources.

6.3.2 Lexicon expansion

Our seed lexicon is based on genres that are different from online discussions. Different language style varieties are prevalent in different genres to perform the same linguistic functions [206]. To account for the genre differences in stance markers, we expanded the seed lexicon using automated techniques based on distributional statistics. This is similar to prior work on the expansion of sentiment lexicons [207, 208].

We used word embeddings [209] to find words that are distributionally similar

Table 6.2: Stance lexicon: seed and expanded terms.

Example stance markers from Biber and Finegan [1]	
Seed term	Expanded terms
significantly	considerably, substantially, dramatically
certainly	surely, frankly, definitely
incredibly	extremely, unbelievably, exceptionally
Example dialog act markers from Jurafsky et al. [199]	
Seed term	Expanded terms
nope	nah, yup, nevermind
great	fantastic, terrific, excellent

to the markers in the initial seed lexicon. We trained word embeddings on a corpus of 25 million Reddit comments and a vocabulary of 100K most frequent words on Reddit using the structured skip-gram models of WANG2VEC [210], which augments WORD2VEC [209] by accounting for word order information. We used these embeddings to find terms that are similar to each of the seed stance marker and expanded our lexicon by including terms that have cosine similarity of at least 0.75 with respect to the seed marker. Examples of seed markers and related terms expanded using the word embeddings are shown in Table 6.2. The final stance lexicon with the expanded terms contains 812 unique markers.

6.4 Linguistic Dimensions of Stancetaking

Motivated by prior work of Biber [202] on the linguistic analyses of genre variation, we operationalize stancetaking through a multi-dimensional analysis on the distributional statistics of stance markers across subreddits. In a multi-dimensional framework, each dimension of variation can be viewed as a spectrum, which is a better choice than dichotomous differentiations to explain linguistic expressions of stancetaking. We perform the multi-dimensional analysis using singular value decomposition, an approach which has been applied successfully to a wide range of problems in natural

language processing and information retrieval [211]. While Bayesian topic models are an appealing alternative, singular value decomposition is fast and deterministic, with a minimal number of tuning parameters. The resulting latent dimensions from our multi-dimensional analysis can be considered as the grouping of stance markers based on the underlying communicative functions associated with their usage patterns in multiple subreddits.

6.4.1 Extracting Stance Dimensions

Motivated by our interest in inter-community stylistic variation on Reddit and the premise that the distributional differences of linguistic markers reflect socially meaningful communicative norms of the communities, we perform our analysis using the co-occurrences patterns of stance markers and subreddits.⁵

Singular value decomposition is often used in combination with a transformation of the co-occurrence counts by pointwise mutual information [212]. This transformation ensures that each cell in the matrix indicates how much more likely a stance marker is to co-occur with a given subreddit than would happen by chance under an independence assumption. Because negative PMI values tend to be unreliable, we use the positive PMI (PPMI) values, by replacing all negative PMI values with zeros [213]. Therefore, we obtain stance dimensions by applying singular value decomposition to the matrix constructed as follows:

$$X_{m,s} = \left(\log \frac{\Pr(\text{marker} = m, \text{subreddit} = s)}{\Pr(\text{marker} = m) \Pr(\text{subreddit} = s)} \right)_+.$$

Truncated singular value decomposition performs the approximate factorization $X \approx U\Sigma V^\top$, where each row of the matrix U is a k -dimensional description of each stance marker, and each row of V is a k -dimensional description of each subreddit.

⁵We performed a pilot study using the co-occurrences of stance markers and individual users and the resulting dimensions appeared to be less stylistically coherent.

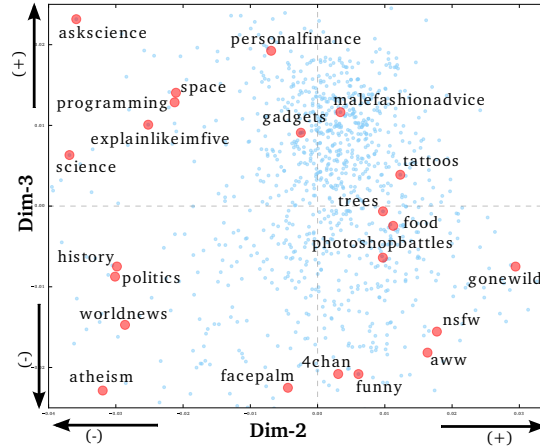


Figure 6.1: Mapping of subreddits in the second and third stance dimensions, highlighting especially popular subreddits.

We included the 7,589 subreddits that received at least 1,000 comments in 2014.

6.4.2 Results: Stance Dimensions

From the SVD analysis, we extracted the six principal latent dimensions that explain the most variation in our dataset.⁶ The decision to include only the first six dimensions was based on the strength of the singular values corresponding to the dimensions. Table 6.3 shows the top five stance markers for each extreme of the six dimensions. The stance dimensions convey a range of concepts, such as involved versus informational language, narrative versus dialogue-oriented writing, standard versus non-standard variation, and positive versus negative affect. Figure 6.1 shows how some of the popular subreddits vary across two of these interactional dimensions. ‘Not Safe For Work (NSFW)’ subreddits *r/gonewild* and *r/aww* map high on dimension-2 and low on dimension-3, indicating involved and informal style of discourse; *r/askscience* and *r/space* map low on dimension-2 and high on dimension-3 indicating informational and formal style.

⁶The choice of six latent dimensions was based on analysis of the associated scree plot [214].

Table 6.3: For each of the six dimensions extracted by our method, we show the five words with the highest loadings.

Dim-1	-	beautifully, pleased, thanks, spectacular, delightful
	+	just, even, all, no, so
Dim-2	-	suggests that, demonstrates, conclude, demonstrated, demonstrate
	+	lovely, awww, hehe, aww, haha
Dim-3	-	funnier, hilarious, disturbing, creepy, funny
	+	thanks, ideally, calculate, estimate, calculation
Dim-4	-	phenomenal, bummed, enjoyed, fantastic, disappointing
	+	hello, thx, hehe, aww, hi
Dim-5	-	lovely, stunning, wonderful, delightful, beautifully
	+	nvm, cmon, smh, lmao, disappointing
Dim-6	-	stunning, fantastic, incredible, amazing, spectacular
	+	anxious, stressed, exhausted, overwhelmed, relieved

6.5 Intrinsic Validation

Evaluating model output against gold-standard annotations is appropriate when there is some notion of a correct answer. As stancetaking is a multidimensional concept, we have taken an unsupervised approach. Therefore, we use evaluation techniques based on the notion of *validity*, which is the extent to which the operationalization of a construct truly captures the intended quantity or concept. Validation techniques for unsupervised content analysis are widely found in the social science literature [215, 216] and have also been recently used in the NLP and machine learning communities [e.g., 203, 217, 204].

We used several methods to validate the stance dimensions extracted from the corpus of Reddit comments. This section describes intrinsic evaluations, which test whether the extracted stance dimensions are linguistically coherent and meaningful, thereby testing the *construct* or *content validity* of the proposed stance dimensions [216]. Extrinsic evaluations are presented in § 6.6.

6.5.1 Word Intrusion Task

A word intrusion task is used to measure the coherence and interpretability of a group of words. Human raters are presented with a list of terms, all but one of which are selected from a target concept; their task is to identify the intruder. If the target concept is internally coherent, human raters should be able to perform this task accurately; if not, their selections should be random. Word intrusion tasks have previously been used to validate the interpretability of topic models [203] and vector space models [217].

We deployed a word intrusion task on Amazon Mechanical Turk (AMT), in which we presented the top four stance markers from one end of a dimension, along with an intruder marker selected from the top four markers of the opposite end of that dimension. In this way, we created four word intrusion tasks for each end of the six dimensions.

Worker selection: We required that the AMT workers (“turkers”) have completed a minimum of 1,000 HITs and have at least 95% approval rate. Furthermore, because our task is based on English language texts, we required the turkers to be native speakers of English living in one of the majority English speaking countries. As a further requirement, we required the turkers to obtain a qualification which involves an English comprehension test similar to the questions in standardized English language tests. These requirements are based on best practices identified in Callison-Burch and Dredze [218].

Task specification: Each AMT human intelligent task (HIT) consists of twelve word intrusion tasks, one for each end of the six dimensions. We provided minimal instructions regarding the task, and did not provide any examples, to avoid introducing

bias.⁷ As a further quality control mechanism, each HIT included three questions which ask the turkers to pick the best synonym for a given word from a list of five answers, where one answer was clearly correct. Turkers who gave incorrect answers for these were to be excluded, but this situation did not arise in practice. Altogether each HIT consists of 15 questions, and was paid US\$1.50. Five different turkers performed each HIT.

Results: We measured the interrater reliability using Krippendorff’s α [219] and the *model precision* metric defined by Chang et al. [203]. Results on both metrics were encouraging. We obtained a value of $\alpha = 0.73$, on a scale where $\alpha = 0$ indicates chance agreement and $\alpha = 1$ indicates perfect agreement. The model precision was 0.82; chance precision is 0.20. To offer a sense of typical values for this metric, Chang et al. [203] report model precisions in the range 0.7–0.83 in their analysis of topic models. Overall, these results indicate that the multi-dimensional analysis has succeeded at identifying dimensions that reflect natural groupings of stance markers.

6.5.2 Pre-registered Hypotheses

Construct validity was also assessed using a set of pre-registered hypotheses. The practice of pre-registering hypotheses before an analysis and testing the correctness is widely used in the social sciences; it was adopted by Sim et al. [204] to evaluate the induction of political ideological models from text. Before performing the multi-dimensional analysis, we identified two groups of hypotheses that are expected to hold with respect to the latent stancetaking dimensions using our prior linguistic knowledge:

- **Hypothesis I:** Stance markers that are synonyms should not appear on the

⁷The prompt for the word intrusions task was: “Select the intruder word/phrase: you will be given a list of five English words/phrases and asked to pick the word/phrase that is least similar to the other four words/phrases when used in online discussion forums”.

Table 6.4: Results for pre-registered hypothesis that stance dimensions will not split synonym pairs.

Stance Dimension	Number of synonym pairs	
	On same end	On opposite ends
DIMENSION 1	6	3
DIMENSION 2	12	2
DIMENSION 3	2	1
DIMENSION 4	11	0
DIMENSION 5	10	2
DIMENSION 6	10	0
Total	51/59	8/59

opposite ends of a stance dimension.

- **Hypothesis II:** If at least one stance marker from a predefined *stance feature group* appears on one end of a stance dimension, then other markers from the same feature group will tend not to appear at the opposite end of the same dimension.

Synonym Pairs: For each marker in our stance lexicon, we extracted synonyms from Wordnet, focusing on markers that appear in only one Wordnet synset, and not including pairs in which one term was an inflection of the other.⁸ Our final list contains 73 synonym pairs (e.g., *eventually/finally*, *grateful/thankful*, *yea/yeah*). Of these pairs, there were 59 cases in which both terms appeared in either the top or bottom 200 positions of a stance dimension. In 51 of these cases, the two terms appeared on the same side (the chance rate would be 50%), supporting the hypothesis and further validating the stance dimensions. More details are shown in Table 6.4.

Stance Feature Groups: Biber and Finegan [1] group stance markers into twelve “feature groups”, such as certainty adverbs, doubt adverbs, affect expressions, and

⁸It is possible that inflections are semantically similar, because by definition they are changes in the form of a word to mark distinctions such as tense, person, or number. However, different inflections of a single word form might be used to mark different stances (e.g., some stances might be associated with the past while others might be associated with the present or future).

Table 6.5: Results for preregistered hypothesis that stance dimensions will align with stance feature groups of Biber and Finegan [1].

Feature group	#Stance marker	χ^2	p -value	Reject null?
Certainty adv.	38	16.94	$4.6e^{-03}$	✓
Doubt adv.	23	13.21	$2.2e^{-02}$	×
Certainty verbs	36	48.99	$2.2e^{-09}$	✓
Doubt verbs	55	30.45	$1.2e^{-05}$	✓
Certainty adj.	28	29.73	$1.7e^{-05}$	✓
Doubt adj.	12	14.80	$1.1e^{-02}$	×
Affect exp.	227	97.17	$2.1e^{-19}$	✓

hedges. Ideally, the stance dimensions should preserve these groupings. To test this, for each of the seven feature groups with at least ten stance markers in the lexicon, we counted the number of terms appearing among the top 200 positions in both ends (high/low) of each dimension. Under the null hypothesis, the stance dimensions are random with respect to the feature groups, so we would expect roughly an equal number of markers on both ends. As shown in Table 6.5, for five of the seven feature groups, it is possible to reject the null hypothesis at $p < .007$, which is the significance threshold after correcting for multiple comparisons. This indicates that the stance dimensions are aligned with predefined stance feature groups.

6.6 Extrinsic Evaluation

The evaluations in the previous section test internal validity; we now describe evaluations testing whether the stance dimensions are relevant to external social and interactional phenomena.

6.6.1 Predicting Cross-posting in Subreddits

Online communities can be considered as *communities of practice* [220], where members come together to engage in shared linguistic practices. These practices evolve simultaneously with membership, coalescing into shared norms. We hypothesize that

Table 6.6: Examples of subreddit pairs that have large and small amount of overlap of contributing members.

Cross-Community Participation	
High-Scoring Pairs	Low-Scoring Pairs
r/blog, r/announcements	r/gonewild, r/leagueofflegends
r/pokemon, r/wheredidthesodago	r/soccer, r/nosleep
r/politics, r/technology	r/programming, r/gonewild
r/LifeProTips, r/dataisbeautiful	r/nfl, r/leagueofflegends
r/Unexpected, r/JusticePorn	r/Minecraft, r/personalfinance

users of Reddit have preferred interactional styles, and that participation in subreddit communities is governed not only by topic interest, but also by these interactional preferences. The proposed stancetaking dimensions provide a simple measure of interactional style, allowing us to test whether it is predictive of community membership decisions.

Classification task: The basic task is to classify pairs of subreddits as *high-crossover* or *low-crossover*, depending on whether individuals are especially likely or unlikely to participate in both subreddits. Individuals are considered to participate in a subreddit if they contribute posts or comments. We compute the pointwise mutual information (PMI) with respect to cross-participation among the 100 most popular subreddits. For each subreddit s , we identify the five highest and lowest PMI pairs $\langle s, t \rangle$, and add these to the high-crossover and low-crossover sets, respectively. Example pairs are shown in Table 6.6. After eliminating redundant pairs, we identify 437 unique high-crossover pairs, and 465 unique low-crossover pairs. All evaluations are based on multiple random training/test splits over this dataset.

Classification approaches: A simple classification approach is to predict that subreddits with similar text will have high crossover. We measure similarity using TF-IDF weighted cosine similarity, using the 8,000 most frequent words on reddit

Table 6.7: Accuracy for prediction of subreddit cross-participation.

Model	Accuracy
BOW-COSINE	66.13%
STANCE-COSINE	64.31%
BOW-SVD	77.48%
STANCE-SVD	84.93%

(STANCE-COSINE), and using the stance lexicon (BOW-COSINE). The similarity threshold between high-crossover and low-crossover pairs was estimated on the training data. We also tested the relevance of multi-dimensional analysis: in the STANCE-SVD model, we used the latent stance dimensions; in the BOW-SVD model, we applied the same multi-dimensional analysis to the 8,000 word vocabulary. For each pair of subreddits, we computed a feature set of the absolute difference across the top six latent dimensions, and applied a logistic regression classifier. Regularization was tuned by internal cross-validation.

Results: Table 6.7 shows average accuracies for these models. The stance-based SVD features are considerably more accurate than the BOW-based SVD features, indicating that interactional style does indeed predict cross-posting behavior. Both are considerably more accurate than the models based on cosine similarity.

6.6.2 Stancetaking Dimensions and Social Phenomena

The utility of the induced stance dimensions depends on their correlation with social phenomena of interest. Prior work has used crowdsourcing to annotate texts for politeness and formality. We evaluate the stancetaking properties of these annotated texts.

Data: We used the politeness corpus of Wikipedia edit requests from Danescu-Niculescu-Mizil et al. [50], which includes the textual content of the edit requests,

along with scalar annotations of politeness. Following the original authors, we compare the text for the messages ranked in the first and fourth quartiles of politeness scores. For formality, we used the corpus from Pavlick and Tetreault [51], focusing on the blogs domain, which is most similar to our domain of Reddit. Each sentence in this corpus was annotated for formality levels from -3 to $+3$. We considered only the sentences with mean formality score greater than $+1$ (more formal) and less than -1 (less formal).

Stance dimensions: For each document in the above datasets, we compute the stance properties, as follows: for each dimension, we compute the total frequency of the hundred most positive terms and the hundred most negative terms, and then take the difference. Instances containing no terms from either list are excluded. We focus on stance dimensions two and five, which are summarized in Table 6.3. Dimension two contrasts informational and argumentative language against emotional and non-standard language. Dimension five contrasts positive and formal language against non-standard and somewhat negative language.

Results: A kernel density plot of the resulting differences is shown in Figure 6.2. The effect sizes of the resulting differences are quantified using Cohen’s d statistic [221]. Effect sizes for all differences are between 0.3 and 0.4, indicating small-to-medium effects — except for the evaluation of formality on dimension five, where the effect size is close to zero. The relatively modest effect sizes are unsurprising, given the short length of the texts. However, these differences lend insight to the relationship between formality and politeness, which may seem to be closely related concepts. On dimension two, it is possible to be polite while using non-standard language such as *hehe* and *awww*, so long as the sentiment expressed is positive; however, these markers are not consistent with formality. On dimension five, we see that positive sentiment terms such as *lovely* and *stunning* are consistent with politeness, but not

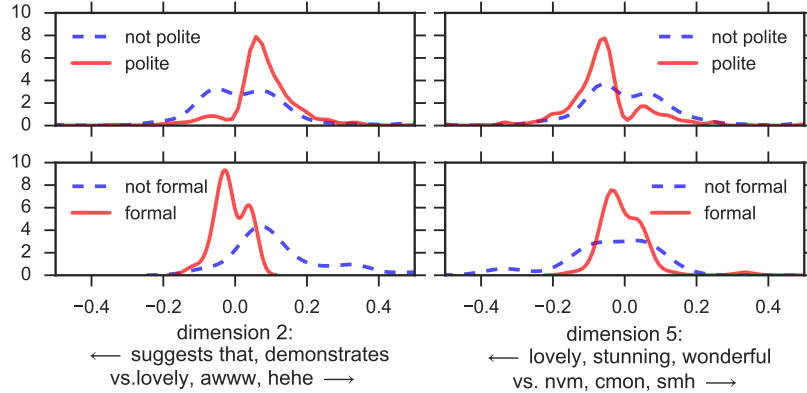


Figure 6.2: Kernel density distributions for stance dimensions 2 and 5, plotted with respect to annotations of politeness and formality.

with formality. Indeed, the distribution of dimension five indicates that both ends of dimension five are consistent only with informal texts.

Overall, these results indicate that interactional phenomena such as politeness and formality are reflected in our stance dimensions, which are induced in an unsupervised manner. Future work may consider the utility of these stance dimensions to predict these social phenomena, particularly in cross-domain settings where lexical classifiers may overfit.

6.7 Conclusions

Stancetaking provides a general perspective on the various linguistic phenomena that structure social interactions. In this chapter I described a quantitative operationalization of the construct of stancetaking to measure linguistic style at scale. Specifically, I described how we extracted a set of several hundred stance markers, building on previously-identified lexicons and using word embeddings to perform lexicon expansion. Then I explained how we used a multidimensional analysis to group these markers into stance dimensions and validated the linguistic coherence and interpretability of the extracted stance dimensions. Then I described how we further validated the model through extrinsic evaluations and demonstrated the utility of the extracted stance

dimensions. This multidimensional stancetaking lexicon captures intra-community linguistic variation along six interactional styles. This lexicon can be used to characterize interactional styles of subreddits and users, and this enables the investigation of several online community writing style related research questions. In the next chapter (Chapter 7), I present a study of linguistic style-shifting in online multi-community environments in which I use the multidimensional stance lexicon to characterize the linguistic styles of communities and individual members. Further, our hope is that this lexicon can also be adapted to capture the interactional styles in other online communities and social media in general.

CHAPTER 7

MULTI-COMMUNITY STYLE VARIATION: LINGUISTIC STYLE-SHIFTING

Continuing the focus on community-level linguistic style variation from the last chapter, in this chapter I explore how community norms and individual writing styles interact in a multi-community setting. Online communities such as Reddit and Stack Exchange house multiple sub-communities on various topics, and also sometimes multiple sub-communities on the same topic [222]. While these sub-communities are housed within the parent sites, most of the time the sub-communities are governed independently by the members of each community through the creation of their own community practices and guidelines [39]. However, the members of these multi-communities create a single account in the parent site and participate in multiple communities using the same account.¹ Similar to community writing styles, individual members also have their own writing style, which of course is shaped by several social contexts as we see in this dissertation. Now, what happens when individuals participate in multiple communities, such as different subreddits on Reddit? Do they use a uniform writing style or do they change their writing style depending on which community they participate in? Another related question is, when individuals participate in multiple communities, do they bring writing style norms from one community to another? This is the focus of this chapter and the work presented in this chapter considers linguistic variation along the broad dimensions of *community norms*, *intra-person*, and *inter-community variation*.

In this chapter, first I will provide some background on multi-communities and

¹While it is possible that some members create multiple accounts for different subreddits, the common practice is to use a single account to participate in multiple subreddits [117, 223].

community norms to motivate the research questions we study. Then I will describe the dataset used for the analysis and the details of four style-shifting models which we formulate to study the research questions. Following that, I report and discuss the results. I conclude the chapter with a discussion of future work.

7.1 Background

Although large scale multi-community analysis is difficult in offline settings, the availability of data capturing interactions in online multi-community environments enables us to study individual member behavior, community linguistic styles, and social dynamics [59]. In online multi-community environments people participate in more than one community [48], and within each of these communities norms emerge over time as members engage in social interactions [40, 224]. As new members join a community and old members gain more authority, the community norms go through continuous change over time. This is in accordance with the notion of “community of practice”, a well-known concept in both sociology [225] and sociolinguistics [41, 220].

The interaction between community linguistic norms, such as community jargons and domain-specific terms, and social dynamics has been investigated in prior work, in the context of individual communities [40, 224]. What happens when people are part of more than one community? Are their interaction styles same in multiple communities or do they adapt to the linguistic style of each community they participate in? How does members’ linguistic style affect the feedback they receive from the community? Large-scale multi-community style accommodation has not been investigated in prior work, and this gap motivates our research questions in this study.

RQ1: Do users participating in multiple subreddits shift style to match the writing style of the subreddit they participate in?

If there is style-shifting, then what are the factors that influence the stylistic

choices of users when they participate in multiple communities? Do community policies or practices influence users' choice of writing style? To shed light on these related questions, we study the following research question:

RQ2: How does the amount of subreddit norm enforcement relate to the amount of style shifting in subreddits?

Multi-community environments, such as Reddit, also enable the study of persistence of linguistic style over consecutive utterances. Persistence is the tendency to repeat a recently used variant in speech or writing. Prior studies of language variation in the linguistics literature have found evidence for persistence in conversational speech [226, 227], for example, persistence in the usage of pronouns [226, 228]. However, we are not aware of any quantitative studies on the persistence of linguistic style. Recent literature argues that a model of language variation should consider both the surface linguistic variation as well as the influence of cognitive forces shaping the variation [227]. Motivated by this literature, in RQ3, we look at persistence of linguistic style when users post to multiple subreddits within a short period of time.

RQ3: Does recency of posting to a subreddit influence the style choices of subsequent posts?

7.2 Dataset

For this study, we use the set of comments posted in Reddit in 2014 obtained from the public archive of Reddit data.² From this dataset we retrieved all comments posted in the top 100 subreddits by subscribers³ and extracted all usernames (i.e., authors of comments). We filtered this list of usernames as follows: removed the 'deleted' username which indicates all accounts that were deleted, removed bots by checking

²<http://files.pushshift.io/reddit/>

³<http://redditlist.com/>

for the presence of the term ‘bot’ in the username, and removed throwaway accounts (anonymous accounts, mostly used for one time posts [229]) by removing usernames containing any of the following variants as done in prior work: ‘thrw’, ‘throwaway’, ‘throw’, and ‘thraway’ [230]. To extract active users, we selected users who have posted at least 100 comments in 2014. As our focus in this work is style-shifting when users post to multiple subreddits, we selected only the users who have participated in at least four different subreddits. After these filters, the resulting set contained 9,921 users. We subsampled a set of 769 users for our analysis and collected their timeline of posts.⁴

7.3 Methods

In this section, I describe how we quantify linguistic style, formulate style-shifting models to study our research questions, implement the models, and evaluate the models.

7.3.1 Measuring Linguistic Style

We use two linguistic measures: (1) the multidimensional stance lexicon developed in the previous chapter (Chapter 6) and (2) lexicons from the linguistic inquiry word count (LIWC) [61]. The stance lexicon is developed to measure the stylistic aspect of language. LIWC is a psycholinguistic measure and has been widely used to characterize the language of online communities (e.g., [46, 224, 230]). In addition to the style lexicon of stance makers, we also use LIWC as a measure in order to compare and contrast the patterns of style-shifting in terms of the use of stance markers and psycholinguistic markers. For the stance measure, we computed the log transformed proportion of the coverage of 100 stance markers from the high and low ends of each of the six stance dimensions (§ 6.4). For the LIWC measure, we counted the

⁴We used subsampling to reduce the data size so that the model running time is reasonable.

number of markers from the following LIWC categories used in prior work on linguistic accommodation [46, 231]: *Article, Certainty, Conjunction, Discrepancy, Exclusive, Inclusive, Indefinite, Negation, Preposition, Quantifier, Tentative, 1st person singular, 1st person plural, and 2nd person pronoun.*

7.3.2 Style-shifting Models

To study the patterns of linguistic style-shifting in multi-community environments, we formulate a set of statistical models. We model the writing style of user u 's post in subreddit r as a function of several covariates. To answer RQ1, we compare the following three models:

No style-shifting:

$$\mathbf{M}_0 : \mathbf{y}_{u,r} \sim f(\boldsymbol{\mu}_u) \tag{7.1}$$

where $\mathbf{y}_{u,r}$ is the style of user u 's posts in subreddit r and $\boldsymbol{\mu}_u$ is an indicator for user u . Bold face notations denote vectors.

Same shifting for all subreddits:

$$\mathbf{M}_1 : \mathbf{y}_{u,r} \sim f(\boldsymbol{\mu}_u + \boldsymbol{\nu}_r) \tag{7.2}$$

where $\boldsymbol{\nu}_r$ is an indicator for subreddit r .

Varying shifting per subreddit:

$$\mathbf{M}_2 : \mathbf{y}_{u,r} \sim f(\alpha_r \boldsymbol{\mu}_u + \boldsymbol{\nu}_r); \alpha_r \in \mathbb{R}_+ \tag{7.3}$$

where α_r is the amount of adaptation of user u in subreddit r . A large value for α_r indicates low style-shifting and vice versa.

To study RQ2, we use the α_r parameter obtained from \mathbf{M}_2 , and analyze its correlation with the following measures of subreddit norm enforcement: (1) number of moderators per active posters in the subreddit,⁵ (2) number of moderators per subscribers, (3) number of moderators per comment, (4) proportion of deleted comments, and (5) average number of deleted comments per moderator.

In RQ3, we are interested in understanding the cognitive aspect of linguistic style-shifting in a multi-community environment by looking at the linguistic style of a user’s consecutive posts in multiple subreddits. We look at the recency or style persistence effect by modeling a user’s linguistic style in the current post to a subreddit as follows: augmenting \mathbf{M}_2 to include the previous subreddits the user has posted to within a period of time prior to the current post. Specifically, we include the subreddit where user u posted immediately before posting to subreddit r in time t (we indicate this as $r(t)$ and the previous subreddit as $r(t - 1)$ where $r(t - 1) \neq r(t)$). To measure persistence over time, we consider the following time bins as time elapsed between the last post in $r(t - 1)$ and the current post in $r(t)$: 1 hour, 6 hours, 12 hours, 1 day, >1 day (up to 10 days).

Persistence of style across subsequent subreddits:

$$\begin{aligned}
 \mathbf{M}_3 : \quad \mathbf{y}_{u,t} &\sim f(\alpha_{r(t)}\boldsymbol{\mu}_u + \boldsymbol{\nu}_{r(t)} + \\
 &\quad 1(\delta_t < t_1)\beta_1\boldsymbol{\nu}_{r(t-1)} + \\
 &\quad 1(t_1 < \delta_t < t_2)\beta_2\boldsymbol{\nu}_{r(t-1)} + \\
 &\quad 1(t_2 < \delta_t < t_3)\beta_3\boldsymbol{\nu}_{r(t-1)} + \dots
 \end{aligned}
 \tag{7.4}$$

where $\delta_t = T(t) - T(t - 1)$, the time difference between the current and immediately previous post. If there is a persistence effect, we would expect the β coefficients to decrease (i.e., $\beta_1 > \beta_2 > \beta_3 \dots$). That means, the shorter the time difference between

⁵users who have at least 10 posts in 2014.

the current post and the immediately previous post, the higher the influence from the writing style of the immediately previous subreddit.

7.3.3 Model Implementation

We implement the models formulated in the previous section using the RStan package [232] for Bayesian statistical inference. Below we provide implementation details for \mathbf{M}_2 . Other models are implemented similarly.

$$\begin{aligned}
 \mathbf{y}_{u,r} &\sim \mathcal{N}(\alpha_r \boldsymbol{\mu}_u + \boldsymbol{\nu}_r, \sigma_y^2) \\
 \boldsymbol{\mu}_u &\sim \mathcal{N}(0, \sigma_u^2) \\
 \boldsymbol{\nu}_r &\sim \mathcal{N}(0, \sigma_r^2) \\
 \alpha_r &\sim \text{Beta}(a, b) \\
 \sigma_u &\sim \text{Gamma}(c, d) \\
 \sigma_r &\sim \text{Gamma}(c, d)
 \end{aligned}
 \tag{7.5}$$

We used $a, b = 2$, and $c, d = 0.001$ for initialization. We fit each model using 10 chains each with 4,000 iterations (2,000 warmup iterations).⁶ Other RStan parameters include: “NUTS” (No-U-Turn sampler) [233] for sampling algorithm, 10 CPU cores for parallelization, and default values for other parameters. We verified model convergence using the Markov chain traceplots and the “Rhat” statistics. For the LIWC models, we model $\mathbf{y}_{u,r}$ as a vector binomial distribution.

7.3.4 Model Evaluation

We compare the performance of models \mathbf{M}_0 - \mathbf{M}_3 using the expected predictive accuracy [234] using the “loo”⁷ R package. The “loo” package estimates the pointwise

⁶Choice of these parameters are based on preliminary modeling, convergence diagnostics, and reasonable running time.

⁷<http://mc-stan.org/users/interfaces/loo>

out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values.⁸

We verify the following hypotheses with regard to each of our research questions:

H₁: If M_1 performs better than M_0 , that would indicate that users deviate from their own writing style and shift style in same amount when posting to different subreddits.

H₂: If M_2 performs better than M_1 , that would indicate that users change their writing style in varying amounts when posting to different subreddits.

H₃: If M_3 performs better than M_2 and if there is a decrease in the β coefficients, that would indicate persistence of users’ linguistic style over time (i.e., recency effect)

7.4 Results

Results for the performance of models M_0 - M_3 are shown in Table 7.1.

Table 7.1: Model evaluation results. We used the multidimensional stance lexicon and LIWC to characterize linguistic style. *Diff(acc.)*: difference in expected predictive accuracy of the two models, *Std. Error*: standard error of the difference. When comparing two models, *A* vs. *B*, a negative value for *Diff(acc.)* indicates that model *A* performs better than model *B*.

Models	Stance		LIWC	
	<i>Diff(acc.)</i>	<i>Std. Error</i>	<i>Diff(acc.)</i>	<i>Std. Error</i>
M_1 vs. M_0	-1718.5	83.1	-6409.8	126.0
M_2 vs. M_1	-436.3	84.9	-146.2	21.9
M_3 vs. M_2	43.5	14.6	89.9	15.2

⁸The specific implementation uses Pareto-smoothed importance sampling for efficient computation [234].

Table 7.2: List of subreddits that enable both high and low style-shifting, ranked based on the α_r parameters of \mathbf{M}_2 . Low α_r indicates more subreddit influence (i.e., more style-shifting) while high α_r indicates less style-shifting.

Stance	High α_r	r/outoftheloop, r/realgirls, r/lifehacks, r/writingprompts, r/relationships
	Low α_r	r/internetisbeautiful, r/android, r/listentothis, r/mildlyinteresting, r/gameofthrones
LIWC	High α_r	r/lifehacks, r/listentothis, r/internetisbeautiful, r/relationships, r/upliftingnews
	Low α_r	r/earthporn, r/gadgets, r/imgoingtohellforthis, r/nsfw, r/fitness

7.4.1 RQ1: Style-shifting

Comparing the performance of \mathbf{M}_1 vs. \mathbf{M}_0 , we find a significant improvement in the predictive accuracy of \mathbf{M}_1 compared to \mathbf{M}_0 , which supports our hypothesis \mathbf{H}_1 . This suggests that the users deviate from their own writing style and shift style in same amount when posting to multiple subreddits.

7.4.2 RQ2: Style-shifting and Subreddit Attributes

Comparing the performance of \mathbf{M}_2 vs. \mathbf{M}_1 , we find a significant improvement in the predictive accuracy of \mathbf{M}_2 compared to \mathbf{M}_1 which supports our hypothesis \mathbf{H}_2 . This suggests that the users change their writing style in varying amounts when posting to different subreddits. Table 7.2 shows a list of subreddits based on their α_r parameter from \mathbf{M}_2 . A low value for α_r indicates high subreddit influence (i.e., high amount of style-shifting when posting to that subreddit), while a high value for α_r indicates low style-shifting. We measured the similarity between the ordering of subreddits based on the α_r parameters for both the Stance and LIWC models using Spearman rank-order correlation coefficient. While we would expect similar patterns for both the Stance and LIWC models, we do not observe any significant correlation between

the ranks of the α_r parameters ($\rho = 0.12$, p -value = 0.24).

To further investigate the level of style-shifting (i.e., rank of α_r parameters) and how it relates to the amount of norm enforcement in the subreddit, we look at the ranked correlation between the α_r parameters and a set of measures to characterize the amount of norm enforcement. Results are shown in Table 7.3 and we find no significant correlations between the measures of subreddit moderation (or norm enforcement) and style-shifting parameters α_r .

Table 7.3: Spearman ranked correlation between subreddit style shifting parameter α_r and measures of subreddit moderation.

Measure	Spearman ranked correlation			
	Stance		LIWC	
	r	p -value	r	p -value
Moderators per active posters	0.1661	0.1265	0.1406	0.1965
Moderators per subscribers	0.1886	0.0718	0.0460	0.6630
Moderators per comment	0.2008	0.0550	0.0289	0.7845
Proportion of deleted comments	-0.0197	0.8489	0.0168	0.8711
Deleted comments per moderator	-0.2020	0.0534	0.0037	0.9717

7.4.3 RQ3: Style-shifting and Style Persistence

When comparing the performance of M_3 vs. M_2 , we did not find significant improvement over modeling the style persistence between consecutive posts in multiple subreddits. This finding does not provide evidence in support of our hypothesis H_3 , which we formulated as: linguistic style of a user persists over time when posting to multiple subreddits.

7.5 Discussion and Future Work

In this work we study the patterns of linguistic style-shifting in the online multi-community environment of Reddit. Using data from the top 100 subreddits by subscribers, we build four models of linguistic style-shifting to study our research

questions. Comparing the performance of models $M_0 - M_2$ we find that the users do change their writing style to adapt to the writing style of the subreddit they post to. We hypothesized that the amount of style-shifting in subreddits could be related to amount of moderation (or norm enforcement) of the subreddit, however, we did not find enough evidence to support this hypothesis. Focusing on the cognitive aspects of linguistic style-shifting, we hypothesized that when users post to multiple subreddits in a short interval of time, there would be some amount of style persistence between consecutive posts in different subreddits. We did not find evidence for this hypothesis from the model outcomes. To summarize, we find that users of Reddit shift style depending on the subreddit they post to; however we did not find evidence of style persistence over time when users post to multiple subreddits.

Future Work. For our analyses presented here, we selected top 100 subreddits by subscribers. Because Reddit users are subscribed to some of these subreddits by default, and most of these subreddits are very popular, the interactional styles of these subreddits might not be unique or could be less characteristic of community norms. One possible alternative to this data sampling method is to focus on subreddits that are less popular (e.g., subreddits in the range of top 100 - 200 in terms of subscribers) and building a member overlap graph to select subreddits which are not central nodes in the member overlap graph. Repeating our model analyses using such alternative data sampling could either validate the current results or provide more insights into the differences regarding the unique vs. less characteristic linguistic styles of subreddits. Other data sampling or selection methods could also be evaluated.

Another direction for future work is to look at how style-shifting is related to the feedback users receive from the subreddits in terms of replies and upvotes. For example, whether adapting to the writing style of the subreddit leads to better feedback from the community. Future work could study this as a causal question.

CHAPTER 8

EFFECTS OF NORM ENFORCEMENT ON ONLINE WRITING STYLE¹

The focus of the previous two chapters are broadly about online community norms and community related stylistic variation. In this chapter I continue the broad focus on community norms and look at one of the key aspects in which online writing differs from other forms of writing: *moderation* or *norm enforcement*. Online content contributed by individuals in different platforms can be moderated for several reasons, including inappropriateness [117], abusive language [235], and violations of community policies [119]. Moderation can vary from minor revisions and removal of content to suspension and banning of contributors [140]. Prior work has investigated the effects of norm enforcement related to the design of online communities (§ 2.4.2); however, the effect of norm enforcement on individual user behavior has not been explored much in large scale settings. What happens when the content contributed to an online community is moderated or corrected for not adhering to community writing style norms? Does moderation help to converge community content to desired stylistic norms? Does moderation change the writing style of individual members to make them better converge to community norms? Or does it make them contribute less and leave the community?

In this chapter, I present a study in which we study these questions in the context of Wikipedia’s writing style norms of neutral point of view (NPOV).² First, I will provide brief background about Wikipedia, motivation to participate, editor roles, and norm

¹Content of this chapter is based on Umashanthi Pavalanathan, Han Xiaochuang, and Jacob Eisenstein. “Mind Your POV: Convergence of Articles and Editors Towards Wikipedia’s Neutrality Norm.”, Proceedings of ACM Human Computer Interaction. 2, CSCW 2018.

²https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

enforcement. Next I formulate two research questions in which we focus on the effects of NPOV norm enforcement both at Wikipedia article-level and at individual editor-level. Then I provide a list of terms related to Wikipedia and our study. Following that, I explain the dataset including the details of how we identified NPOV tagging and treatment editors, and how we constructed the datasets for both the article-level and editor-level analysis. Then I explain our method for identifying biased language, the causal inference method for measuring the effects of treatment on biased language usage, and the analyses to measure the treatment effects. After presenting the results, I discuss the results along with the implications of the work and conclude the chapter with a discussion on challenges and limitations. The study presented in this chapter focuses on the dimensions of *community norms* and *intra-person variation*.

8.1 Background

Wikipedia is the largest online collaborative content creation community. The English Wikipedia contains nearly 5.6 million articles and nearly 140 thousand actively contributing accounts [236]. Similar to the norms or rules of other online communities [38], Wikipedia also has set of norms regarding content and member participation [126]. The three core content policies of Wikipedia are: “neutral point of view (NPOV)”, “verifiability”, and “no original research” [127]. Wikipedia provides strict guidelines about NPOV, and non-complying articles are marked with an NPOV tag by editors. Such NPOV-tagged articles are listed under the “NPOV disputed” Wikipedia category. While the NPOV tagging system has been in place for over a decade, its impact on Wikipedia has not been quantitatively measured. We study the effectiveness of NPOV tagging as a language norm enforcement strategy in Wikipedia, measuring its impact on articles as well as individual editors. Specifically, we ask whether tagging helps articles and editors to converge towards the community’s prescribed writing style, as a step toward the larger goal of understanding the effects of norm enforcement on

online writing.

8.1.1 Norm Enforcement in Wikipedia

Wikipedia provides detailed guidelines of the preferred writing styles through a manual of style [118]. The manual of style contains guidelines not only for formatting such as capitalization, punctuations, and abbreviations, but also for grammar and vocabulary. The *Neutral Point of View (NPOV)* guideline states that “*All encyclopedic content on Wikipedia must be written from a neutral point of view (NPOV), which means representing fairly, proportionately, and, as far as possible, without editorial bias, all of the significant views that have been published by reliable sources on a topic.*” The guideline further states that “*This policy is non-negotiable, and the principles upon which it is based cannot be superseded by other policies or guidelines, nor by editor consensus.*” These editorial policies are enforced by the community of editors to maintain high quality of the articles.

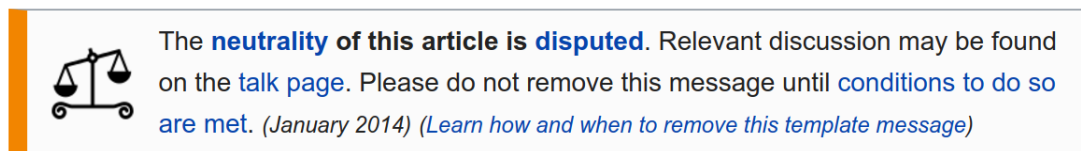


Figure 8.1: An NPOV tag displayed at the top of an article.

Wikipedia articles which do not follow the NPOV guidelines are marked by editors with tags such as `{{POV}}`, `{{NPOV}}`, and `{{POV-section}}` so that they can be included in the editing workflow (Figure 8.1). From the explicit tagging of the non-NPOV content in Wikipedia articles and the availability of the content before and after NPOV revisions, we can develop a better understanding of the acceptable writing style in Wikipedia. The example below from our dataset shows a portion of a sentence before and after it was corrected for NPOV. Usage of **dictated** in (a) suggests a point of view and it is replaced with **overseen** in (b) to make the language more neutral.

- (a) “BL Republic is **dictated** by UberConsul Lars and High Marshall Bill ...”
- (b) “BL Republic is **overseen** by UberConsul Lars and High Marshall Bill ...”

8.1.2 Motivation for Participation and Editor Roles in Wikipedia

The scale and success of Wikipedia as a volunteer-run collaborative content producing community have attracted a large body of research about various aspects of Wikipedia including motivations to contribute [237, 238, 239], formation of community [240, 241, 242], content quality [243, 244], editorial and authorship attributes [245], sharing and collaborative knowledge building [243], learning [237], coordinations of complex tasks [246], and editor functional roles [247, 248]. Prior studies on the motivations for voluntary contribution to Wikipedia provide several explanations including the ideology for contributing to Wikipedia as a variant of open-source application [238], internal self-concept motivation [249], cognitive (e.g., learning new things or intellectual challenge) and affective reasons (e.g., pleasure) [250], and the sense of individual efficacy [124].

Another line of work focused on different editor roles that emerge in Wikipedia [124, 251, 252, 253, 247]. Bryant et al. [124] find that while novice editors contribute edits to articles related to their domain of expertise, expert editors contribute towards improving the quality of Wikipedia itself. Arazy et al. [247] identified functional roles in Wikipedia such as technical administrator, border patrol, quality assurance technicians, administrators, and directors, and find that editor roles determine activity patterns across variety of editing tasks. Yang et al. [254] automatically identify various editor roles based on low-level edit types and find that different editor roles contribute differently in terms of edit categories and articles in different quality stages require different types of editors. In this work, we focus on editor roles that make contributions to improve the quality of Wikipedia articles, particularly articles in the NPOV dispute category.

8.1.3 Research Questions

We study the effects of NPOV tagging and revision as norm enforcement strategies to make Wikipedia an objective source of knowledge at the article level and at the editor level.

RQ1: What are the article-level effects of NPOV tagging?

RQ2: What are the editor-level effects of NPOV correction?

RQ2a: What effect does NPOV correction have on the writing style of the editors?

RQ2b: What effect does NPOV correction have on the engagement of editors?

We address these questions using a corpus of nearly 7,500 articles in the *NPOV dispute* category, and we identify revisions in which a NPOV tag was added or removed, or in which content was corrected for NPOV. We then identify the first revision in which an article received a NPOV tag as the treatment point for the article. Similarly we identify editors whose content contributions were corrected for NPOV as treatment editors and identify the treatment revisions. As a measure of language, we use a set of lexicons associated with non-objective or biased language.³ Considering editor contribution and article revisions before and after the treatment points, we analyze the effect of treatment using an interrupted time series analysis [165].

8.1.4 Terminology

Below, I provide a brief summary of different terminology used in this study:

- **Edit** refers to any changes made to a Wikipedia article.

³Lexical analysis is chosen because of the existence of well-validated lexicons, as described below. We leave for future work the analysis of bias in dimensions such as syntax, semantics, and pragmatics.

- **Revision** is a version of a Wikipedia article resulted from editing the previous revision.
- **Article** contains chronologically ordered revisions.
- **Reverts** in Wikipedia refers to edits that restored the current revision to a previous revision.
- **NPOV tagging** refers to the addition of any NPOV templates to the articles.
- **NPOV correction** refers to a set of edits that are made to correct for NPOV errors. These are detected from the comment metadata in a revision.
- **Treatment** refers to an intervention made to the subject of interest (i.e., article, editor).
- **Pre-treatment** refers to the period prior to the treatment.
- **Post-treatment** refers to the period after the treatment.
- **Treatment to the articles** refers to the addition of an NPOV tag.
- **Treatment to the editors** refers to editors' contribution being corrected for NPOV.
- **Biased language** refers to words/phrases which are not neutral and may introduce a point of view or attitude.

8.2 Datasets

In this section I describe the datasets and the extraction of treatment articles and editors. Figure 8.2 summarizes the data pipeline and methods used in this study.

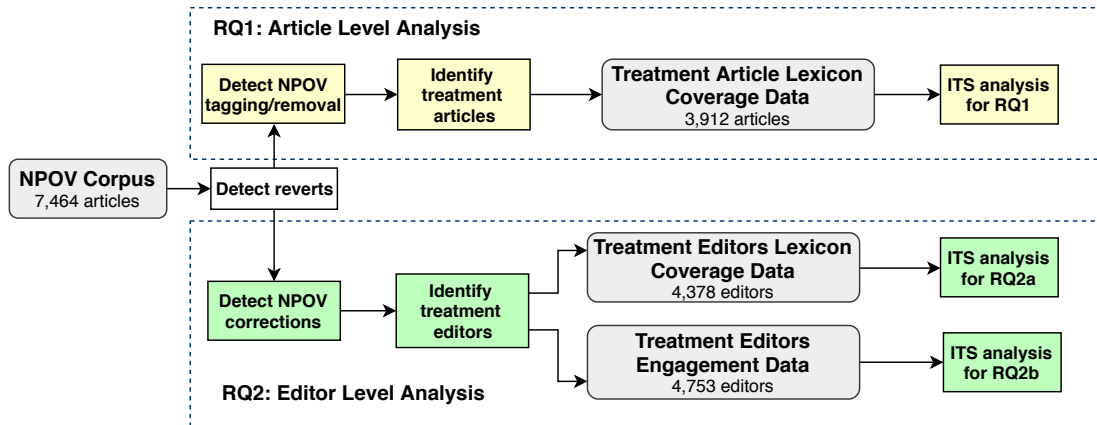


Figure 8.2: Flowchart showing the data and method set up for each research question.

8.2.1 Wikipedia NPOV Corpus

We begin with the NPOV corpus from Recasens et al. [137], which they used to build a classifier to detect bias inducing terms in phrases. The NPOV corpus contains 7,464 articles from the English Wikipedia with a total of 2.7 million revisions. It was constructed by retrieving all the articles that were in the *NPOV dispute* category⁴ in early 2013, together with their full revision history. Wikipedia editors are encouraged to identify and revise biased content to achieve neutral tone, and several NPOV tags are used to mark biased content.⁵ Articles that are marked with any of the NPOV tags will be listed in Wikipedia’s category of *NPOV disputes*. Each version of the article is considered as a *revision*, and an article is a set of chronologically ordered revisions. Each revision of the article contains the version of the article content along with metadata such revision ID, timestamp, contributor, revision comment, and an SHA-1 hash key.

⁴https://en.wikipedia.org/wiki/Category:All_NPOV_disputes

⁵e.g., `{{POV}}`, `{{POV-check}}`, `{{POV-section}}`, etc. When such tag is added, a template such as the following is displayed in the article page: “The neutrality of this article is disputed. Relevant discussion may be found on the talk page. Please do not remove this message until the dispute is resolved.”

8.2.2 Identifying NPOV Tagging and Removal

Several NPOV tags are used to mark biased content and there is no standardized template or metadata available to directly detect tag addition and removal. Therefore, to identify in which revision an article was tagged as violating NPOV and when that tag was removed, we used a set of NPOV-tag patterns based on a preliminary inspection of a small subset of NPOV tagged articles.⁶ In order to detect the addition or removal of an NPOV tag in the absence of any readily available metadata, we need to compare each consecutive revisions and then inspect the difference in the textual content for an addition or removal of an NPOV tag. To automatically extract the changes between two consecutive revisions in terms of added and removed content, we used the Diff Match and Patch library,⁷ which was also used by Recasens et al. [137]. If any of the NPOV tags was present in the added content, then we mark that revisions as an NPOV tag addition point for that article. Similarly, if any of the NPOV tags was present in the removed content, we mark that revision as an NPOV tag removal point for that article.

Accurate identification of valid tag additions and removals is challenging due to the presence of vandalism and reverts.⁸ From our initial inspections of a small set of NPOV tagged articles, we identified several cases of vandalism (e.g., NPOV tags were removed intentionally or accidentally, and then added back within the next few revisions; revisions containing NPOV tags were reverted to a prior revision without a tag). To minimize the effect of vandalism in the accurate identification of NPOV tag addition and removal, we used two heuristics. First, we used the SHA-1 hash key present in the revision metadata to track reverts and disregarded reverted revisions.

⁶`{.*(POV|Pov|PoV|Point Of View|pov|NPOV|Npov|npov|NEUTRALITY|Neutrality|neutrality|NEUTRAL|Neutral|neutral).*`}

⁷<http://code.google.com/p/google-diff-match-patch>

⁸Vandalism and edit wars are primary challenges for Wikipedia editors. There is a large body of work regarding this. For example, see Geiger and Ribes [122], Shachaf and Hara [255], Potthast et al. [256].

Second, if an NPOV tag was added immediately after a prior tag removal, we consider the last tag removal as unreliable. If such changes appear within five revisions after a tag removal, we merge the consecutive tag addition-removal pairs and consider only the latest revision as a valid tag removal revision. In this way, we identified 6,512 articles that had at least one NPOV tag addition and 1,095 articles with multiple NPOV tag additions.

8.2.3 Identifying Treatment Editors

For our editor-level analysis, we define treatment as editing part(s) of an article’s content to correct for NPOV. Treatment editors are editors who originally contributed the portion of the text that was corrected for NPOV. To identify whether a revision was corrected for NPOV, we first check if the comment metadata for that revision contains “NPOV” or “POV” or any case variations. If one of these strings appear in the comment and if that revision was not identified as an NPOV tag addition or removal revision, then we consider it to be a *NPOV correcting* revision. While Recasens et al. [137] checked for only the presence of NPOV strings in the comment to detect NPOV corrections, in our initial inspection we found that sometimes comments of NPOV tagging and removal revisions also contain these NPOV strings. Therefore, we additionally checked if the revision had a tag addition or removal as we do not consider those as NPOV correcting revisions.⁹

Once we have identified NPOV correcting revisions, we trace back previous revisions to identify the treatment editors—editors who originally contributed the portion of text that was corrected for NPOV. As a preprocessing step we removed revisions which are reverts of previous revisions in order to identify the editor who originally contributed a span of text. Otherwise, the editor who made the revert could have

⁹Note that an NPOV removal revision could include corrections for NPOV and removal of NPOV tag. However, in our initial inspections, we found that NPOV issues are often solved in revisions prior to the NPOV tag removal revision. In this way, few NPOV correcting instances could have been missed, but we aimed for precision rather than recall.

been attributed to a portion of text. To trace the treatment editors, we first built a detailed record of which part of the article was contributed by which editor at the character granularity level from the beginning of the article revision history. We then checked all of the NPOV correcting revisions, extracted their first *diff* compared to the previous revision, and looked into the record to identify which editor originally contributed to that *diff-ed* language.¹⁰

Automated scripts or bots are used in Wikipedia to perform repetitive and mundane tasks such as minor fix for syntax and adding dates to NPOV tags. Because our interest is in studying the writing style of human editors, we removed bots by checking for the presence of “bot” in the username and presence for the “bot” editor group attributes.¹¹ In total, we identified 5,645 treatment editors after removing 20 bot accounts. Note that each treatment editor could be NPOV corrected more than once. On average, a treatment editor is corrected for NPOV 1.9 times, and 32.4% of the treatment editors are corrected for NPOV more than once.

8.2.4 Dataset for Article-Level Analysis

In RQ1 we study the effect of NPOV tagging on the linguistic style of the articles. For this analysis we use the revisions of articles in the NPOV corpus (§ 8.2.1) and characterize the linguistic style of each of the revisions. To observe the biased language trend over time, we limited this dataset to include NPOV articles which have at least 40 revisions and considered up to 40 revisions before and after the treatment for the regression analysis¹² after the treatment (i.e., addition of an NPOV tag).

¹⁰Note that recent work such as Flöck and Acosta [257] proposed other sophisticated algorithms to attribute authorship of revisioned content. Future work could consider employing such algorithms.

¹¹Editor attributes are obtained from the Media Wiki API <https://www.mediawiki.org/wiki/API:Query>. Examples of editor group attributes include “reviewer”, “administrator”, and “bot”

¹²We also considered the number of revision threshold window W of $[10 \leq W < 20]$ and $[20 \leq W < 40]$ and the results are similar qualitatively. The number of articles included in the analysis for each of these thresholds are 1054, 1095, and 1763 for the revision windows $[10 \leq W < 20]$, $[20 \leq W < 40]$, and $[W > 40]$ respectively.

8.2.5 Dataset for Editor-Level Analysis

To study the editor-level effects of NPOV correction, we collected each treatment editor’s textual contribution to the Wikipedia article namespace¹³ by querying the Media Wiki API.¹⁴ In RQ2a we study the effect of NPOV correction on the writing style of the treatment editors. To characterize treatment user writing style before and after treatment, we first obtained up to 150 revisions to which the editors contribute (in all Wikipedia articles, not restricting to the initial NPOV corpus) before and after their first treatment point. To control for any platform level changes such as revisions to NPOV policies, we restricted the pre and post treatment revisions to be within an year from the treatment. When then queried for the parent revision of each of the editors’ revisions and extracted the text added by the treatment editors using the *diff* library. Since our interest is in the language use, we considered only the revisions with any content addition.¹⁵ In RQ2b we study the changes in treatment editor engagement before and after NPOV correction. To characterize treatment editor engagement we obtained all the revisions (in all Wikipedia articles, not restricting to the initial NPOV corpus) they have contributed to during two months before and after the treatment point.

8.3 Methods

In this section I describe the methods we used for preprocessing of Wikipedia data for our analyses, identifying biased language, and measuring effect of treatment.

¹³Wikipedia includes multiple namespaces such as article page, article talk page, user page, user talk page, etc. Since our interest in to characterize editor writing style in Wikipedia articles, we restricted our dataset to the article namespace. Prior work, such as Elia [258], has found significant differences between the linguistic style of article text and talk page text due to the conversational nature of talk pages.

¹⁴<https://www.mediawiki.org/wiki/API:Query>

¹⁵Examples of other revision contributions include addition of links, templates, and citations.

8.3.1 Identifying Biased Language

Bias in Wikipedia can emerge from several factors including language style, editors' point of view, cited sources, coverage of topics, etc. Prior work focused on political bias [259], cultural bias [260], gender bias [261], topic bias [262], and authoritative bias [263]. Our goal in RQ1 and RQ2a is to study the effects of NPOV tagging in article-level and editor-level language. Particularly, we are interested in characterizing the bias in articles and editor contribution in terms of the linguistic style that may introduce bias. The Wikipedia Manual of Style [118] states that “*There are no forbidden words or expressions on Wikipedia, but certain expressions should be used with caution, because they may introduce bias. Strive to eliminate expressions that are flattering, disparaging, vague, or endorsing of a particular view point.*” Prior work [e.g., 264, 137] has addressed the task of identifying biased language in Wikipedia at different levels. However, we are not aware of a reliable system that can detect bias in terms of linguistic style in Wikipedia articles. Therefore, we use a set of linguistic style lexicons as a proxy to characterize biased language in Wikipedia.¹⁶

The Wikipedia Manual of Style provides a list of *words to watch* in a prescriptive manner [265], indicating style words that may introduce bias. We compiled these style words into a lexicon called *Words to Watch*. Prior work of Recasens et al. [137] introduced the task of detecting bias inducing terms in phrases from Wikipedia articles and used a set of pre-compiled style lexicons that are indicative of expressions of attitude or point of view. These lexicons include hedges, factive verbs, assertive verbs, positive words, and negative words. We use these nine lexicons in addition to the words from Wikipedia manual.¹⁷ A list of all the lexicons we used is shown in Table 8.1 with a description, sources, and example terms. To characterize the amount of biased

¹⁶We limit our analysis only to lexicon, while syntax, semantics, and pragmatic measures could also be studied.

¹⁷Note that Recasens et al. [137] used a bias lexicon created from the NPOV articles. We do not include it for our analysis as this lexicon is not entirely indicative of linguistic style and included content words such as *communist*, *historian*, and *migration*.

language in a text, we compute the coverage of each lexicon word per token in the text.

8.3.2 Measuring the Effect of Treatment on Biased Language Usage

Our goal in RQ1 is to investigate whether NPOV tag addition has an effect on the level of biased language in the NPOV tagged articles. Similarly, our goal in RQ2a is to investigate whether NPOV correction has an effect on the level of biased language usage of treatment users. Both of these are causal questions. To answer them, we apply the Interrupted Time Series technique (ITS) for causal inference [165]. This method has been used in recent studies on hate speech [149] and conspiratorial discussions[166]. Interrupted Time Series analysis is a quasi-experimental design that can be used to evaluate the longitudinal effects of an intervention (i.e., treatment), through segmented regression modeling. The term “quasi-experimental” refers to an absence of randomization. ITS is a tool for analyzing observational data where complete randomization, or case-control design, is not affordable or possible. In its basic form, an ITS is modeled using a regression model (e.g., linear, logistic or Poisson) that includes only three time-based covariates as shown in Equation 8.1. The regression coefficients of these covariates estimate the pre-treatment trend, the level change at the treatment point, and the trend change from pre-treatment to post-treatment.

$$\mathbf{y}_t = \beta_0 + \beta_1 \mathbf{t} + \beta_2 \mathbf{x}_t + \beta_3 \mathbf{t} \mathbf{x}_t + \epsilon \tag{8.1}$$

Here \mathbf{t} is the time elapsed since the start of the study; \mathbf{x}_t is a dummy variable indicating the pre-treatment period (coded 0) or the post-treatment period (coded 1); \mathbf{y}_t is the outcome at time t . The pre-treatment slope, β_1 , quantifies the trend for the outcome before the treatment; the level change, β_2 , estimates the change in level that can be attributed to the treatment; the change in slope, β_3 , quantifies the difference

Table 8.1: List of Lexicons Used to Characterize Bias Language.

Lexicon Name	Description	Example Terms
Words to Watch	Style words that may introduce bias (Wikipedia [265])	<i>fortunately, notable, often, speculate</i>
Hedges	Used to reduce one’s commitment to the truth of a proposition (Hyland [266])	<i>apparent, seems, unclear, would</i>
Assertives	Complement clauses that assert a proposition (Hooper [267])	<i>allege, hypothesize, verify, claim</i>
Positive Words	Positive sentiment terms (Liu et al. [268])	<i>achieve, inspire, joyful, super</i>
Negative Words	Negative sentiment terms (Liu et al. [268])	<i>criticize, foolish, hectic, weak</i>
Factives	Terms that presuppose the truth of their complement clause (Kiparsky and Kiparsky [269])	<i>regret, amuse, strange, odd</i>
Implicatives	Imply the truth or untruth of their complement, depending on the polarity of the main predicate (Karttunen [270])	<i>avoid, hesitate, refrain, attempt</i>
Report Verbs	Used to indicate that discourse is being quoted or paraphrased (Recasens et al. [137])	<i>praise, claim, dispute, feel</i>
Strong Subjectives	Add strong subjective force to the meaning of a phrase (Riloff and Wiebe [108])	<i>celebrate, dishonor, overkill, worsen</i>
Weak Subjectives	Add weak subjective force to the meaning of a phrase (Riloff and Wiebe [108])	<i>widely, unstable, although, innocently</i>

between the pre-treatment and post-treatment slopes. This model of ITS assumes that without the treatment, the pre-treatment trend would continue unchanged into the post-treatment period, and there are no external factors systematically affecting the trends. One important step before modeling an ITS analysis is to hypothesize how the treatment would impact the outcome if it were effective, particularly whether the change will be in the slope of the trend, a change in the level, or both.

Why no control? Our goal in this study is to measure the longitudinal effects of NPOV norm enforcement on writing style in Wikipedia. We use interrupted time series analysis, which is a widely used approach to model such longitudinal effects through a within-subject analysis. Another causal inference approach used in many other observational studies compare matched pairs of treatment and control subjects [e.g., 159, 160]. The impact of the treatment can be assessed through comparisons of these groups [164]. This technique requires that the treatment and control groups are identical in all attributes that could possibly affect the outcome. In the case of Wikipedia articles and editors, it is difficult to even enumerate these attributes, let alone measure and match them. Furthermore, text documents (articles and edits) are inherently high-dimensional objects, which poses statistical challenges for matching [271]. We therefore focus on within-subject analysis through the interrupted time series paradigm.

8.4 RQ1: Article-Level Effects of NPOV Tagging

Quantifying Article Writing Style. We use the dataset we created for the article level analysis (§ 8.2.4) to study RQ1. Before measuring the writing style of each article revision, we performed a set of preprocessing steps to extract textual contribution of the article revision. We removed Wikipedia markup templates, URLs, and other non-textual content. We then tokenized the text and extracted tokens with two or

more characters. To quantify article writing style, we extracted the coverage of tokens from each of the ten bias-related lexicons (§ 8.3.1) and computed lexicon coverage rate as the number of lexicon tokens per total tokens in the text of each article revision. We considered the latest 40 pre-treatment revisions and the earliest 40 post-treatment revisions of the NPOV articles for the ITS analysis.

ITS Modeling for Article Writing Style. When an article has an active NPOV tag, it stays on top of the article for all revisions until its removal. As this can be considered a continuous treatment, we would expect the editors to be cautious for biased language and the articles to have a continuous drop in biased language coverage. Therefore, to measure the effect of an article having an NPOV tag on the biased language coverage, we hypothesize a slope change in the article-level biased language usage after treatment. To model this, we use the ITS regression model in Equation 8.1 and use a linear model to fit in the R software.

Results: RQ1. Results of the ITS analysis for RQ1 is shown in Table 8.2 and the trends are shown in Figure 8.3. Through the ITS analysis, we observe that when an active NPOV tag is on the article, there is a statistically significant change in the slope of the trend of nine out of ten lexicons used to characterize biased language.

Robustness Check: The dataset used in this analysis was collected from the snapshot of articles in Wikipedia’s *NPOV dispute* category in early 2013. Some of the articles contain include NPOV issues that are inherently difficult to resolve and could be tagged for a long period of time. As this could be a potential confound in the editing behavior of such articles, as a robustness check, we re-ran the analysis on a subset of the data after removing articles that are NPOV tagged for more than three years. The distribution of time-intervals between the addition of the NPOV tag and the data collection time, and the result of the regression analysis for the subset of

Table 8.2: Results of RQ1 Analysis. Article lexicon coverage is computed for the textual content of the article results from each revision. Coefficient β_3 indicates the change in slope after treatment. Statistical significance after correcting for multiple comparisons using Benjamini and Hochberg [2] adjustment for false discovery rate are shown; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Percentage change is computed as the change in post-treatment lexicon coverage after 40 revisions.

Lexicon	β_3	p -value		% change
Words to Watch	-1.47e-05	1.08e-13	***	-5.24
Hedges	-4.83e-06	1.30e-06	***	-3.54
Assertives	-3.12e-06	4.40e-09	***	-5.29
Positive Words	-4.47e-05	< 2e-16	***	-7.63
Negative Words	-2.35e-05	< 2e-16	***	-4.52
Factives	8.91e-07	2.84e-01		+0.49
Implicatives	-3.41e-06	3.50e-08	***	-4.80
Report Verbs	-3.42e-06	2.99e-04	**	-1.79
Strong Subjectives	-5.86e-05	< 2e-16	***	-8.56
Weak Subjectives	-4.53e-05	< 2e-16	***	-3.11

articles after removing potential outliers are shown Appendix C. The treatment effect remains significant after excluding these long-running disputes.

8.5 RQ2: Editor-Level Effects of NPOV Correction

Next, we explore the editor-level effects of NPOV correction through the following research questions:

RQ2: What are the editor-level effects of NPOV correction?

RQ2a: What effect does NPOV correction have on the writing style of the treatment editors?

RQ2b: What effect does NPOV correction have on the engagement of treatment editors?

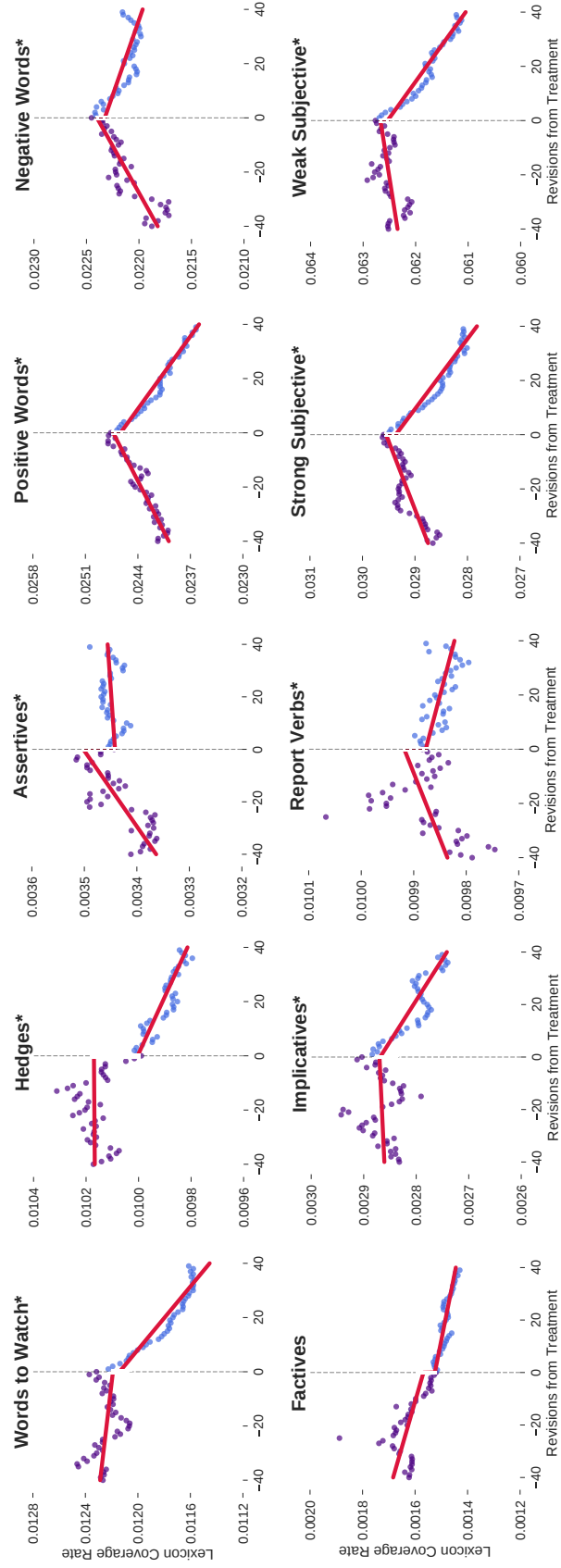


Figure 8.3: Article level lexicon coverage with respect to the treatment of NPOV tagging. We computed the average lexicon coverage rate of treatment articles' 40 revisions before and after the treatment revision. Lexicons for which we observe a statistically significant drop in slope ($p < 0.05$) are marked with an asterisk (*).

8.5.1 Editor Writing Style: RQ2a

Quantifying Editor Writing Style. We use the dataset we created for the editor level analysis (§ 8.2.5) to study RQ2a. Before measuring editor writing style, we performed a set of preprocessing steps to extract textual contributions of the treatment editors. We removed Wikipedia markup templates, URLs, and other non-textual content. We then tokenized the text and extracted tokens with two or more characters. To quantify editor writing style, we extracted the coverage of tokens from each of the ten bias-related lexicons (§ 8.3.1) and computed lexicon coverage rate as the number of lexicon tokens per total tokens in the text contributed by an editor in a revision. We considered the latest 40 pre-treatment edits and the earliest 40 post-treatment edits of the treatment editors for the ITS analysis.

Controlling for Editor Experience and Talk Page Discussion. Other factors may confound measurement of the effect of NPOV correction on editor-level writing. One such factor is the experience of the editor: previous work shows that inexperienced editors respond differently to norm-enforcement activities such as edit reverts (e.g., [152]). To control for editor experience, we include it as a binary predictor in the regression model. Based on the criteria used in prior work [252], we consider editors who have contributed at least 250 edits prior to the treatment as *experienced editors* and others as *inexperienced editors*. In our dataset, 2,847 out of 4,378 editors had at least 250 edits prior to the treatment.

In addition to correcting content for NPOV, sometimes editors also post to the talkpage of the articles or the talkpage of the editor whose contribution was revised to discuss about the correction. These posts trigger an email prompt to the original editor. Thus, the talk page discussion acts as an additional “correction” method that could potentially amplify the treatment of correcting content with NPOV tags in the comment metadata. The discussion between the correcting and corrected editors

during the treatment period could influence the future behavior of the corrected editor in terms of language usage and engagement. Therefore, we also include the presence of talk page discussion as an additional binary predictor, which indicates whether any of the following criteria is satisfied during a week prior or after the treatment: (1) correcting editor posts on corrected editor’s talk page; or (2) corrected editor posts on correcting editor’s talk page; or (3) both corrected editor and correcting editor post on the article talk page. In our dataset, 1,216 of the 4,378 NPOV corrections were accompanied by a talk page discussion.

ITS Modeling for Editor Writing Style. Unlike an active NPOV tag on the article, which stays as a continuous treatment for all later revisions until its removal, the editor who gets an NPOV correction only receives treatment once.¹⁸ Therefore, we hypothesize a change in the *level* of the trend of average editor writing style after treatment. We use the ITS regression model in Equation 8.1, excluding the term $\beta_3 \mathbf{x}_t$ (since we are only hypothesizing a level change) and using a linear model fit in the R software (*Model-I*). In addition to the terms in the ITS equation, as discussed in the previous subsection, we also used a second model (*Model-II*) including two binary control predictors for editor experience (*experienced*) and talk page discussion during treatment (*discussion*). We also add an interaction term between each of these additional predictors and the post-treatment indicator (i.e., interaction with the dummy term \mathbf{x}_t).

Results: RQ2a. To understand the trend of the change due to treatment, we first visualize the level change in average editor writing style from *Model-I* in Figure 8.4. NPOV correction is associated with a small level decrease in the usage of four out of ten lexicons used to characterize biased language: positive words, negative words,

¹⁸While 32.4% of the treatment editors are corrected more than once, repeated corrections are rare in the short window of 40 edits adjacent to the treatment.

strong subjectives, and weak subjective. No significant treatment effect is observed for the remaining six lexicons. *Model-II* includes the control predictors for editor experience and talk page discussion. Relevant output from *Model-II* is shown in Table 8.3. The trend of level change (β_2) remains the same after including the control predictors. The coefficients for the *experienced* predictor are negative and significant for eight of the ten lexicons, which suggests that experienced editors use a lower amount of non-neutral language.¹⁹ The coefficients for the effect of discussion after treatment ($\mathbf{d} : \mathbf{x}_t$) are not significant except for the negative words lexicon. None of the coefficients for the effect of editor experience after treatment ($\mathbf{e} : \mathbf{x}_t$) are significant, suggesting that editor experience does not influence the effect of NPOV correction on writing style.

8.5.2 Editor Engagement: RQ2b

Quantifying Editor Engagement. To study RQ2b, we define engagement of a Wikipedia editor as any non-minor revisions they made to the article pages. We use the editor-level engagement dataset we created (§ 8.2.5) for this purpose and considered all treatment editors who have contributed at least ten revisions during the two months prior to the treatment.

ITS Modeling for Editor Engagement. Similar to the hypothesis for editor level lexicon coverage, we hypothesize a change in the level of editor engagement after treatment. We model this using the ITS regression model in Equation 8.1 excluding the slope term $\beta_3 \mathbf{t} \mathbf{x}_t$, but with editor fixed effects. Because the dependent variable in this model is a count variable (i.e., number of revisions per day), we would use a Poisson regression model. However, summary statistics of the dependent variable show an excessive amounts of zeros ($> 40\%$) and over-dispersion (variance \gg mean).

¹⁹This finding also validates our choice of lexicons to characterize non-neutral language as we would expect experienced editors to better adhere to community writing style norms compared to newbies.

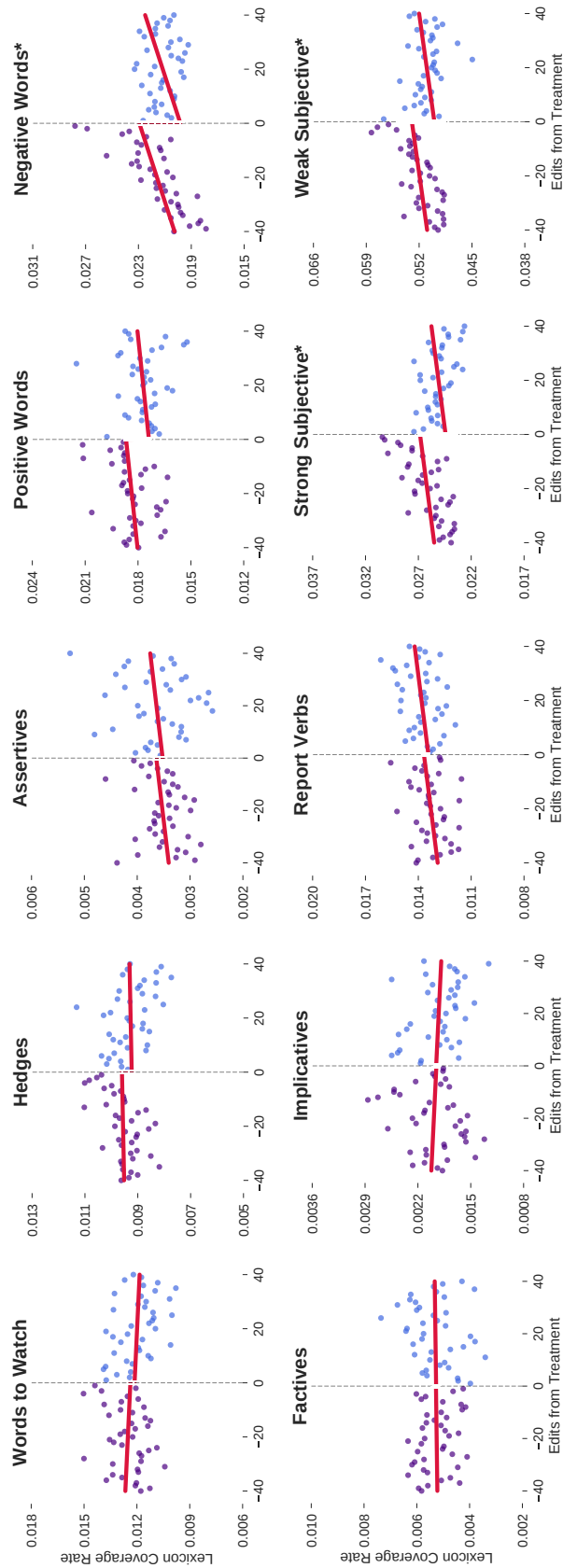


Figure 8.4: Editor level lexicon coverage with respect to the treatment of NPOV correction. We computed the average lexicon coverage rate of treatment users' textual contributions 40 revisions before and after the treatment revision. Lexicons for which we observe a statistically significant level drop ($p < 0.05$) are marked with an asterisk (*).

Table 8.3: Results of RQ2 Analysis after controlling for editor experience and talk page discussion during treatment. Editor lexicon coverage is computed for their textual contribution in each revision. Coefficient β_2 indicates the change in level after treatment; **d** indicates whether there was any talk page discussion during treatment (reference: *discussed* = 0); **e** indicates whether the editor is considered experienced or not (reference: *experienced* = 0); \mathbf{x}_t is the indicator variable for post-treatment; $\mathbf{d} : \mathbf{x}_t$ and $\mathbf{e} : \mathbf{x}_t$ are the interaction terms. Statistical significance after correcting for multiple comparisons using Benjamini and Hochberg [2] adjustment for false discovery rate are shown; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Percentage change is computed as the level change following the treatment. Note that none of the coefficients for the $\mathbf{e} : \mathbf{x}_t$ interaction term are significant at $p < 0.05$.

Lexicon	β_2	% change	$\mathbf{d} : \mathbf{x}_t$	\mathbf{e}	$\mathbf{e} : \mathbf{x}_t$
Words to Watch	-4.98E-04	-1.92	-5.90E-04	-2.98E-03	*** 6.72E-04
Hedges	-8.76E-04	-3.85	-7.78E-05	-1.38E-03	*** 6.59E-04
Assertives	1.58E-04	-3.32	-3.05E-04	-5.63E-04	* -2.46E-04
Positive Words	-2.27E-03	** -6.96	-1.70E-04	-3.47E-03	*** 1.48E-03
Negative Words	-2.55E-03	** -13.54	-1.92E-03	* -3.66E-03	*** 5.41E-05
Factives	3.51E-04	-0.10	-5.44E-04	1.48E-03	*** -3.78E-04
Implicatives	-2.78E-06	0.16	-1.72E-04	-6.44E-04	*** 2.22E-04
Report Verbs	3.27E-04	-1.86	-7.11E-04	6.02E-04	-6.03E-04
Strong Subjectives	-2.48E-03	** -9.02	1.59E-04	-5.99E-03	*** -7.43E-06
Weak Subjectives	-2.70E-03	* -5.63	-2.18E-04	-7.20E-03	*** -1.88E-04

Table 8.4: Results of RQ2b Analysis: predictor coefficients of the count model of the zero-inflated negative binomial regression. The dependent variable, editor engagement is computed as number of edits during each of the three-day period. We considered 60 days prior to the treatment and 60 days after the treatment. *Post-treatment* is the indicator for whether the observation is post (=1) or pre (=0) treatment. *Timebin* is the three-day window; we consider 20 timebin windows pre- and post-treatment, which includes 120 days in total. *Discussion* is the indicator whether there was talk page discussion between correcting and corrected editors during treatment (*Discussion* = 1) or not. *Experienced* is the indicator for whether the editor is experienced (>250 edits, *Experienced* = 1) or not. Interaction terms include: Post-treatment : Discussion(=1) and Post-treatment : Experienced(=1). Statistical significance after correcting for multiple comparisons using Benjamini and Hochberg [2] adjustment for false discovery rate are shown; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

Predictor	Coefficient	<i>p</i> -value	
Post-treatment	0.4675	<2e-16	***
Timebin	0.0080	<2e-16	***
Discussion(=1)	0.2018	<2e-16	***
Experienced(=1)	1.8937	<2e-16	***
Post-treatment : Discussion(=1)	0.0491	0.0404	*
Post-treatment : Experienced(=1)	-0.6618	<2e-16	***

Therefore we used zero-inflated negative binomial regression model, as implemented in the *pscl* package²⁰ in the R software. Similar to RQ2a, we include binary controls for *experienced* and *discussion*.

Results: RQ2b. Results of the ITS analysis for RQ2b are shown in Table 8.4. The interaction between the level change and control predictors *experienced* and *discussion* are shown in Figure 8.5. These results show a significant increase in engagement for inexperienced editors (editors with less than 250 edits at treatment time). The decrease in engagement for experienced editors is small after including the overall temporal trend. Talk page discussion during treatment is associated with a small increase in engagement for both experienced and inexperienced editors.

²⁰<https://cran.r-project.org/web/packages/pscl/pscl.pdf>

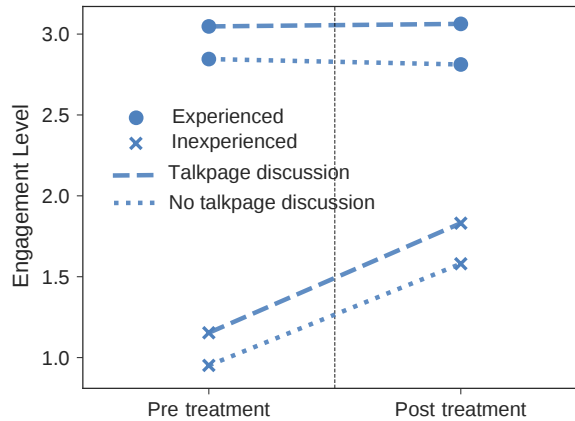


Figure 8.5: Visualization of RQ2b regression coefficients showing the pre- and post-treatment engagement levels for the control factors *experienced* and *discussion*.

8.6 Discussion

Next, we reflect on the article-level and editor-level effects of NPOV tagging, discuss the differences in success for articles and editors, discuss implications of our findings to Wikipedia and other online communities with writing style norms and conclude with challenges and limitations.

8.6.1 RQ1: Article-level Effects of NPOV Tagging

When an article has an active NPOV tag, we observe a statistically significant, gradual reduction in biased language coverage for all lexicons except factives.²¹ The reduction is larger for Wikipedia words to watch (5.2%), strong subjectives (8.6%), positive words (7.6%), and assertives (5.3%), compared to other lexicons. Using these lexicons as a proxy for biased or non-NPOV language, it is evident that non-objective language used in Wikipedia articles decreases as editors make more revisions. This demonstrates the effectiveness of marking Wikipedia articles with NPOV tag in reducing the prevalence of non-neutral language.

The dataset we used for this analysis is collected from a single snapshot of the

²¹Note that compared to other lexicons, the set of factives is small in size — containing only 27 terms — and hence the coverage of this lexicon is small.

articles in the *NPOV dispute* category. Some of the articles which are controversial and hard to resolve could be NPOV tagged for a long period of time could affect the editing behavior. As a robustness check we repeated the analysis on the subset of the data after removing articles which has an active NPOV tag for more than three years. As we report in Appendix C, the main results hold after removing these potential outliers.

8.6.2 RQ2: Editor-level Effects of NPOV Correction

When an editor is corrected for NPOV, there is a significant reduction in their negative words usage (13.5%); however, we do not observe any statistically significant change for six of the ten lexicons relating to biased language, including Wikipedia’s own list of “words to watch” (RQ2a). These trends remain even after controlling for editor experience and talk page discussion during treatment. For RQ2b, we observe an increase in engagement for inexperienced editors after NPOV correction. One possible explanation is that NPOV correction helps inexperienced users become aware that their contributions are monitored by others, and this awareness could motivate them to contribute more. These results for RQ2 suggest that when corrected for NPOV, the quality of the writing style of editors does not improve significantly (at least, as measured by all but one of our lexicons), while it leads to a increase in editing activities for inexperienced editors. These findings partially conflict with the observations of Halfaker et al. [152], who found that revert actions demotivate new editors and reduce the quantity of work, even as they increase the overall quality of contributions. Halfaker et al. [152] also found that reverts affect editors differently based on the experience of the editor who makes the revert. Future work could perform similar analysis to further understand the effects of NPOV correction on engagement based on the experience of correcting editors.

8.6.3 Possible Reasons for the Differences in the Effectiveness of Norm Enforcement

While NPOV tagging helps articles to converge to the desired writing style, we do not find statistically significant evidence that these corrections encourage editors to adopt the desired writing style. One possible reason for the significant improvement in writing style of the NPOV tagged articles is the active contribution of a group of editors who are dedicated to improve the overall quality of Wikipedia articles. According to prior work on the classification of different types of work (§ 8.1.2), editing tasks, and roles in Wikipedia, editors perform distinct roles, corresponding to distinct types of contributions. For example, Bryant et al. [124] found that while novice editors contribute only to articles related to their expertise, expert editors contribute to improve the overall quality of Wikipedia articles. Arazy et al. [247] identified several roles of Wikipedia editors, including quality assurance technicians, who are editors contributing towards patrolling Wikipedia and ensuring content quality. Further, Yang et al. [254] found that articles in different quality stages require different types of editors. Tools such as SuggestBot [272] recommend specific editing tasks (e.g., cleanup and rewrite) to editors based on their previous editing patterns and interests. These prior findings support our hypothesis that the NPOV tagged articles are of interest to specific editor roles, who are dedicated to make revisions to improve NPOV tagged articles.

NPOV tags are highly visible to Wikipedia editors via the banner template displayed in the article, often on the top, and the NPOV tagged articles are listed in a specific Wikipedia category (i.e., *NPOV disputes* category). These visible actions could attract more editors to contribute towards improving the tagged articles. However, the treatment of NPOV correction of editor contribution is not visible to the same extent because editors are not directly notified about their contribution being corrected for NPOV. Editors keep track of the revisions to their previously contributed articles via

the “watchlist” tool [124], but this requires active actions from the editors.

8.6.4 Implications for Wikipedia and Online Language Moderation

Our findings in RQ2a show that the current regime of tag-and-correct does not help editors to significantly improve their adherence to the NPOV norm. This suggests the need for additional interventions targeted at editors. Possible strategies include explicit notification of their textual contribution being corrected for NPOV language, reminders about writing style norms, and incentives when they progress. Findings from RQ2b shows an increase in engagement after treatment for inexperienced editors; talk page discussion during treatment is also associated with a small increase in engagement for all editors. The increase in engagement for inexperienced editors could be due to the signal that the community is aware of their contributions. Interventions such as explicit notification of NPOV correction could include personalized messages [154] that could further increase the engagement of inexperienced editors.

Other online collaborative content creation communities who wish to enforce linguistic style norms may consider implementing interventions similar to Wikipedia’s treatment of NPOV. Further community-level and individual-level interventions could help to converge to expected writing style norms.

In this work, we focused on NPOV, which is one of the three core content policies of Wikipedia. The other policies are “Verifiability” and “No original research”. These three policies jointly determine the quality of content acceptable in Wikipedia articles. Similar to NPOV tagging, these policies are enforced using tags such as `{{citation needed}}`, `{{verification needed}}`, `{{original research}}`, and `{{synthesis}}`. Our findings suggest that the NPOV tagging helps articles to converge to neutral language, but we did not find significant changes at individual editor language. As we discuss in § 8.6.3, various editor roles and the types of work performed by different editor roles could be one of the mechanisms by which NPOV tagging helps to improve article

quality. A similar study could be performed to test the effectiveness of tagging for the other quality control policies of Wikipedia.

8.6.5 Limitations and Future Work

Limitations with characterizing biased language. We use a set of style lexicons including Wikipedia’s “words to watch” to characterize non-neutral linguistic style. While similar approaches are used in prior work, these lexicons may not accurately and entirely capture non-neutral language. Machine learning could be applied to this problem by training on the edits that led to the application of an NPOV-related tag and correction [137]. We limit our analysis of language style to lexical methods; however, other aspects of language such as syntax, semantics, or pragmatics could also be studied (e.g., active vs. passive form).

Potential issues in accurately identifying the treatment groups. The presence of reverts and vandalism imposed challenges in correctly identifying treatment articles and editors. We used a set of heuristics to reduce these issues (§ 8.2.2). However, our approach might not be perfect and could attribute an NPOV correction for an editor who is not the original contributor of the corrected text. This will result in an underestimation of treatment effects. Sophisticated algorithms to detect vandalism and reverts [129, 122], and algorithms to attribute authorship of revised content [257] could be used in future work to improve accurate identification of treatment editors.

Furthermore, the non-standard nature of NPOV tags and variations in the tags limit the recall of detecting tag addition and tag removal revisions. To detect NPOV corrections, following the approach of Recasens et al. [137], we searched for occurrences of “NPOV”, “POV”, or any case variations (§ 8.2.3) in the revision comments. However, other idiomatic terms such as “point of view”, “bias”, and “weasel words” could also be used to indicate NPOV corrections. Failing to detect these variants could result in an

underestimation of the treatment effects.

Controlling for platform-level changes over time. While we limit the time range of our analyses to control for significant platform-level changes, there is a possibility for internal and external factors, such as influx of new editors, site upgrades, or policy changes, to influence the findings. The article dataset used in this work was collected in 2013, and the policies and normative behavior in Wikipedia may have changed since then. Future work could consider replicating this work using datasets collected in different time periods to analyze any longitudinal changes in normative behavior in the platform.

Restricting article level analysis to unresolved articles. Our article dataset was collected from the snapshot of articles in the *NPOV dispute* in 2013. We use this single snapshot data and the articles in this dataset have an active NPOV tag (i.e., an unresolved NPOV issue). As a robustness check, we re-ran the regression for RQ1 using a subset of data after removing articles with long standing disputes and the effect of NPOV tagging treatment remains significant. Using additional article data collection, the article level analysis could be expanded to cases where an NPOV tag was added and removed or where articles went through multiple NPOV tag additions/removals. This could provide insights about the effects after tag removal and repeated offenders, and future work could focus on this direction. From RQ1 we find that when an article has an active NPOV tag, subsequent revisions result in a decrease in bias language. Note that this finding does not imply that a tag is *required* to reduce biased language in articles.

8.7 Conclusions

In this chapter, I have described our study on the effects of norm enforcement on writing style on Wikipedia. Specifically, we investigated the effectiveness of neutral point of view (NPOV) norm enforcement both at the platform level (i.e., articles) and at the individual member level (i.e., editors). Our work is the first quantitative investigation of the effectiveness of the Wikipedia NPOV tagging system which has been in place for more than a decade. In this chapter, I first provided a brief background about Wikipedia and the norm enforcement practices of Wikipedia to motivate our broad research question, which is about the effects of NPOV norm enforcement on Wikipedia writing style. Then I described our dataset, the challenges in identifying NPOV tagging and correction, and our approach to extracting treatment articles and editors. Following that I explained how we characterized non-neutral or biased language style using a set of lexicons from prior work and Wikipedia’s manual of style. Then I explained the causal inference methodology we used in this work, interrupted time series analysis (ITS), and its appropriateness for our study. After providing the details of the specific analyses to study both of our research questions, I presented the results along with a discussion of implications and limitations.

An important aspect in which online writing differs from other forms of writing is the possibility of content being moderated or corrected for not adhering to the community norms. Looking at the NPOV tagging system in Wikipedia as a case study, we investigated the effectiveness of writing style norm enforcement in online writing. Using a causal inference analysis we find that after an article is tagged for NPOV, there is a significant decrease in biased language in the article, as measured by several lexicons. However, for individual editors, NPOV corrections and talk page discussions yield no significant change in the usage of words in most of these lexicons, including Wikipedia’s own list of “words to watch.” This suggests that NPOV tagging

and discussion does improve content, but has less success enculturating editors to the site's linguistic norms. Wikipedia is a collaboratively contributed content creation community. A similar study could be done to compare these findings from Wikipedia to the effectiveness of linguistic style moderation in other online communities such as discussion forums (e.g., Reddit, Stack Overflow).

This chapter concludes the series of case studies through which I investigated how social context influence the stylistic variation in online writing. In the next chapter, I summarize these case studies and provide directions for future work.

CHAPTER 9

CONCLUSION AND FUTURE WORK

The central thesis of this dissertation is that “using data-driven computational algorithms and causal statistical analysis on observational data we can extract insights about how social context influences online writing.” I provide evidence for this claim through the following findings:

- Using a data-driven method to automatically extract a large number of non-standard linguistic variables from a large collection of tweets, and focusing on the audience navigation features of Twitter, I show evidence for audience-modulated stylistic variation in Twitter.
- Through a causal statistical analysis on the adoption of a new form of stylistic innovation in online writing, emojis, I show that the new technological affordances in online social platforms influence the stylistic choices in online writing.
- By quantitatively operationalizing the sociolinguistic construct of stancetaking, I show that unsupervised methods can be used to characterize the interactional meaning in language at scale in order to study intra-person and inter-community stylistic variation in online writing.
- Through a causal statistical analysis to study the effects of norm enforcement in Wikipedia, I show that the norm enforcement actions influence online community’s convergence of linguistic style towards the desired writing style norms.

There are several future directions which could further provide evidence for this thesis and provide more insights into linguistic style variation in online writing:

Longitudinal Patterns of Variation and Language Change. This thesis work has primarily focused on stylistic variation at a specific time point or over a short period of time. I studied the social meaning of stylistic variation considering the interplay between the need to convey varied social meanings in online interpersonal interactions and the affordances in technology-mediated platforms. The affordances of social platforms and the goals of social media users are changing rapidly and hence the stylistic patterns in online writing. Therefore, for a deeper understanding of stylistic variation in online platforms and its social dynamics, we need to investigate the longitudinal patterns of stylistic variation in online writing. Language change is a fundamentally social phenomenon [273]. By combining longitudinal analysis with evolving social network information, future work could track the language change in progress in social media. Although large-scale studies of change over time will involve several challenges, as the amount of longitudinal social media data is keeps increasing, it is a potential avenue for future research.

Beyond and Inside Lexicon. My thesis work has primarily focused on lexical variation (i.e., variation at the word level). However, stylistic variation exists at other layers as well: semantic, syntactic, and phonological. Sociolinguistic inquiry of variation has focused on variation at these other layers, but such studies have primarily focused on a single or a handful of variants. A limited number of computational work focused on other layers of variation (e.g., [94, 274]). An important extension of my thesis work is to investigate stylistic variation in terms of syntactic and orthographic markers of linguistic style and the social dynamics of this variation.

Linguistic and Social Dynamics of Online Communities. The availability of large amounts of longitudinal data from various online communities enables the study of linguistic and social dynamics of communities, which are hard to study in the offline settings. In the previous three chapters of this thesis, I considered community linguistic

norms and how people adapt to those norms. There are several other questions about the linguistic and social dynamics of online communities that are unexplored. Some examples include: (1) how can we track community-level and user-level linguistic norm changes over time? (2) how does the amount and speed of linguistic norm adaptation affect the success of a community? (3) how do exogenous shocks, such as the influx of new users, changes in community policies, and external events related to the topic of the community, affect community linguistic norms? (4) how does community's tie-strength influence changes in community linguistic norms? Future work could investigate these questions to further understand the linguistic style variation in online communities.

Variation Across Multiple Social Platforms. Current studies of linguistic style variation are limited to individuals' language usage in a single social platform. However, different social platforms provide various communicative and linguistic affordances that could influence the language use. Therefore, using a single data source to study any online writing phenomenon has the risk of introducing various biases and potentially limit the generalizability of findings [157]. For example, as discussed in § 2.3.2, prior work found different notions of audience in the social platforms such as Twitter [34] and Facebook [104]. These varied perceptions of audience could lead to varied patterns of linguistic style-shifting. Therefore, future work should consider comparison of individuals' language usage across multiple social platforms and the generalizability of findings across platforms or data sources. A major challenge in studying this is the limited data availability; networked comment platforms such as Disqus¹ could be potential data sources.

Effects of Topic on Style Variation. While various social attributes and social context influence stylistic variation in online writing, the topic of discourse is a potential

¹<https://help.disqus.com/what-is-disqus/what-is-disqus>

confound that could influence stylistic variation in online discourse. For example, in their study of style-shifting using Scottish variants, Shoemark et al. [107] found that both topic and audience influence the choice of Scottish variants. While we found (Chapter 4) that larger audience size (i.e., when using hashtags) inhibits the usage of local non-standard variants in general discourse on Twitter, Stewart et al. [106] found the reverse when analyzing tweets on the specific topic of Catalonia Referendum. Isolating the effects of topic on stylistic variation is challenging and it is an interesting direction for future work.

Appendices

APPENDIX A

LINGUISTIC VARIABLES

A.1 Example tweets with linguistic variables

Table A.1: Examples of each of the linguistic variables in bold, with glosses in parentheses.

- (1) **lml** (love my life) thank u
- (2) that was **od** (very) crazy
- (3) **dat** (that) was odee (very) loud
- (4) this line **deadass** (really) **raps** (wraps) around the whole store
- (5) my sis **sed** (said) i wouldd lol
- (6) o **werd** (really) ? lol i really like your style
- (7) **ima** (I am a) la girl but these other ones be reaal **fk**n (fuking) ratchet man
- (8) i'm so complicated **smh** (somehow) **ion** (I don't) understand myself so **yu** (you) definitely can't understand me !!
- (9) **finna** (fixing to) get ah (a) nap in and party **lata** (later)
- (10) **imm** (I am) 2 hours late **fa** (for) **skool** (school).
- (11) ohh **lawd** (Lord) i can't stand to be lied to !
- (12) say **wats** (what's) wrong **wit** (with) **oomf** (one of my follower) she got life messed up .. #shitno
- (13) **lil** (little) man **wya** (where you at) i'm **bouta** (about) come scoop u
- (14) this **mf** (motherfucker) behind me is breathing so hard. i wanna **ctfu** (crack the fuck up) .
- (15) im ready **fa** (for) college
- (16) **fa** (for) **sho** (sure), hoe (derogatory term for woman, or promiscuous person of any gender)
- (17) lol **jawns** (people) be lying
- (18) bored as **shyt** (shit) ... **lls** (laughing like shit)
- (19) yeah , **yu** (you) act like **yu** (you) don't **noe** (know) me !
- (20) right lol . like **tf** (the fuck) is this bahaha
- (21) **dang** (damn) didn't even mention ques .. hmm
- (22) its copyrighted lol, **sike** (just kidding) u can use it
- (23) #whatsthepoint of going to school tomorrow .? **lbvs** (laughing but very serious) .
- (24) i **slick** (sort of, might) wanna move to memphis one day
- (25) i'm **slick** (really) serious af (as fuck)
- (26) these **hoes** (women) **b** (be) hatin hard fr **fr** (for real)
- (27) oh **ard** (alright) just checkin lol

A.2 Lexical variable frequencies for each MSA

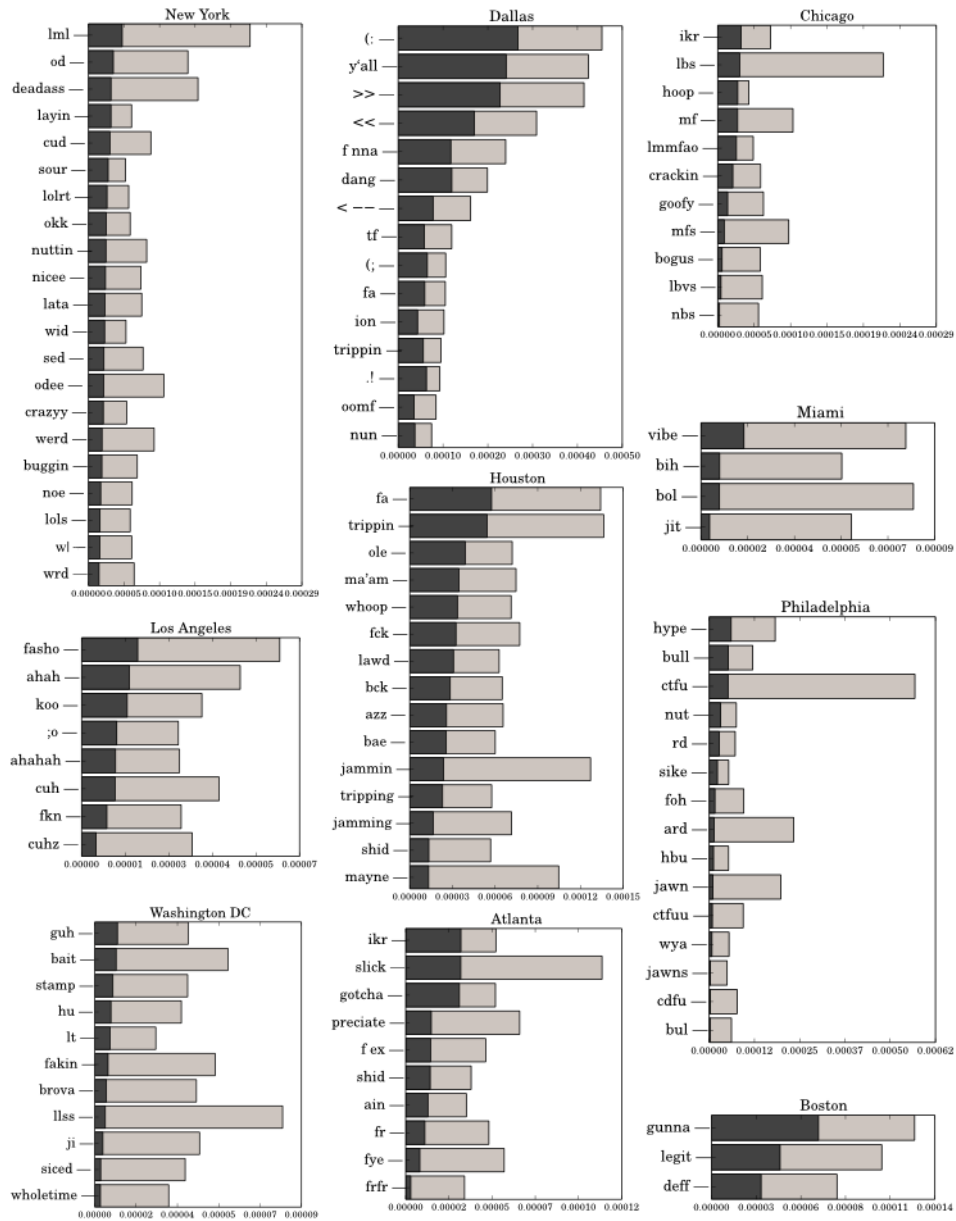


Figure A.1: Frequencies of each linguistic variable from each metropolitan statistical area (MSA), with the local frequency in gray and the national frequency in black.

APPENDIX B
EMOTICON TOKENS

Table B.1: Top-twenty extracted emoticon tokens and their cumulative frequencies.

March 2014		March 2015	
Emoticon	Cumulative %	Emoticon	Cumulative %
:)	40.60	:)	40.31
:(53.15	:(51.85
;))	61.23	:D	59.36
:-)	68.20	;))	66.37
:D	74.97	:-)	72.50
:/	78.55	:/	77.64
:P	80.60	:?(79.75
;)	82.30	:P	81.77
:p	83.57	;)	83.54
-_-	84.81	:-)	84.57
:-)	85.92	:p	85.56
:o	86.51	^^	86.32
:O	87.03	=)	86.96
=)	87.50	-_-	87.53
^^	87.91	*^*	88.09
:-D	88.32	:-D	88.63
:-/	88.66	:O	89.17
;D	88.99	:o	89.71
-_- -	89.29	(:-	90.03
(:-	89.51	:-/	90.34

APPENDIX C
ROBUSTNESS CHECK FOR ARTICLE-LEVEL EFFECTS OF NPOV
TAGGING

The distribution of time-intervals between the addition of the NPOV tag and the data collection time is shown in Figure C.1.

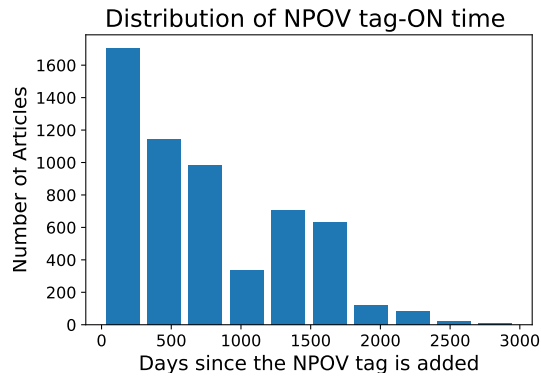


Figure C.1: The distribution of time-intervals between the addition of the NPOV tag and the data collection time.

Table C.1 shows results of the ITS analysis for RQ1 on the subset of data after removing articles which are NPOV tagged for more than three years. The treatment effect remains significant after excluding these long-running disputes.

Table C.1: Results of RQ1 Analysis robustness check on the subset of data after removing articles which are tagged for more than three years. Article lexicon coverage is computed for the textual content of the article results from each revision. Coefficient β_3 indicates the change in slope after treatment. Statistical significance after correcting for multiple comparisons using Benjamini and Hochberg [2] adjustment for false discovery rate are shown; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Percentage change is computed as the change in post-treatment lexicon coverage after 40 revisions.

Lexicon	β_3	<i>p</i> -value		% change
Words to Watch	-1.33e-05	3.43e-15	***	-5.14
Hedges	-3.50e-06	3.39e-06	***	-2.60
Assertives	-3.41e-06	4.54e-11	***	-5.43
Positive Words	-2.69e-05	< 2e-16	***	-4.32
Negative Words	-1.20e-05	< 7.23e-12	***	-2.29
Factives	5.37e-07	6.57e-01		+1.86
Implicatives	-1.96e-06	5.00e-05	***	-3.46
Report Verbs	-6.61e-06	1.20e-05	**	-3.83
Strong Subjectives	-3.68e-05	< 2e-16	***	-5.13
Weak Subjectives	-2.97e-05	< 2e-16	***	-1.70

REFERENCES

- [1] D. Biber and E. Finegan, “Styles of stance in english: Lexical and grammatical marking of evidentiality and affect,” *Text*, vol. 9, no. 1, pp. 93–124, 1989.
- [2] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.
- [3] M. Bakhtin, “The dialogic imagination, trans,” *Caryl Emerson and Michael Holquist (Austin: University of Texas Press, 1981)*, vol. 69, 1981.
- [4] D. Crystal, *Language and the Internet*, 2nd ed. Cambridge University Press, sep 2006.
- [5] J. Eisenstein, “What to do about bad language on the internet,” in *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, 2013, pp. 359–369.
- [6] S. C. Herring, “A faceted classification scheme for computer-mediated discourse,” *Language@ internet*, vol. 4, no. 1, 2007.
- [7] J. Androutsopoulos, “Language change and digital media: a review of conceptions and evidence,” in *Standard Languages and Language Standards in a Changing Europe*, N. Coupland and T. Kristiansen, Eds. Oslo: Novus, 2011.
- [8] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: a content-based approach to geo-locating twitter users,” in *cikm*, 2010, pp. 759–768.
- [9] J. Eisenstein, B. O’Connor, N. A. Smith, and E. P. Xing, “A latent variable model for geographic lexical variation,” in *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, 2010, pp. 1277–1287.
- [10] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, “Effects of age and gender on blogging,” in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, vol. 6, 2006, pp. 199–205.
- [11] S. Rosenthal and K. McKeown, “Age prediction in blogs: A study of style, content, and online behavior in pre- and Post-Social media generations,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 763–772.
- [12] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, ““how old do you think i am?” a study of language and age in twitter,” in *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 2013, pp. 439–448.

- [13] S. C. Herring and J. C. Paolillo, “Gender and genre variation in weblogs,” *Journal of Sociolinguistics*, vol. 10, no. 4, pp. 439–459, 2006.
- [14] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, “Classifying latent user attributes in twitter,” in *Proceedings of Workshop on Search and mining user-generated contents*, 2010.
- [15] J. D. Burger and J. C. Henderson, “An exploration of observable features related to blogger age.” in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006.
- [16] D. Bamman, J. Eisenstein, and T. Schnoebelen, “Gender identity and lexical variation in social media,” *Journal of Sociolinguistics*, vol. 18, no. 2, pp. 135–160, 2014.
- [17] J. Eisenstein, A. Ahmed, and E. P. Xing, “Sparse additive generative models of text,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2011, pp. 1041–1048.
- [18] M. A. Halliday and R. Hasan, *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford University Press, 1989.
- [19] A. Bell, “Language style as audience design,” *Language in Society*, vol. 13, no. 2, pp. 145–204, 1984.
- [20] P. Eckert, *Jocks and burnouts: Social categories and identity in the high school*. Teachers College Press, 1989.
- [21] —, *Linguistic variation as social practice*. Blackwell, 2000.
- [22] L. Stark and K. Crawford, “The conservatism of emoji: Work, affect, and communication,” *Social Media + Society*, vol. 1, no. 2, 2015.
- [23] J. Chen, H. Xu, and A. B. Whinston, “Moderated online communities and quality of user-generated content,” *Journal of Management Information Systems*, vol. 28, no. 2, pp. 237–268, 2011.
- [24] J. K. Chambers and N. Schilling, *The handbook of language variation and change*. John Wiley & Sons, 2013, vol. 129.
- [25] J. Milroy and L. Milroy, *Authority in language: Investigating standard English*. Routledge, 2012.
- [26] P. Eckert and J. R. Rickford, *Style and sociolinguistic variation*. Cambridge University Press, 2001.
- [27] N. C. Dorian, “Defining the speech community,” *Sociolinguistic variation in speech communities*, pp. 25–33, 1982.

- [28] D. Hymes, “Models of the interaction of language and social life: toward a descriptive theory,” *Intercultural discourse and communication: The essential readings*, pp. 4–16, 2005.
- [29] J. J. Gumperz, “The speech community,” *Linguistic anthropology: A reader*, vol. 1, p. 66, 2009.
- [30] E. Dresner and S. C. Herring, “Functions of the nonverbal in CMC: Emoticons and illocutionary force,” *Communication Theory*, vol. 20, no. 3, pp. 249–268, 2010.
- [31] M. Danesi, *The semiotics of emoji: The rise of visual language in the age of the internet*. Bloomsbury Publishing, 2016.
- [32] P. Eckert, “Variation, convention, and social meaning,” in *Annual Meeting of the Linguistic Society of America. Oakland CA*, vol. 7, 2005.
- [33] J. J. Gumperz, *Language and social identity*. Cambridge University Press, 1982, vol. 2.
- [34] A. E. Marwick and d. boyd, “I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience,” *New Media & Society*, vol. 13, no. 1, pp. 114–133, 2011.
- [35] H. Rheingold, *The virtual community: Homesteading on the electronic frontier*. MIT press, 2000.
- [36] J. Preece and D. Maloney-Krichmar, “Online communities: Design, theory, and practice,” *Journal of Computer-Mediated Communication*, vol. 10, no. 4, p. JCMC10410, 2005.
- [37] A. J. Kim, *Community building on the web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc., 2000.
- [38] G. Burnett and L. Bonnici, “Beyond the faq: Explicit and implicit norms in usenet newsgroups,” *Library & Information Science Research*, vol. 25, no. 3, pp. 333–351, 2003.
- [39] F. Casey, J. A. Jiang, J. McCann, K. Frye, and J. R. Brubaker, “Reddit rules! characterizing an ecosystem of governance,” in *ICWSM*, 2018.
- [40] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, “No country for old members: User lifecycle and linguistic change in online communities,” in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 307–318.
- [41] P. Eckert, “Communities of practice,” *Encyclopedia of language and linguistics*, vol. 2, no. 2006, pp. 683–685, 2006.

- [42] U. Pavalanathan and J. Eisenstein, “Confounds and consequences in geotagged twitter data,” in *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, September 2015. [Online]. Available: <http://www.aclweb.org/anthology/D/D15/D15-1256.pdf>
- [43] W. Labov, *The social stratification of English in New York City*. Cambridge, U.K.: Cambridge University Press, 2006.
- [44] P. Eckert, “(ay) goes to the city. exploring the expressive use of variation,” *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pp. 47–68, 1995.
- [45] J. C. Paolillo, “Language variation on internet relay chat: A social network approach,” *Journal of Sociolinguistics*, vol. 5, no. 2, pp. 180–213, 2001.
- [46] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, “Mark my words! linguistic style accommodation in social media,” in *Proceedings of the Conference on World-Wide Web (WWW)*, 2011, pp. 745–754.
- [47] T. Tran and M. Ostendorf, “Characterizing the language of online communities and its relation to community reception,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.
- [48] C. Tan and L. Lee, “All who wander: On the prevalence and characteristics of multi-community engagement,” in *Proceedings of the Conference on World-Wide Web (WWW)*, 2015, pp. 1056–1066.
- [49] B. L. Monroe, M. P. Colaresi, and K. M. Quinn, “Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict,” *Political Analysis*, vol. 16, no. 4, pp. 372–403, 2008.
- [50] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, “A computational approach to politeness with application to social factors,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 250–259.
- [51] E. Pavlick and J. Tetreault, “An empirical analysis of formality in online communication,” *Transactions of the Association for Computational Linguistics (TACL)*, vol. 4, pp. 61–74, 2016.
- [52] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg, “Echoes of power: Language effects and power differences in social interaction,” in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 699–708.
- [53] E. Gilbert, “Phrases that signal workplace hierarchy,” in *Proceedings of Computer-Supported Cooperative Work (CSCW)*, 2012, pp. 1037–1046.

- [54] V. Prabhakaran, O. Rambow, and M. Diab, “Predicting overt display of power in written dialogs,” in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2012, pp. 518–522.
- [55] B. Johnstone and S. F. Kiesling, “Indexicality and experience: Exploring the meanings of /aw/-monophthongization in pittsburgh1,” *Journal of sociolinguistics*, vol. 12, no. 1, pp. 5–33, 2008.
- [56] A. Gatt and E. Kraehmer, “Survey of the state of the art in natural language generation: Core tasks, applications and evaluation,” *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.
- [57] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, “Normalization of non-standard words,” *Computer speech & language*, vol. 15, no. 3, pp. 287–333, 2001.
- [58] A. Ritter, C. Cherry, and B. Dolan, “Unsupervised modeling of twitter conversations,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 172–180.
- [59] D. Nguyen, A. S. Doğruöz, C. P. Rosé, and F. de Jong, “Computational sociolinguistics: A survey,” *Computational Linguistics*, vol. 42, no. 3, pp. 537–593, September 2016.
- [60] Y. Yang, M.-W. Chang, and J. Eisenstein, “Toward socially-infused information extraction: Embedding authors, mentions, and entities,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.
- [61] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: LIWC and computerized text analysis methods,” *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.
- [62] M. Silverstein, “Indexical order and the dialectics of sociolinguistic life,” *Language & Communication*, vol. 23, no. 3, pp. 193–229, 2003.
- [63] P. Eckert, “Variation and the indexical field,” *Journal of Sociolinguistics*, vol. 12, no. 4, pp. 453–476, 2008.
- [64] J. T. Irvine, *Style and sociolinguistic variation*. Cambridge University Press, 2001, ch. "Style" as distinctiveness: the culture and ideology of linguistic differentiation, pp. 21–43.
- [65] M. Bucholtz and K. Hall, “Language and identity,” in *A Companion to linguistic anthropology*, A. Duranti, Ed. Blackwell, 2004, pp. 369–394.
- [66] M. Bucholtz, “From stance to style,” *Stance: Sociolinguistic Perspectives*, p. 146, 2009.

- [67] A. Jaffe, "Introduction: Non-standard orthography and non-standard speech," *Journal of Sociolinguistics*, vol. 4, no. 4, pp. 497–513, 2000.
- [68] J. Carey, "Paralanguage in computer mediated communication," in *Proceedings of the 18th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1980, pp. 67–69.
- [69] Y. M. Kalman and D. Gergle, "Letter repetitions in computer-mediated communication: A unique link between spoken and online language," *Computers in Human Behavior*, vol. 34, pp. 187–193, 2014.
- [70] M. Bucholtz, N. Bermudez, V. Fung, L. Edwards, and R. Vargas, "Hella nor cal or totally so cal? the perceptual dialectology of california," *Journal of English Linguistics*, vol. 35, no. 4, pp. 325–352, 2007.
- [71] B. Schading, R. Schading, and V. R. Slayton, *A Civilian's Guide to the US Military: A Comprehensive Reference to the Customs, Language & Structure of the Armed Forces*. Writer's Digest Books, Incorporated, 2007.
- [72] U. Pavalanathan, V. V. Datla, S. Volkova, L. Charles-Smith, M. Pirrung, J. J. Harrison, A. Chappell, and C. D. Corley, "Discourse, health and well-being of military populations through the social media lens." in *AAAI Workshop: WWW and Population Health Intelligence*, 2016.
- [73] J. Androutsopoulos, "Non-standard spellings in media texts: The case of German fanzines," *Journal of Sociolinguistics*, vol. 4, no. 4, pp. 514–533, 2000.
- [74] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "Diffusion of lexical change in social media," *PLoS ONE*, vol. 9, November 2014.
- [75] L. L. Rezabek and J. J. Cochenour, "Visual cues in computer-mediated communication: Supplementing text with emoticons." *Journal of Visual Literacy*, vol. 18, no. 2, 1998.
- [76] A. Wolf, "Emotional expression online: Gender differences in emoticon use," *CyberPsychology & Behavior*, vol. 3, no. 5, pp. 827–833, 2000.
- [77] D. Derks, A. E. Bos, and J. Von Grumbkow, "Emoticons and social interaction on the internet: the importance of social context," *Computers in human behavior*, vol. 23, no. 1, pp. 842–849, 2007.
- [78] T. Schnoebelen, "Do you smile with your nose? Stylistic variation in Twitter emoticons," *University of Pennsylvania Working Papers in Linguistics*, vol. 18, no. 2, p. 14, 2012.
- [79] J. Park, V. Barash, C. Fink, and M. Cha, "Emoticon style: Interpreting differences in emoticons across cultures." in *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 2013.

- [80] C. Kelly, “Do you know what i mean > :(: A linguistic study of the understanding of emoticons and emojis in text messages,” pp. kmstad University, School of Education, Humanities and Social Science, 2015.
- [81] P. K. Novak, J. Smailovic, B. Sluban, and I. Mozetic, “Sentiment of emojis,” *CoRR*, vol. abs/1509.07761, 2015.
- [82] H. Cramer, P. de Juan, and J. Tetreault, “Sender-intended functions of emojis in us messaging,” in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 2016, pp. 504–509.
- [83] H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht, “Blissfully happy” or “ready to fight”: Varying interpretations of emoji,” *Proceedings of ICWSM 2016*, 2016.
- [84] F. Barbieri, G. Kruszewski, F. Ronzano, and H. Saggion, “How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics,” in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 531–535.
- [85] N. Ljubešić and D. Fišer, “A global analysis of emoji usage,” in *Proceedings of the 10th Web as Corpus Workshop*, 2016, pp. 82–89.
- [86] R. P. López and F. Cap, “Did you ever read about frogs drinking coffee? investigating the compositionality of multi-emoji expressions,” in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 113–117.
- [87] S. M. Seyednezhad and R. Menezes, “Understanding subject-based emoji usage using network science,” in *Workshop on Complex Networks CompleNet*. Springer, 2017, pp. 151–159.
- [88] J. Ge and U. Gretzel, “Emoji rhetoric: a social media influencer perspective,” *Journal of Marketing Management*, pp. 1–24, 2018.
- [89] P. Eckert, “Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation,” *Annual Review of Anthropology*, vol. 41, pp. 87–100, 2012.
- [90] W. Labov, “The linguistic variable as a structural unit.” *Washington Linguistics Review*, 1966.
- [91] W. A. Wolfram, “A sociolinguistic description of detroit negro speech,” *Urban Language Series, No. 5.*, 1969.
- [92] P. Trudgill, S. Trudgill *et al.*, *The social differentiation of English in Norwich*. CUP Archive, 1974, vol. 13.

- [93] B. Wing and J. Baldrige, “Simple supervised document geolocation with geodesic grids,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 955–964.
- [94] G. Doyle, “Mapping dialectal variation by querying social media,” in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 2014, pp. 98–106.
- [95] S. L. Blodgett, L. Green, and B. O’Connor, “Demographic dialectal variation in social media: A case study of african-american english,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1119–1130.
- [96] J. Eisenstein, B. O’Connor, N. A. Smith, and E. P. Xing, “Mapping the geographical diffusion of new words,” ArXiv, Tech. Rep. 1210.5268, oct 2012.
- [97] N. Coupland, *Style: Language Variation and Identity*. Cambridge University Press, jul 2007.
- [98] H. Giles, J. Coupland, and N. Coupland, *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, 1991.
- [99] J. Androutsopoulos and E. Ziegler, “Exploring language variation on the internet: Regional speech in a chat community,” in *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe, ICLaVE*, vol. 2, 2004, pp. 99–111.
- [100] J. Milroy and L. Milroy, “Linguistic change, social network and speaker innovation,” *Journal of linguistics*, vol. 21, no. 02, pp. 339–384, 1985.
- [101] J. Androutsopoulos, “Code-switching in computer-mediated communication,” *Pragmatics of computer-mediated communication*, pp. 667–694, 2013.
- [102] I. Johnson, “Audience design and communication accommodation theory: Use of Twitter by Welch-English biliterates,” in *Social Media and Minority Languages: Convergence and the Creative Industries*, E. H. G. Jones and E. Uribe-Jongbloed, Eds. Bristol, U.K.: Multilingual Matters, 2013, pp. 99–118.
- [103] D. Nguyen, D. Trieschnigg, and L. Cornips, “Audience and the use of minority languages on twitter.” in *ICWSM*, 2015, pp. 666–669.
- [104] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer, “Quantifying the invisible audience in social networks,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 21–30.
- [105] P. Shoemark, D. Sur, L. Shrimpton, I. Murray, and S. Goldwater, “Aye or naw, whit dae ye hink? scottish independence and linguistic identity on social media,” in *Proceedings of the 15th Conference of the European Chapter of the*

Association for Computational Linguistics: Volume 1, Long Papers, vol. 1, 2017, pp. 1239–1248.

- [106] I. Stewart, Y. Pinter, and J. Eisenstein, “Si o no, que penses? catalonian independence and linguistic identity on social media,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, vol. 2, 2018, pp. 136–141.
- [107] P. Shoemark, J. Kirby, and S. Goldwater, “Topic and audience effects on distinctively scottish vocabulary usage in twitter data,” in *Proceedings of the Workshop on Stylistic Variation*, 2017, pp. 59–68.
- [108] E. Riloff and J. Wiebe, “Learning extraction patterns for subjective expressions,” in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 2003, pp. 105–112.
- [109] J. Wiebe, T. Wilson, and C. Cardie, “Annotating expressions of opinions and emotions in language,” *Language resources and evaluation*, vol. 39, no. 2, pp. 165–210, 2005.
- [110] J. W. Du Bois, “The stance triangle,” in *Stancetaking in discourse*, R. Englebretson, Ed. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2007, pp. 139–182.
- [111] A. Jaffe, *Stance: Sociolinguistic Perspectives*. Oxford University Press, 2009.
- [112] S. F. Kiesling, “Style as stance,” *Stance: sociolinguistic perspectives*, pp. 171–194, 2009.
- [113] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, “Knowledge sharing and yahoo answers: Everyone knows something,” in *Proceedings of the Conference on World-Wide Web (WWW)*, 2008, pp. 665–674.
- [114] J. Hessel, C. Tan, and L. Lee, “Science, askscience, and badscience: On the coexistence of highly related communities,” in *Proceedings of the International Conference on Web and Social Media (ICWSM)*. Menlo Park, California: AAAI Publications, 2014, pp. 171–180.
- [115] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, “Group formation in large social networks: Membership, growth, and evolution,” in *Proceedings of Knowledge Discovery and Data Mining (KDD)*, 2006, pp. 44–54.
- [116] M. A. Hogg and S. A. Reid, “Social identity, self-categorization, and the communication of group norms,” *Communication theory*, vol. 16, no. 1, pp. 7–30, 2006.
- [117] K. Bergstrom, ““don’t feed the troll”: Shutting down debate about community expectations on reddit.com,” *First Monday*, vol. 16, no. 8, 2011.

- [118] Wikipedia, “Wikipedia:manual of style,” 2018, [Online; accessed 15-April-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style
- [119] C. Lampe and P. Resnick, “Slash (dot) and burn: distributed moderation in a large online conversation space,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 543–550.
- [120] S. Sood, J. Antin, and E. Churchill, “Profanity use in online communities,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1481–1490.
- [121] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Antisocial behavior in online discussion communities,” in *ICWSM*, 2015, pp. 61–70.
- [122] R. S. Geiger and D. Ribes, “The work of sustaining order in wikipedia: the banning of a vandal,” in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 2010, pp. 117–126.
- [123] Y. Ren, F. M. Harper, S. Drenner, L. Terveen, S. Kiesler, J. Riedl, and R. E. Kraut, “Building member attachment in online communities: Applying theories of group identity and interpersonal bonds,” *Mis Quarterly*, pp. 841–864, 2012.
- [124] S. L. Bryant, A. Forte, and A. Bruckman, “Becoming wikipedia: transformation of participation in a collaborative online encyclopedia,” in *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*. ACM, 2005, pp. 1–10.
- [125] S. Chancellor, A. Hu, and M. De Choudhury, “Norms matter: Contrasting social support around behavior change in online weight loss communities,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018.
- [126] Wikipedia, “Wikipedia:expectations and norms of the wikipedia community,” 2018, [Online; accessed 15-April-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Expectations_and_norms_of_the_Wikipedia_community
- [127] —, “Neutral point of view,” 2018, [Online; accessed 15-April-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view
- [128] J. M. Reagle Jr, ““be nice”: Wikipedia norms for supportive communication,” *New Review of Hypermedia and Multimedia*, vol. 16, no. 1-2, pp. 161–180, 2010.
- [129] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl, “Creating, destroying, and restoring value in wikipedia,” in *Proceedings of the 2007 international ACM conference on Supporting group work*. ACM, 2007, pp. 259–268.

- [130] D. Hymes, “On communicative competence,” *sociolinguistics*, vol. 269293, pp. 269–293, 1972.
- [131] P. Bourdieu, *Language and symbolic power*. Harvard University Press, 1991.
- [132] N. Lipka and B. Stein, “Identifying featured articles in wikipedia: writing style matters,” in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1147–1148.
- [133] K. Al Khatib, H. Schütze, and C. Kantner, “Automatic detection of point of view differences in wikipedia.” in *COLING*. Citeseer, 2012, pp. 33–50.
- [134] M. Anderka, B. Stein, and N. Lipka, “Predicting quality flaws in user-generated content: the case of wikipedia,” in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 981–990.
- [135] M. Anderka and B. Stein, “A breakdown of quality flaws in wikipedia,” in *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*. ACM, 2012, pp. 11–18.
- [136] M. Harpalani, M. Hart, S. Singh, R. Johnson, and Y. Choi, “Language of vandalism: Improving wikipedia vandalism detection via stylometric analysis,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 83–88.
- [137] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, “Linguistic models for analyzing and detecting biased language,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 1650–1659.
- [138] S. Kiesler, R. Kraut, P. Resnick, and A. Kittur, “Regulating behavior in online communities,” *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press, Cambridge, MA, 2012.
- [139] M. B. V. Dyke, “Reddit users revolt after site bans “fat people hate” and other communities,” *Buzzfeed News*, 2015.
- [140] J. Grimmelmann, “The virtues of moderation,” *Yale JL & Tech.*, vol. 17, p. 42, 2015.
- [141] N. A. Draper, “Distributed intervention: networked content moderation in anonymous mobile spaces,” *Feminist Media Studies*, pp. 1–17, 2018.
- [142] H. News, “Hacker news faq,” 2018, [Online; accessed 22-July-2018]. [Online]. Available: <https://news.ycombinator.com/newsfaq.html>

- [143] M. Rahman, B. Carbutar, J. Ballesteros, G. Burri, and D. H. Chau, “Turning the tide: Curbing deceptive yelp behaviors,” in *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, 2014, pp. 244–252.
- [144] C. Lampe, P. Zube, J. Lee, C. H. Park, and E. Johnston, “Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums,” *Government Information Quarterly*, vol. 31, no. 2, pp. 317–326, 2014.
- [145] I. Beschastnikh, T. Kriplean, and D. W. McDonald, “Wikipedian self-governance in action: Motivating the policy lens.” in *ICWSM*, 2008.
- [146] A. Di Iorio, F. Vitali, and S. Zacchiroli, “Wiki content templating,” in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 615–624.
- [147] M. Jan Piskorski and A. Gorbatâi, “Testing coleman’s social-norm enforcement mechanism: Evidence from wikipedia,” *American Journal of Sociology*, vol. 122, no. 4, pp. 1183–1222, 2017.
- [148] K. Wise, B. Hamman, and K. Thorson, “Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate,” *Journal of Computer-Mediated Communication*, vol. 12, no. 1, pp. 24–41, 2006.
- [149] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, “You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, p. 31, 2017.
- [150] U. Matzat and G. Rooks, “Styles of moderation in online health and support communities: An experimental comparison of their acceptance and effectiveness,” *Computers in Human Behavior*, vol. 36, pp. 65–75, 2014.
- [151] K. K. Seo, “Utilizing peer moderating in online discussions: Addressing the controversy between teacher moderation and nonmoderation,” *The American Journal of Distance Education*, vol. 21, no. 1, pp. 21–36, 2007.
- [152] A. Halfaker, A. Kittur, and J. Riedl, “Don’t bite the newbies: how reverts affect the quantity and quality of wikipedia work,” in *Proceedings of the 7th international symposium on wikis and open collaboration*. ACM, 2011, pp. 163–172.
- [153] A. Halfaker, R. S. Geiger, J. T. Morgan, and J. Riedl, “The rise and decline of an open collaboration system: How wikipedia’s reaction to popularity is causing its decline,” *American Behavioral Scientist*, vol. 57, no. 5, pp. 664–688, 2013.
- [154] R. S. Geiger, A. Halfaker, M. Pinchuk, and S. Walling, “Defense mechanism or socialization tactic? improving,” in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

- [155] L. Muchnik, S. Aral, and S. J. Taylor, “Social influence bias: A randomized experiment,” *Science*, vol. 341, no. 6146, pp. 647–651, 2013.
- [156] G. King, J. Pan, and M. E. Roberts, “Reverse-engineering censorship in china: Randomized experimentation and participant observation,” *Science*, vol. 345, no. 6199, p. 1251722, 2014.
- [157] A. Olteanu, E. Kiciman, and C. Castillo, “A critical review of online social data: Biases, methodological pitfalls, and ethical boundaries,” in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 785–786.
- [158] S. Gouws, D. Metzler, C. Cai, and E. Hovy, “Contextual bearing on linguistic variation in social media,” in *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011, pp. 20–29.
- [159] P. R. Rosenbaum and D. B. Rubin, “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
- [160] C. E. Frangakis and D. B. Rubin, “Principal stratification in causal inference,” *Biometrics*, vol. 58, no. 1, pp. 21–29, 2002.
- [161] V. L. Dos Reis and A. Culotta, “Using matched samples to estimate the effects of exercise on mental health from twitter,” in *Proceedings of the twenty-ninth AAAI conference on artificial intelligence*, 2015, pp. 182–188.
- [162] M. De Choudhury, S. Sharma, and E. Kiciman, “Characterizing dietary choices, nutrition, and language in food deserts via social media,” in *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*. ACM, 2016, pp. 1157–1170.
- [163] D. B. Rubin, E. A. Stuart *et al.*, “Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions,” *The Annals of Statistics*, vol. 34, no. 4, pp. 1814–1826, 2006.
- [164] A. Abadie, “Semiparametric difference-in-differences estimators,” *The Review of Economic Studies*, vol. 72, no. 1, pp. 1–19, 2005.
- [165] J. L. Bernal, S. Cummins, and A. Gasparrini, “Interrupted time series regression for the evaluation of public health interventions: a tutorial,” *International journal of epidemiology*, vol. 46, no. 1, pp. 348–355, 2017.
- [166] T. M. Mattia Samory, “Conspiracies online: User discussions in a conspiracy community following dramatic events,” in *International AAAI Conference on Web and Social Media*, 2018.

- [167] G. Caldarelli, A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni, and G. Riotta, “A multi-level geographical study of italian political elections from twitter data,” *PloS one*, vol. 9, no. 5, p. e95809, 2014.
- [168] D. A. Broniatowski, M. J. Paul, and M. Dredze, “National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic,” *PloS one*, vol. 8, no. 12, p. e83672, 2013.
- [169] B. Gonçalves and D. Sánchez, “Crowdsourcing dialect characterization through twitter,” *PloS one*, vol. 9, no. 11, p. e112074, 2014.
- [170] J. Eisenstein, “Written dialect variation in online social media,” in *Handbook of Dialectology*, C. Boberg, J. Nerbonne, and D. Watt, Eds. Wiley, 2017.
- [171] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsoulis, “Discovering geographical topics in the twitter stream,” in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 769–778.
- [172] B. Han, P. Cook, and T. Baldwin, “Text-based twitter user geolocation prediction.” *Journal of Artificial Intelligence Research (JAIR)*, vol. 49, pp. 451–500, 2014.
- [173] C. Mallinson, B. Childs, and G. Van Herk, *Data collection in sociolinguistics: methods and applications*. Routledge, 2013.
- [174] J. Ritchie, J. Lewis, C. M. Nicholls, R. Ormston *et al.*, *Qualitative research practice: A guide for social science students and researchers*. Sage, 2013.
- [175] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, “Understanding the demographics of twitter users.” in *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 2011, pp. 554–557.
- [176] B. Hecht and M. Stephens, “A tale of cities: Urban biases in volunteered geographic information,” in *Proceedings of the International Conference on Web and Social Media (ICWSM)*. Menlo Park, California: AAAI Publications, 2014, pp. 197–205.
- [177] P. A. Longley, M. Adnan, and G. Lansley, “The geotemporal demographics of twitter usage,” *Environment and Planning A*, vol. 47, no. 2, pp. 465–484, 2015.
- [178] M. Malik, H. Lamba, C. Nakos, and J. Pfeffer, “Population bias in geotagged tweets,” in *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*. The AAAI Press, 2015, pp. 18–27.
- [179] S. Yardi, D. Romero, G. Schoenebeck *et al.*, “Detecting spam in a twitter network,” *First Monday*, vol. 15, no. 1, 2009.
- [180] A. Lenhart, “Mobile access shifts social media use and other online activities,” Pew Research Center, Tech. Rep., April 2015.

- [181] K. Zickuhr, “Location-based services,” Pew Research Center, Tech. Rep., September 2013.
- [182] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldrige, “Supervised text-based geolocation using language models on an adaptive grid,” in *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, 2012, pp. 1500–1510.
- [183] J. Chang, I. Rosenn, L. Backstrom, and C. Marlow, “ePluribus: Ethnicity on social networks,” in *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 2010, pp. 18–25.
- [184] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, “Discriminating gender on twitter,” in *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, 2011, pp. 1301–1309.
- [185] E. Finegan and D. Biber, “Register and social dialect variation: An integrated approach,” *Sociolinguistic perspectives on register*, pp. 315–347, 1994.
- [186] M. Naaman, H. Becker, and L. Gravano, “Hip and trendy: Characterizing emerging trends on twitter,” *J. Am. Soc. Inf. Sci.*, vol. 62, no. 5, pp. 902–918, 2011.
- [187] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?” in *Proceedings of the Conference on World-Wide Web (WWW)*, 2010, pp. 591–600.
- [188] D. Jurgens, “That’s what friends are for: Inferring location in online social media platforms based on social relationships.” *ICWSM*, vol. 13, pp. 273–282, 2013.
- [189] S. A. Tagliamonte, *Analysing Sociolinguistic Variation*. Cambridge University Press, 2006.
- [190] D. E. Ho, K. Imai, G. King, and E. A. Stuart, “MatchIt: Nonparametric preprocessing for parametric causal inference,” *Journal of Statistical Software*, vol. 42, no. 8, pp. 1–28, 2011.
- [191] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist’s companion*. Princeton University Press, 2008.
- [192] M. A. Walker, P. Anand, R. Abbott, and R. Grant, “Stance classification using dialogic properties of persuasion,” in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2012, pp. 592–596.
- [193] V. Freeman, R. Wright, G.-A. Levow, Y. Luan, J. Chan, T. Tran, V. Zayats, M. Antoniak, and M. Ostendorf, “Phonetic correlates of stance-taking,” *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 2175–2175, 2014.

- [194] E. Ochs, “Constructing social identity: A language socialization perspective,” *Research on language and social interaction*, vol. 26, no. 3, pp. 287–306, 1993.
- [195] E. Kärkkäinen, “Stance taking in conversation: From subjectivity to intersubjectivity,” *Text & Talk-An Interdisciplinary Journal of Language, Discourse Communication Studies*, vol. 26, no. 6, pp. 699–731, 2006.
- [196] T. Keisanen, “Stancetaking as an interactional activity: Challenging the prior speaker,” *Stancetaking in discourse: Subjectivity, evaluation, interaction*, pp. 253–81, 2007.
- [197] K. Precht, “Stance moods in spoken english: Evidentiality and affect in british and american conversation,” *Text - Interdisciplinary Journal for the Study of Discourse*, vol. 23, no. 2, pp. 239–258, 2003.
- [198] P. R. White, “Beyond modality and hedging: A dialogic view of the language of intersubjective stance,” *Text - Interdisciplinary Journal for the Study of Discourse*, vol. 23, no. 2, pp. 259–284, 2003.
- [199] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, “Lexical, prosodic, and syntactic cues for dialog acts,” in *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, 1998, pp. 114–120.
- [200] P. J. Stone, *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, 1966.
- [201] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, “The development and psychometric properties of LIWC2015,” The University of Texas at Austin, Tech. Rep., 2015.
- [202] D. Biber, *Variation across speech and writing*. Cambridge University Press, 1991.
- [203] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei, “Reading tea leaves: How humans interpret topic models,” in *Neural Information Processing Systems (NIPS)*, Vancouver, 2009, pp. 288–296.
- [204] Y. Sim, B. Acree, J. H. Gross, and N. A. Smith, “Measuring ideological proportions in political speeches,” in *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, 2013.
- [205] E. Gilbert, “Widespread underprovision on reddit,” in *Proceedings of Computer-Supported Cooperative Work (CSCW)*, 2013, pp. 803–808.
- [206] D. Biber, J. Egbert, and M. Davies, “Exploring the composition of the searchable web: a corpus-based taxonomy of web registers,” *Corpora*, vol. 10, no. 1, pp. 11–45, 2015.

- [207] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in *Proceedings of the Association for Computational Linguistics (ACL)*, Madrid, Spain, 1997, pp. 174–181.
- [208] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, “Inducing domain-specific sentiment lexicons from unlabeled corpora,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.
- [209] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.
- [210] W. Ling, C. Dyer, A. W. Black, and I. Trancoso, “Two/too simple adaptations of word2vec for syntax problems,” in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Denver, CO, May 2015, pp. 1299–1304.
- [211] T. Landauer, P. W. Foltz, and D. Laham, “Introduction to latent semantic analysis,” *Discourse Processes*, vol. 25, pp. 259–284, 1998.
- [212] J. A. Bullinaria and J. P. Levy, “Extracting semantic representations from word co-occurrence statistics: A computational study,” *Behavior research methods*, vol. 39, no. 3, pp. 510–526, 2007.
- [213] Y. Niwa and Y. Nitta, “Co-occurrence vectors from corpora vs. distance vectors from dictionaries,” in *Proceedings of the International Conference on Computational Linguistics (COLING)*, Kyoto, Japan, 1994, pp. 304–309.
- [214] R. B. Cattell, “The scree test for the number of factors,” *Multivariate behavioral research*, vol. 1, no. 2, pp. 245–276, 1966.
- [215] R. P. Weber, *Basic content analysis*. Sage, 1990, no. 49.
- [216] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev, “How to analyze political attention with minimal assumptions and costs,” *American Journal of Political Science*, vol. 54, no. 1, pp. 209–228, 2010.
- [217] B. Murphy, P. P. Talukdar, and T. Mitchell, “Learning effective and interpretable semantic models using non-negative sparse embedding,” in *Proceedings of the International Conference on Computational Linguistics (COLING)*, Mumbai, India, 2012, pp. 1933–1949.
- [218] C. Callison-Burch and M. Dredze, “Creating speech and language data with amazon’s mechanical turk,” in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 1–12.
- [219] K. Krippendorff, “Computing krippendorff’s alpha reliability,” *Departmental papers (ASC)*, p. 43, 2007.

- [220] P. Eckert and S. McConnell-Ginet, “Think practically and look locally: Language and gender as community-based practice,” *Annual review of anthropology*, vol. 21, pp. 461–490, 1992.
- [221] J. Cohen, *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.
- [222] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier, “Evolution of reddit: from the front page of the internet to a self-referential community?” in *Proceedings of the 23rd international conference on world wide web*. ACM, 2014, pp. 517–522.
- [223] K. E. Anderson, “Ask me anything: what is reddit?” *Library Hi Tech News*, vol. 32, no. 5, pp. 8–11, 2015.
- [224] D. Nguyen and C. P. Rosé, “Language use as a reflection of socialization in online communities,” in *Proceedings of the Workshop on Language Analysis in Social Media*, 2011, pp. 76–85.
- [225] E. Wenger, *Communities of practice: Learning, meaning, and identity*. Cambridge university press, 1998.
- [226] D. Sankoff and S. Laberge, “Statistical dependence among successive occurrences of a variable in discourse,” *Linguistic variation: models and methods*, pp. 119–126, 1978.
- [227] M. Tamminga, “Persistence in phonological and morphological variation,” *Language Variation and Change*, vol. 28, no. 3, pp. 335–356, 2016.
- [228] L. Mayol, “An account of the variation in the rates of overt subject pronouns in romance,” *Spanish in Context*, vol. 9, no. 3, pp. 420–442, 2012.
- [229] A. Leavitt, “This is a throwaway account: Temporary technical identities and perceptions of anonymity in a massive online community,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 317–327.
- [230] M. De Choudhury and S. De, “Mental health discourse on reddit: Self-disclosure, social support, and anonymity.” in *ICWSM*, 2014.
- [231] G. Doyle, D. Yurovsky, and M. C. Frank, “A robust framework for estimating linguistic alignment in twitter conversations,” in *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 2016, pp. 637–648.
- [232] Stan Development Team, “RStan: the R interface to Stan,” 2018, r package version 2.17.3. [Online]. Available: <http://mc-stan.org/>

- [233] M. D. Hoffman and A. Gelman, “The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, 2014.
- [234] A. Vehtari, A. Gelman, and J. Gabry, “Practical bayesian model evaluation using leave-one-out cross-validation and waic,” *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.
- [235] W. Warner and J. Hirschberg, “Detecting hate speech on the world wide web,” in *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 2012, pp. 19–26.
- [236] Wikipedia, “Wikipedia,” 2018, [Online; accessed 15-April-2018]. [Online]. Available: <https://en.wikipedia.org/wiki/Wikipedia>
- [237] A. Forte and A. Bruckman, “From wikipedia to the classroom: Exploring online publication and learning,” in *Proceedings of the 7th international conference on Learning sciences*. International Society of the Learning Sciences, 2006, pp. 182–188.
- [238] O. Nov, “What motivates wikipedians?” *Communications of the ACM*, vol. 50, no. 11, pp. 60–64, 2007.
- [239] J. Schroer and G. Hertel, “Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it,” *Media Psychology*, vol. 12, no. 1, pp. 96–120, 2009.
- [240] Y. Benkler, *The wealth of networks: How social production transforms markets and freedom*. Yale University Press, 2006.
- [241] A. Cifforilli, “Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of wikipedia,” *first monday*, vol. 8, no. 12, 2003.
- [242] J. Voss, “Measuring wikipedia,” in *Proceedings of 10th International Conference of the International Society for Scientometrics and Informetrics*, 2005.
- [243] A. Lih, “Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource,” *Nature*, vol. 3, no. 1, 2004.
- [244] T. Chesney, “An empirical examination of wikipedia’s credibility,” *First Monday*, vol. 11, no. 11, 2006.
- [245] W. Emigh and S. C. Herring, “Collaborative authoring on the web: A genre analysis of online encyclopedias,” in *System Sciences, 2005. HICSS’05. Proceedings of the 38th Annual Hawaii International Conference on*. IEEE, 2005, pp. 99a–99a.

- [246] B. Keegan, D. Gergle, and N. Contractor, “Hot off the wiki: Structures and dynamics of wikipedia’s coverage of breaking news events,” *American Behavioral Scientist*, vol. 57, no. 5, pp. 595–622, 2013.
- [247] O. Arazy, F. Ortega, O. Nov, L. Yeo, and A. Balila, “Functional roles and career paths in wikipedia,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 1092–1105.
- [248] A. Forte and A. Bruckman, “Why do people write for wikipedia? incentives to contribute to open-content publishing,” *Proc. of GROUP*, vol. 5, pp. 6–9, 2005.
- [249] H.-L. Yang and C.-Y. Lai, “Motivations of wikipedia content contributors,” *Computers in human behavior*, vol. 26, no. 6, pp. 1377–1383, 2010.
- [250] S. Rafaeli and Y. Ariel, “Online motivational factors: Incentives for participation and contribution in wikipedia,” *Psychological aspects of cyberspace: Theory, research, applications*, pp. 243–267, 2008.
- [251] T. Kriplean, I. Beschastnikh, and D. W. McDonald, “Articulations of wikiwork: uncovering valued work in wikipedia through barnstars,” in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 2008, pp. 47–56.
- [252] K. Panciera, A. Halfaker, and L. Terveen, “Wikipedians are born, not made: a study of power editors on wikipedia,” in *Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 2009, pp. 51–60.
- [253] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith, “Finding social roles in wikipedia,” in *Proceedings of the 2011 iConference*. ACM, 2011, pp. 122–129.
- [254] D. Yang, A. Halfaker, R. E. Kraut, and E. H. Hovy, “Who did what: Editor role identification in wikipedia.” in *ICWSM*, 2016, pp. 446–455.
- [255] P. Shachaf and N. Hara, “Beyond vandalism: Wikipedia trolls,” *Journal of Information Science*, vol. 36, no. 3, pp. 357–370, 2010.
- [256] M. Potthast, B. Stein, and R. Gerling, “Automatic vandalism detection in wikipedia,” in *European conference on information retrieval*. Springer, 2008, pp. 663–668.
- [257] F. Flöck and M. Acosta, “Wikiwho: Precise and efficient attribution of authorship of revisioned content,” in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 843–854.
- [258] A. Elia, “An analysis of wikipedia digital writing,” in *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, 2006.

- [259] S. Greenstein and F. Zhu, “Collective intelligence and neutral point of view: the case of wikipedia,” National Bureau of Economic Research, Tech. Rep., 2012.
- [260] E. S. Callahan and S. C. Herring, “Cultural bias in wikipedia content on famous persons,” *Journal of the Association for Information Science and Technology*, vol. 62, no. 10, pp. 1899–1915, 2011.
- [261] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier, “It’s a man’s wikipedia? assessing gender inequality in an online encyclopedia.” in *ICWSM*, 2015, pp. 454–463.
- [262] O. Ferschke, I. Gurevych, and M. Rittberger, “The impact of topic bias on quality flaw prediction in wikipedia,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 721–730.
- [263] S. Das, A. Lavoie, and M. Magdon-Ismael, “Manipulation among the arbiters of collective intelligence: How wikipedia administrators mold public opinion,” *ACM Transactions on the Web (TWEB)*, vol. 10, no. 4, p. 24, 2016.
- [264] L. Herzig, A. Nunes, and B. Snir, “An annotation scheme for automated bias detection in wikipedia,” in *Proceedings of the 5th Linguistic Annotation Workshop*. Association for Computational Linguistics, 2011, pp. 47–55.
- [265] Wikipedia, “Wikipedia words to watch,” 2018, [Online; accessed 15-April-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch
- [266] K. Hyland, *Metadiscourse: Exploring interaction in writing*. A&C Black, 2005.
- [267] J. B. Hooper, *On assertive predicates*. Indiana University Linguistics Club, 1974.
- [268] B. Liu, M. Hu, and J. Cheng, “Opinion observer: analyzing and comparing opinions on the web,” in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 342–351.
- [269] P. Kiparsky and C. Kiparsky, “Fact. ed. m. bierwisch and k. heidolph,” *Progress in Linguistics*, 1970.
- [270] L. Karttunen, “Implicative verbs,” *Language*, pp. 340–358, 1971.
- [271] G. King and R. Nielsen, “Why propensity scores should not be used for matching,” *Copy at <https://gking.harvard.edu/files/gking/files/psnot.pdf>*, vol. 378, 2016.
- [272] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl, “Suggestbot: using intelligent task routing to help people find work in wikipedia,” in *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 2007, pp. 32–41.

- [273] W. Labov, *Principles of Linguistic Change*. Wiley-Blackwell, 2001, vol. 2: Social Factors.
- [274] D. Bamman, C. Dyer, and N. A. Smith, “Distributed representations of geographically situated language,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2014, pp. 828–834.