COMPUTING FOR SOCIAL SCIENCE: CHARACTERIZING, QUANTIFYING, AND ANALYZING SOCIAL PHENOMENA IN TECHNOLOGY MEDIATED COMMUNICATIONS

A Dissertation
Presented to
The Academic Faculty

by

Clayton J. Hutto

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

Georgia Institute of Technology
December, 2018

COMPUTING FOR SOCIAL SCIENCE: CHARACTERIZING, QUANTIFYING, AND ANALYZING SOCIAL PHENOMENA IN TECHNOLOGY MEDIATED COMMUNICATIONS

Approved by:

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

Dr. Erica Briscoe
Georgia Tech Research Institute
Georgia Institute of Technology

Dr. Amy Bruckman
School of Interactive Computing
Georgia Institute of Technology

Dr. Phillip Odom
Georgia Tech Research Institute
Georgia Institute of Technology

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

Dr. Scott Counts
Social Technologies Group
Microsoft Research

Date Approved: [June 20, 2018]

ACKNOWLEDGEMENTS

I would like to thank Eric Gilbert, Amy Bruckman, and Munmun de Choudhury for their continued encouragement, enthusiasm, and their resolute dedication to helping me keep on keeping on, to make something useful, to make it better, to reflect on it and learn something, and to finally finish. I especially thank Eric for his sage guidance at one of our earliest meetings; he advised me that I needed to be a social scientist with better technical chops.

I would like to thank my additional committee members Erica Briscoe, Phillip Odom, and Scott Counts for their support. I am tremendously grateful to Erica and Phillip, whose helpful feedback, insightful critiques, and probing questions were extremely valuable for prompting me to think (and re-think) about the context and content of my work. Their input helped improve the quality of this thesis, and I am thankful for their time and effort.

I thank Dennis Folds for encouraging me to start this journey in the first place, and whose mentorship, wisdom, humor, and stimulating conversation are always appreciated.

I would like to especially thank my family—Angela, Kaleigh, and Mackenzie for their love and unwavering support, for lifting my spirits, and most of all for their patience during my odyssey of good times and rough patches as I pursued a PhD while working full time in the midst of their soccer practices and games, mixed martial arts training and competitions, band concerts, theater performances, and our multitude of other family social activities. Thanks to Mom and Dad for the significant motivation and “drive”. Together, we did it!

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **iii**

LIST OF TABLES **vii**

LIST OF FIGURES **ix**

LIST OF SYMBOLS AND ABBREVIATIONS **xii**

SUMMARY **xv**

CHAPTER 1. Introduction **1**
1.1 Background and Motivation **1**
1.2 Dissertation Overview and Summary of Contributions **3**
1.2.1 Computing and Assessing Digital Predictors of Persistent Social Ties 4
1.2.2 Computing Affect Using Sentiment Analysis for Social Text 5
1.2.3 Scaled-Up Qualitative Data Analysis and Human Validation/Evaluation 7
1.2.4 Computing Bias in the News: Quantifying Bias in Sentence-level Text 8
1.3 Connections, and the Bigger Picture **10**
1.3.1 Computational Social Science: Big Picture 11
1.3.2 Human-Centered Computing: Big Picture 13
1.3.3 Bigger Picture: Human-Centered Computational Social Science 14
1.3.4 Connecting to the Bigger Picture 15

CHAPTER 2. Digital Predictors of Persistent Social Ties **17**
2.1 Chapter Overview **17**
2.2 Study Variables Informed by Social Science Theory **19**
2.2.1 Social Behavior and Follower Growth 19
2.2.2 Message Content and Follower Growth 21
2.2.3 Network Structure and Follower Growth 24
2.2.4 Limitations (and Benefits) of Longitudinal Observations 25
2.3 Dataset and Theory-Motivated Operational Definitions **26**
2.3.1 Data Collection and Reduction 26
2.3.2 Response Variable (Dependent Measure) Operational Definition 27
2.3.3 Predictor Variable Operational Definitions 27
2.4 Analysis and Discussion **32**
2.4.1 Descriptive Statistical Characteristics 32
2.4.2 Relative Prominence of the Factors Predicting Persistent Social Ties 36
2.4.3 Practical Implications 43
2.4.4 Theoretical and Methodological Implications 47
2.4.5 Study Limitations 47
2.5 Chapter Summary **48**

CHAPTER 3. Sentiment Analysis for Social Text **50**
3.1 Chapter Overview **50**

3.2 Sentiment Analysis in Computer and Social Science Scholarship 52
3.2.1 Sentiment Lexicons 52
3.2.2 Machine Learning Approaches 58
3.3 VADER Development, Validation, and Evaluation 59
3.3.1 Constructing and Validating a Valence-Aware Sentiment Lexicon: A Human-Centered Approach 60
3.3.2 Generalizable Heuristics Humans Use to Assess Sentiment Intensity in Text 64
3.3.3 Controlled Experiments to Evaluate Impact of Grammatical and Syntactical Heuristics 66
3.3.4 Ground Truth in Multiple Domain Contexts 67
3.4 Comparing VADER to Other Sentiment Analysis Benchmarks 68
3.5 Chapter Summary 74

CHAPTER 4. Large Scale Human Validation, Evaluation, and QDA 76
4.1 Chapter Overview 76
4.2 Qualitative Coding, Annotations, and Content Analysis 78
4.2.1 Crowdsourcing Qualitative Coding & Content Analysis 78
4.2.2 Crowdsourcing Data Annotations for Machine Learning 80
4.3 Strategies for Eliciting Consistently High Quality Data 81
4.3.1 Challenge 1 – Undisclosed Aptitudes 81
4.3.2 Strategy 1 – Screen Workers for Targeted Knowledge, Skills, or Abilities 82
4.3.3 Challenge 2 – Subjective Interpretation Disparity 82
4.3.4 Strategy 2 – Convergence Modeling (Provide Examples and Train Workers) 83
4.3.5 Challenge 3 –Existing Financial Incentive is to Minimize Time-on-Tasks 83
4.3.6 Strategy 3 – Financially Incentivize Workers to Focus on Quality 83
4.3.7 Challenge 4 – Low Independent (Individual) Agreement 84
4.3.8 Strategy 4 – Aggregate, Iteratively Filter, or Both 84
4.4 Qualitative Data Analysis and Annotation Tasks 85
4.4.1 Task 1: People in Pictures (PP), Median Difficulty = 1 86
4.4.2 Task 2: Sentiment Analysis (SA), Median Difficulty = 2 87
4.4.3 Task 3: Word Intrusion (WI), Median Difficulty = 2 88
4.4.4 Task 4: Credibility Assessment (CA), Median Difficulty = 3 89
4.5 Empirical Evaluation of Intervention Strategies by Tasks 90
4.5.1 Comparative Measures of "Correctness" for Subjective Judgments 91
4.5.2 Statistical Analysis Overview 92
4.5.3 Experiments 92
4.6 Analysis and Discussion 100
4.6.1 Effects of Interventions on Annotation Accuracy Increases as Subjective Judgement Difficulty Increases 101
4.6.2 Person-oriented Strategies Trump Process-oriented Strategies for Encouraging High-Quality Data Analysis and Annotations 101
4.6.3 Why Do More to Get Less? 102
4.6.4 Amazon is not a neutral observer; AMT is getting better 103
4.7 Chapter Summary 104

CHAPTER 5. Computing Bias in Journalistic News 107
5.1 Chapter Overview 107

5.2 Perceptions of Bias in Journalism 108
5.2.1 Hostile Media Bias 110
5.2.2 Manifestations of Bias in Journalism 112
5.3 The Nature of Bias: Types and Forms of Biases 115
5.3.1 Framing Effects and Biases 115
5.3.2 Epistemological Biases 119
5.4 Modeling Bias: Biased Sentence Investigator (BSI) 121
5.4.1 Lexical Level Indicators of Statement Bias 121
5.4.2 Sentence Level Indicators of Statement Bias 126
5.4.3 Coverage Bias at the Sentence Level: CASTER 129
5.5 Existing Computational Approaches for Measuring Bias 130
5.6 BSI: Preliminary Feasibility Evaluation 136
5.6.1 Dataset of Biased and Unbiased Text from News-Like Stories 136
5.6.2 Human Judgements of Bias in Unattributed News Stories 139
5.6.3 Detecting and Computing Degree of Bias in News Stories 139
5.6.4 Preliminary Evaluation Results 142
5.7 BSI: Expanded Study 144
5.7.1 Expanded Dataset 145
5.7.2 Expanded Feature Set and Feature Evaluation 149
5.7.3 Exploring Prediction Models: Linear, Non-Linear, and Machine Learning 159
5.7.4 Comparing BSI to a Parsimonious (Sentiment-Only) Model 165
5.8 Demonstration: Practical Applications of the BSI model 167
5.8.1 Statement Bias Computed at the Sentence Level of News Text 167
5.8.2 Diagnostics: Exposing the Nature of Bias in News Stories 169
5.8.3 Coverage Bias at the Sentence, Article, and Corpus Level 171
5.9 Chapter Summary 173

CHAPTER 6. Conclusions 175

REFERENCES 179

LIST OF TABLES

Table 1	Descriptive statistics for the dependent variable (follower growth) and seventeen of the twenty-two predictor and control variables (details in Section 2.3.3).	33
Table 2	Negative Binomial Auto-Regressive Model Coefficients.	37
Table 3	Example words from two of LIWC's 76 categories. These two categories can be leveraged to construct a semantic orientation-based lexicon for sentiment analysis.	53
Table 4	Example of baseline text with eight test conditions comprised of grammatical and syntactical variations.	66
Table 5	Statistics associated with grammatical and syntactical cues for sentiment intensity.	67
Table 6	VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, opinion news, product reviews).	71
Table 7	Three-class performance (F1 scores) for each machine trained model (and the corpus it was trained on) as tested against every other domain context. (Note: SVM models for the movie and NYT data were too intensive for my multicore CPU with 94GB RAM).	72
Table 8	Median subjective judgment difficulty for each task type.	86
Table 9	Combinatorial space of experiments - Four task types varying in median subjective judgment difficulty (People in Pictures, Sentiment Analysis, Word Intrusion, Credibility Assessment), two classes of Incentives (NB - No Bonus, Bonus), three types of three types of bonus incentive (M – Majority Consensus, B – BTS, C – Competition), six intervention strategies (Control, Baseline, Screen, Train, Both, Iterative Filtering). A total of 34 combinations were explored (marked ✓).	96
Table 10	χ^2 tests of independence for Experiment 1.	96
Table 11	Mean (and Standard Deviation) of 91 human-judgments of perceived bias (scale: 0=Unbiased, 1=Slightly, 2=Moderately, and 3=Extremely Biased).	140

Table 12 Brief descriptions of 26 statistical and machine learning regression models used to predict perceived bias of news articles at the sentence level. 160

Table 13 Comparison of 26 prediction models (ordered by R^2) 164

Table 14 Coefficients, error, *t-test* values, *and* p-values for a multiple linear regression using the BSI full model. F(13,954) = *15.64*, p < 2.2e-16. (Ranked by feature importance using the same ensemble regression-based metric depicted in Figure 26). 165

Table 15 BSI prediction model compared to a model based solely on sentiment. 167

Table 16 Example sentences from a Guardian news story with mean bias ratings (and standard deviations). Bias scale is continuous from 0 (neutral) to 3 (extremely biased). 168

Table 17 Feature Impact Index for each of the five sentences in the example story is calculated using a logistic (sigmoidal regularizing) function on a feature's observed value for a sentence, and then multiplying by the regression coefficient (beta). 170

Table 18 Example of using BSI (with VADER and CASTER) to analyze coverage bias at the article level. 172

LIST OF FIGURES

Figure 1	Computational Social Science	12
Figure 2	Human-Centered Computing	14
Figure 3	Human-Centered Computational Social Science	15
Figure 4	Themes and connections between dissertation chapters.	16
Figure 5	Graphical summary of analysis pipeline for longitudinal study.	32
Figure 6	For most people, negative content makes up about 20% of all tweets, while positive and neutral content each make up about 40% of tweets for most people (left). When people tweet sentiment-laden content, the intensity of positive sentiment is about three times higher than negative sentiment (right).	35
Figure 7	Standardized beta coefficients (β) show the relative effect sizes that each input variable has on follower growth. Green bars indicate positive effects on follower gain, and red bars indicate negative effects (i.e., suppression of follower growth).	38
Figure 8	Process for VADER development, validation, and evaluation.	60
Figure 9	Example of the interface implemented for acquiring valid point estimates of sentiment valence (intensity) for each context-free candidate feature comprising the VADER sentiment lexicon. A similar UI was used for all rating activities described in sections 3.3.1–3.3.4.	62
Figure 10	Sentiment scores from VADER and 11 other highly regarded sentiment analysis tools/techniques on a corpus of over 4K tweets. Although this figure specifically portrays correlation, it also helps to visually depict (and contrast) VADER's classification precision, recall, and F1 accuracy within this domain (see Table 6). Each subplot can be roughly considered as having four quadrants: true negatives (lower left), true positives (upper right), false negatives (upper left), and false positives (lower right).	70
Figure 11	Example of the subjective judgment difficulty user interface.	86
Figure 12	Example pictures for three of the five possible data annotation categories.	87
Figure 13	Example of the sentiment analysis annotation task.	88

Figure 14 Example of a topic list (with the intruder word highlighted with red text for illustration purposes). 89

Figure 15 Example of a tweet along with the five credibility coding/annotation categories modeled according to existing work on credibility annotation categories. 90

Figure 16 (Top panel) Proportion of correct responses across all tasks with respect to crowd. Pairwise comparisons which are statistically significant are shown with connecting lines (all p-values significant at 0.001 after Bonferroni correction). Effect sizes, as measured by Cramer's V coefficient, are indicated using "+" symbols at four levels: +, ++, +++, and ++++ indicate a very weak effect Cramer's V < 0.15, a weak effect Cramer's V ϵ (0.15, 0.2], a moderate effect (Cramer's V ϵ (0.2, 0.25], and moderately strong Cramer's V ϵ (0.25, 0.3], respectively. (Bottom panel) Pearson correlation between expert and crowd annotations across all tasks. 100

Figure 17 Perceptions of bias in journalism continue to be prevalent in America. 109

Figure 18 Perceptions of waning credibility in print and television news [179] reflect sensitivities to "hostile media bias" [176]. 110

Figure 19 Graphical summary of how individual features used in the Biased Sentence Investigator (BSI) computational model are drawn from theoretical underpinnings from psychology, computer-mediated communications [CMC] research, and mass media communications studies. 129

Figure 20 Proportion of variance accounted for by each feature in the improved model using the mean R2 of three regression techniques (feature added to model first, feature added to model last, and feature beta squared). 143

Figure 21 Results of 10-fold cross-validation analysis for fit between observed and predicted values of degree of bias in text. 143

Figure 22 The five news outlets selected for the extended dataset represent a range of political ideological audience preferences [180]. 146

Figure 23 The five news outlets selected for the extended dataset have sizable audiences, and are generally more trusted than untrusted by most Americans [180]. 147

Figure 24 Sentences from opinion-editorial news stories are generally percieved as being more biased than sentences from journalistic news articles. 148

Figure 25 The correlogram graphs the correlation matrix for the initial set of features in the BSI computational model; multicollinearity is not a concern. 151

Figure 26 Relative feature importance according to the proportion of variance accounted for by each feature in a multiple linear regression model using the mean of seven regression-based feature evaluation techniques. 154

Figure 27 Relative feature importance according to the ranked *p*-values from an SVM model with a polynomial kernel (degree=3). 156

Figure 28 Relative feature importance according to an ensemble of unbiased decision trees (i.e., random forest). 158

Figure 29 Graphical comparison of Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R^2 for 26 statistical and machine learning prediction algorithms 163

Figure 30 Comparison of BSI prediction model to a sentiment-only model. 167

LIST OF SYMBOLS AND ABBREVIATIONS

AMT	Amazon Mechanical Turk
ANEW	Affective Norms for English Words
AOB	Actor-Observer Bias
API	Application Programming Interface
BCART	Bootstrap-Aggregated (‘bagged’) Classification and Regression Tree
BGLM	Bayesian Generalized Linear Model
BRNN	Bayesian Regularized Neural Networks
BSI	Biased Sentence Investigator
BTS	Bayesian Truth Serum
CART	Classification and Regression Tree
CIRF	Conditional Inference Random Forest
CIT	Conditional Inference Tree
CMC	Computer Mediated Communications
CSS	Computational Social Science
EGB	Extreme Gradient Boosting
ELM	Extreme Learning Machine
ENet	Elastic Net [Regression]
FAE	Fundamental Attribution Error
FKGL	Flesch-Kincaid Grade Level
GAM	Generalized Additive Model
GI	General Inquirer
GPRBF	Gaussian Process w/ RBF Kernel

GPPK	Gaussian Process w/ Polynomial Kernel
HCC	Human-Centered Computing
HIT	Human Intelligence Task
ICR	Independent Component Regression
KNN	*k*-Nearest Neighbors
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LM	Linear (regression) Model
MARS	Multivariate Adaptive Regression Splines
ME	Maximum Entropy
MLP	Multi-Layer Perceptron
NASA-TLX	National Aeronautics and Space Administration - Task Load Index
NB	Naive Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
NN	Neural Network
PP	People in Pictures
PRF	Parallel Random Forest
QDA	Qualitative Data Analysis
RIX	Readability Index
RBF	Radial Basis Function
RF	Random Forest
RFR	Random Forest by Randomization
RRF	Regularized Random Forest

SA	Sentiment Analysis
SCN	SenticNet
SD	Standard Deviation
SIP	Social Information Processing
SVM	Support Vector Machine
SVMRBF	Support Vector Machines w/ Radial Basis Function Kernel
SVMPLY	Support Vector Machines w/ Polynomial Kernel
SWN	SentiWordNet
TGA	Tree models from Genetic Algorithms
TReDIX	Tweet Reading Difficulty Index
UAE	Ultimate Attribution Error
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary and sEntiment Reasoner
VV&E	Verification, Validation, and Evaluation
WI	Word Intrusion
WotC	Wisdom of the Crowd
WSD	Word-Sense Disambiguation

SUMMARY

Traditional social science methods of analyzing unstructured and semi-structured qualitative content often rely on labor and time intensive methods to transform qualitative data into quantitative representations of phenomena of interest. In order to rapidly conduct such social scientific research on large-scale data, social science researchers need to incorporate computational tools and methods. The Computational Social Science (CSS) paradigm offers useful perspectives for gaining insights from large-scale analyses of demographic, behavioral, social network, and technology-mediated communication data to investigate human activity, relationships, and other phenomena at multiple scales (e.g., individual, organizational, community, social group, and societal). Human-Centered Computing (HCC) complements CSS in this context by offering foundational science for designing, developing, evaluating, and deploying computational artifacts that better support the human endeavors associated with the conduct and practice of CSS research. This dissertation demonstrates theoretical, methodological, and technological contributions resulting from blending traditional social science with computational approaches for the study of human cognition and behavior. Following the CSS paradigm, I build theoretically-informed representations of social constructs—e.g., models of *interpersonal relationships* and the complex cognitive processes related to human perceptions of *sentiment* and *bias*—and use HCC methods and principles to develop and evaluate computational tools that implement those models for the purpose of aiding social science research oriented around large-scale content analysis (e.g., of content from social media networks, product and movie reviews, and newspapers).

CHAPTER 1. INTRODUCTION

In today's technology-mediated world, all sorts of human social and behavioral data are observable on previously unprecedented scales. As of June 2018, users were producing more than 8,000 tweets per second [106], and the internet in general saw more than 2.5 quintillion (2.5×10^{18}) bytes of data created each day [48]. Of course, social scientists still rely heavily on traditional sources of social and behavioral data such as in-person, telephone, or computer assisted interviews, questionnaires and survey instruments, as well as sources of "thick descriptions" [67] of human behavior compiled from ethnographic or anthropological observation research. However, new sources of human social behavior data are now available due to our increased use of mobile phone and personal wearable technology, not to mention the plethora of detailed information about human behavior available for mining from digital communications and online interactions. These data sources allow researchers to conduct human social analytics for insights ranging from investigations at intra-individual scale through inter-personal and group level interactions, to organizational and societal population scale research (c.f., [8,64,66,99,101–104,150]).

1.1 Background and Motivation

In order to be capable of rapidly conducting such social scientific research on larger scales, social scientists need to incorporate computational tools and methods. In direct fulfillment of [135]'s vision in which "…a computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth, depth, and scale…that may reveal patterns of individual and group behaviors" (pp. 721), the general theme for this dissertation is to demonstrate theoretical, methodological, and

technological contributions resulting from blending traditional social science with computational approaches for the study of human behavior – especially social phenomena as viewed through the lens of technology-mediated communications and interactions. Building on established social science theory as motivation, inspiration, and explanation, I incorporate computational and statistical data modeling techniques to be blend insights from *thick data* (most commonly qualitative in form: e.g., digital text) with the concepts of *big data* (typically more quantitative in nature and characterized by massive volume (amount of data), velocity (speed of data in or out), and variety (range of data types and sources)). For example, given the vast amount of rich, qualitative content available in social media platforms such as Twitter, Facebook, and a host of curated news, blogging and microblogging technologies, it is possible to create "social sensors" that monitor important indicators of human behavior on massive scales, in near real-time.

Unfortunately, traditional social science methods rely on labor and time intensive qualitative data analysis techniques to transform qualitative content into quantitative representations of phenomena of interest (e.g., manually reading and coding individual text entries to determine if a person is expressing positive or negative affect, or the extent to which the text may be perceived as biased) [33,194]. In contrast to most typical quantitative methods, qualitative data analysis methods do not easily scale up. Datasets are too large (consider the entire internet of social media, text messages, emails, blogs, news articles, etc.), and they are produced at extreme velocities (e.g., 500 million tweets per day, or status updates from 1.8 billion active Facebook users per day [106]). It is impossible for individual human researchers to even look at all the data, much less analyze it in a timely manner. Thus, it is evident that technological tools and techniques which help social

scientists employ their research methods on large-scales (while reducing the time and labor burdens) will be beneficial to the broader social scientific community.

Additionally, the practical costs associated with traditional interview or survey-based methods of social research usually prohibit long batteries of questions about the named discussions, with the result being that many such studies are restricted to either a small number of questions, a small number of human subjects, or both [73]. Furthermore, the direct probe approach is both intrusive (to participants) and methodologically obtrusive. This obtrusive approach has the disadvantage of being more susceptible to typical over- or under-reporting inaccuracies sometimes associated with self-reports [54], participant response bias resulting from phenomena such as social desirability [170] and researcher-induced expectancy bias [190], or observer effects whereby individuals (often unconsciously) change their behavior when they are aware of being observed (in psychology, this is also called reactivity or the Hawthorne Effect [1]). Clearly, a more unobtrusive means of discerning social phenomena of interest – and doing so on large scales without jeopardizing scientific rigor or risking researcher or participant induced biases – will also be useful to the broader social scientific community.

1.2 Dissertation Overview and Summary of Contributions

The general organization and flow of this dissertation is to first present the kinds of insights about technology-mediated social behavior that are possible when computational techniques blend with traditional social science techniques to characterize, quantify, and analyze persistent social tie formations in a popular online social network, Twitter (see Chapter 2). Next, I delve deeper into the analysis of text-based social media content by

applying human-centered methods to develop, evaluate, and deploy a computational model (called VADER) to support large scale sentiment analysis of online content from social media, news articles, and user-generated reviews of movies and products (see Chapter 3). I support the development and evaluation of VADER (and similar CSS- and HCC- inspired technology) by further refinement of a generalized crowdsource based methodological framework for conducting high-volume human evaluations/validation on large scales without jeopardizing qualitative data analysis quality (see Chapter 4). Finally, I apply the methods, tools, and techniques described above to computationally detect and quantify the degree of perceived bias in journalistic news stories. In short, this dissertation presents the confluence of social science theory building and application with human-centered development, evaluation, and deployment of computational tools to support the systematic and (unobtrusive) study of human behavior as observed via technology-mediated communications and interactions in online content. As such, I argue that this research makes substantial theoretical, methodological, and technical contributions to the fields of Human-Centered Computing and Computational Social Science, as summarized in the subsections below.

1.2.1 Computing and Assessing Digital Predictors of Persistent Social Ties

The work described in Chapter 2 is a multi-disciplinary investigation of predicting (persistent) social tie formations in online networks. Inspired by several theoretical perspectives from various social science disciplines (e.g., behavioral science & psychology, computer mediated communications [CMC], linguistics, network science/social network analysis), I answer the question of *which factors really matter for growing a social media audience*. My approach is to operationalize 22 theoretically-

motivated factors equally distributed into three categories: 1) *message content* (characteristics of the text in social communications, e.g., writing style and linguistic cues such as sentiment that expresses the tone of the message, the readability of the text, and so on), 2) *social interactions* (e.g., behavioral choices and social signals that a person uses to convey specific social impressions or expressions), and 3) attributes of the *social network structure* (e.g., network overlap/structural balance and triadic closure, network size, follower/following, follow-back reciprocity potential). I examine these 22 factors by tracking data from over 500 active Twitter users for 15 months as they collectively tweeted more than a half-million times. I observe a snapshot of each users' changing social network at regular 3-month intervals, in order a) to try to predict the change in audience size, and b) determine which factors – and which theoretical perspective(s) – are best suited to predicting the social tie connections leading to sustained audience growth on social media. The temporal nature of the longitudinal method is crucial because it more strongly suggests causal relationships between the 22 predictor variables and the dependent variable (audience growth) on Twitter. To my knowledge, this research represents the first longitudinal study of persistent social tie formation predictors on Twitter, and it is the first to show that the relative contributions of social *behavior* and message *content* are just as impactful as factors related to social network *structure* for predicting growth of online social networks. The principal contribution of the work described in Chapter 2 is social science theory building and application; moderate methodological contributions emerge as other computational social science researchers leverage many of the operational definitions presented in this work.

1.2.2 Computing Affect Using Sentiment Analysis for Social Text

The research in Chapter 3 capitalizes on an insight highlighted in Chapter 2 that there is both a strong desire and a dire need for better (i.e., more social-scientifically grounded, verified, and validated) computational tools and methods to support systematic study of human emotions, opinions, beliefs, and attitudes as presented within the digital traces of social communications – e.g., within social media message content. However, the inherent *social* nature of social media text poses serious challenges to practical applications of computational sentiment analysis. The research and technological implementation presented in Chapter 3 provides the capability to characterize both the *polarity* (e.g., positive/negative, favorable/unfavorable) and the *intensity* of sentiment expressed in digital social text. I describe the human-centered development, validation, and evaluation of VADER, a sentiment lexicon and parsimonious rule-based computational model for general sentiment analysis.

VADER ("Valence Aware Dictionary and sEntiment Reasoner") is intended to specifically address the challenge of computationally assessing sentiment in social media communications. The research approach unambiguously begins, iteratively integrates, and ends with a host of human-centric methods. The process combines large-scale qualitative content analyses with empirical evaluations (human-subject validation and experimental investigations by leveraging a wisdom-of-the-crowd[1] (WotC) approach [208]), and by incorporating established natural language processing (NLP) techniques. I then compare VADER's lexicon effectiveness to eleven typical state-of-practice benchmark lexicons including Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words

[1] *Wisdom-of-the-crowd* is the process of incorporating aggregated opinions from a collection of individuals to answer a question. The process has been found to be as good as (often better than) estimates from lone individuals, even experts.

(ANEW), General Inquirer, SentiWordNet, and machine learning oriented techniques relying on Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms to produce domain-specific lexicons from sample data. VADER outperforms individual human raters (*F1-Score* = 0.96 and 0.84, respectively), and generalizes more favorably across contexts than any other benchmark. Contributions of the work described in Chapter 3 are principally methodological and technological in nature; VADER provides a foundational building block for computational social science research efforts interested in unobtrusively characterizing the attitudes, opinions, belief expressions, or biases presented in text-based technology-mediated social communications and online content.

1.2.3 Scaled-Up Qualitative Data Analysis and Human Validation/Evaluation

In Chapter 4, I describe a quality-assurance oriented framework for ensuring high value subjective data collection from crowdsourced micro-labor markets. Motivation for this work comes from the research in Chapters 3 and 5, which both rely heavily on Amazon Mechanical Turk (AMT) to provide human-centered verification, validation, and evaluations (VV&E) of computational models, and for rapid qualitative data analysis (QDA) of textual content on large scales without sacrificing analysis quality. The availability of a massive, distributed, transient, anonymous crowd of non-expert individuals willing to perform general human-intelligence micro-tasks for micro-payments is a valuable resource for researchers and practitioners, and has dramatically influenced large scale social science research. However, the very nature of *massive*, *distributed*, *non-expert*, *transient*, and *anonymous* crowds – with associated variances in individual differences of knowledge, skills, aptitudes, and motivations – presents a challenge for obtaining consistent QDA results as well as concerns about low quality analysis. Due to

the somewhat specialized and subjective nature of many qualitative data analysis activities, I develop and test a "person-centric" framework comprised of a collection of strategies that facilitates quality assurance for research-worthy data collection via Amazon Mechanical Turk. I then compare those person-centric strategies to an alternative framework comprised of a collection of "process-centric" strategies for obtaining quality data via AMT. Results point to the advantages of person-oriented strategies over process-oriented strategies. Specifically, I demonstrate that prescreening workers for requisite aptitudes and providing rudimentary training in collaborative qualitative data coding techniques is quite effective, significantly outperforming control and baseline conditions. Interestingly, such strategies can improve qualitative coder annotation accuracy above and beyond common (and more complex) benchmark strategies such as Bayesian Truth Serum (BTS). Using these person-centric strategies results in improved human-produced verification, validation, and qualitative analyses described in Chapter 3 (VADER sentiment analysis tool development) and Chapter 5 (biased statement detection and quantification). Thus, the principal contribution of the research described in Chapter 4 is a generalized methodological framework for obtaining consistent computational model VV&E and high quality QDA via the wisdom of the crowd (WotC, c.f., [208]) for social science research.

1.2.4 Computing Bias in the News: Quantifying Bias in Sentence-level Text

The research described in Chapter 5 incorporates the methods, tools, and techniques from Chapters 2-4, and leverages them for applied research related to computationally detecting and quantifying the degree of bias in sentence-level text of journalistic news stories. Fair and impartial reporting is a prerequisite for objective journalism; the public holds faith in the idea that the journalists we look to for insights about the world around us

are presenting nothing more than neutral, unprejudiced facts. Most news organizations strictly separate *journalistic news* and *editorial* staffs. Bias is, unfortunately, nevertheless ubiquitous in journalism. It is therefore at once both intellectually fundamental to understand the nature of bias and pragmatically valuable to be able to conduct rapid initial review of news stories for the presence of bias. To this end, I construct a computational model to detect bias when it is expressed in news reports and to quantify the intensity of the biased expression. Using the methods described in Chapter 4, human judges provided ground-truth gold standard ratings for the degree of perceived bias (slightly, moderately, or extremely biased) for every sentence across 105 separate news articles to help investigate the factors that influence the perception of bias in real as well as representative (albeit fictitious) news stories. In a preliminary pilot study, I analyze a combination of text-based structural and linguistic information for not only detecting the presence of biased text, but also to construct a model capable of estimating its magnitude. I compare and contrast common linguistic and structural cues of biased language, to develop an initial computational model with greater than 97% accuracy, and accounted for 85.9% of the variance in human judgements of perceived bias in news-like text for a very small dataset comprised of sentences from five news-like stories. Expanding on this initial feasibility study, I further develop a theory-informed computational model called the Biased Sentence Investigator (BSI) that implements a total of 32 measures hierarchically organized into 13 categories. These include sentence-level measures such as *sentiment* and *certainty* as well as lexical-level measures such as *presupposition* language markers (which reflect epistemological bias and presupposed truths), and *value-*, *partisan-*, and *figurative-* language markers (which reflect a blend of biases arising from the framing effects

associated with certain rhetorical devices) to name a few. I next compare 26 different statistical and machine learning regression models using the BSI features to predict the perceived bias of sentences in an annotated dataset of news articles. Implementations range from multiple variations on linear regression models to more complex nonlinear, non-parametric regressions, decision trees, random forests, neural networks, and support vector machines. Extensive feature and model evaluations show that performance of the BSI model and selected features compare favorably to human performance for matching the average perceived bias rating for sentences in real world news stories (for example, the mean Pearson Correlation Coefficient was r=0.565 for BSI using Regularized Random Forest machine learning, compared to r=0.661 for human judges). Finally, I demonstrate the BSI capabilities for investigating statement bias and coverage bias at the sentence and article units of analysis. The principal contributions of the work presented in this chapter are: a) demonstrable application of computational social science methods, tools, and techniques developed in Chapters 2-4 for social science theory building and understanding of bias in journalistic text, and b) technological implementation of a tool capable of rapidly assessing the presence and computing the degree of bias in journalistic news stories.

1.3 Connections, and the Bigger Picture

As stand-alone efforts, the projects and studies discussed in Chapters 2-5 represent very strong theoretical, methodological, and technological contributions. But, how do they relate to each other (especially given that they cover such seemingly disparate research topics, each with their own unique underlying theories and data), and how are they situated within the broader context of Human-Centered Computing and Computational Social Science? To answer the second of these two questions, an understanding of HCC and CSS

would be useful. While a full literature review of either paradigm in its entirety is far beyond the scope of this dissertation, a brief introduction and explanation that provides context for this dissertation is appropriate.

1.3.1 Computational Social Science: Big Picture

In 2009, Lazer and colleagues noted that "…a computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth, depth, and scale…that may reveal patterns of individual and group behaviors" [135]. Five years later, Cioffi-Revilla published the first textbook on the subject, stating that "The new field of Computational Social Science can be defined as the interdisciplinary investigation of the social universe of many scales, ranging from individual actors to the largest groupings, through the medium of computation" [33]. Together, these two definitions reveal a few key aspects of CSS: the first is on recognizing a vast new world of *(human-centered) data* at multiple scales and across time; the second is on *computation* as a means to facilitate collection and analysis of this small- to large-scale data; the third is that such analysis is for the purpose of identifying patterns and working towards a *quantitative understanding* of complex social systems in our social universe.

These three aspects help mitigate some longstanding difficulties for traditional social science. First, with regards to data, many traditional methods from social science are oriented around data collected from surveys, interviews, researcher observations, lab experiments, and (manual) labor and time intensive qualitative data analysis. However, the ever increasing integration of Information and Communication Technology (ICT) into our lives has created unprecedented volumes of data on society's everyday behavior. The

resulting rise of big Human-Centered Data [9] represents exiting new opportunities for social scientists to "observe" complex social systems in a planetary scale "natural lab" [36]. Through computation, CSS facilitates social science work by enhancing the capacity to access, collect, process, and store the data (e.g., via data mining, natural language processing, and other tools for automated data extraction). Second, with regards to analysis, computational algorithms and models helps to formally characterize, operationalize, and otherwise quantify social science concepts and constructs representing patterns of human cognition and behavior ranging from individual decision making to internet scale social networks and communications. Third, with regards to working towards a *quantitative understanding* of complex social systems, CSS improves on traditional social science via experiments and investigations on larger scales, longer time horizons, with greater complexity and realism—either by deploying these algorithms and models in ethical, safe, economical "virtual labs" (e.g., with simulations) or by cyclically feeding them back into the broader ICT "natural lab" for additional data collection or scientific investigation. Figure 1 graphically summaries the above concepts for CSS:

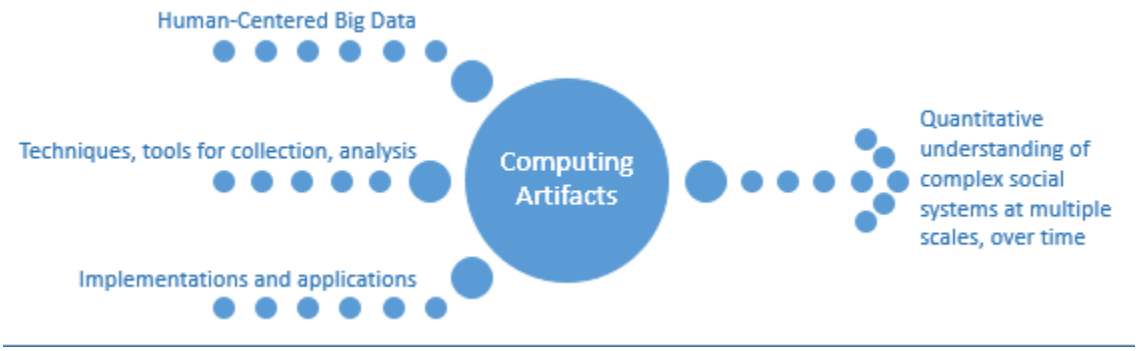


Figure 1: Computational Social Science

1.3.2 Human-Centered Computing: Big Picture

The field of Human-Centered Computing emerged from the convergence of multiple disciplines concerned with both a) understanding humans and b) the analysis, creation, and evaluation of computational artifacts. Each term of the title is important: *Human-Centered* reflects the prominence of human beings (at multiple scales, e.g., individual, group, team, organization, community, or society) as the central focus during the creation and use of technology artifacts; *Computing* reflects the emphasis on computational technology (as opposed to other forms of technology, such as Norman's "everyday things" like teapots and door handles [160], or Bjiker's bicycles, Bakelite, and bulbs [15]). The full title *Human-Centered Computing* reflects a systems view that posits humans and computing artifacts should be considered together as a holistic unit, and that such systems are themselves situated within multi-scaled (from hyper-local to global) contexts composed of physical (or virtual), social, cultural, ethical, economical, and societal systems. This sociotechnical system-of-systems perspective also connotes the idea that societies and technologies co-evolve, influencing and changing each other in their respective evolution processes. At its core, HCC research is focused "on how humans, in various roles and domains, perceive computing artifacts as they design and use them, and on the wider social implications of those artifacts" [157]. In the creation (design and production) of computational artifacts, HCC incorporates computer science as informed by cognitive/behavioral/social psychology, sociology, ethnography, anthropology, design science, human factors, cognitive science, linguistics, communication and media studies, political science, science-technology-society (STS) studies, information science, and other

related fields. Likewise, many of these same disciplines provide the basis for evaluations of those computational artifacts. Figure 2 graphically summarizes the above concepts:

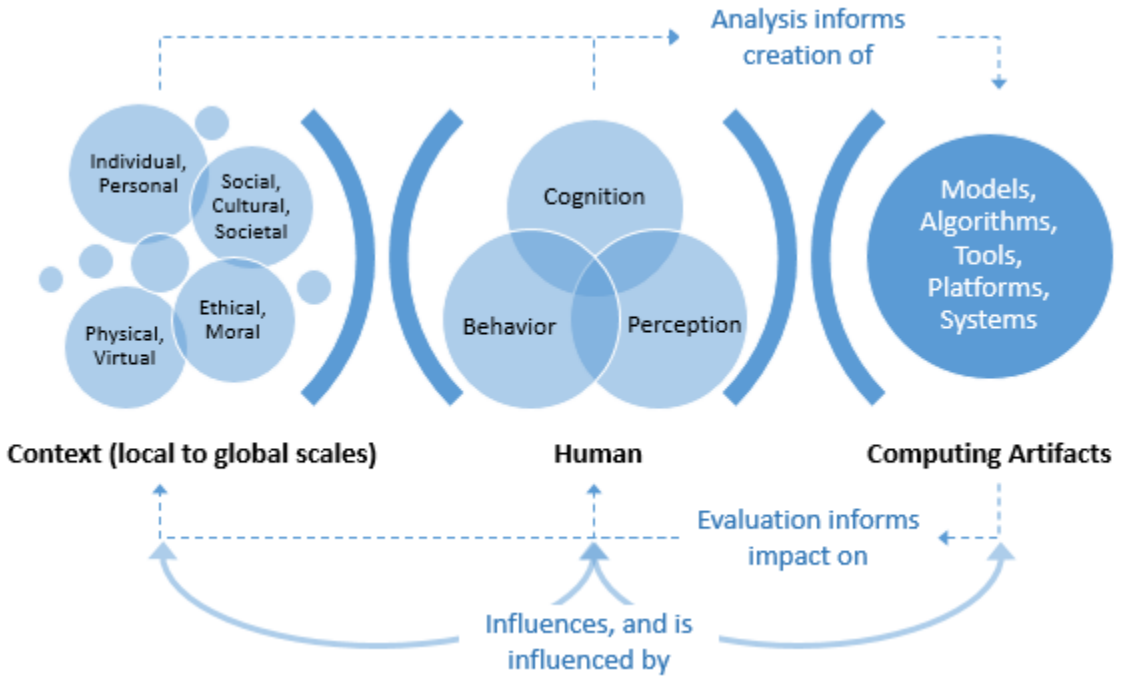


Figure 2: Human-Centered Computing

1.3.3 Bigger Picture: Human-Centered Computational Social Science

The Computational Social Science (CSS) paradigm offers a useful perspective for gaining insights from large-scale analyses of demographic, behavioral, social network, technology-mediated communications, and other online content to investigate human activity, relationships, and social phenomena at multiple scales (e.g., individual, organizational, community, social group, and societal) and over time [32,33,36,135]. In this way, HCC complements CSS by offering foundational science for analyzing, creating, and evaluating computational artifacts that better support the human endeavors associated with the conduct and practice of social science research (see Figure 3).

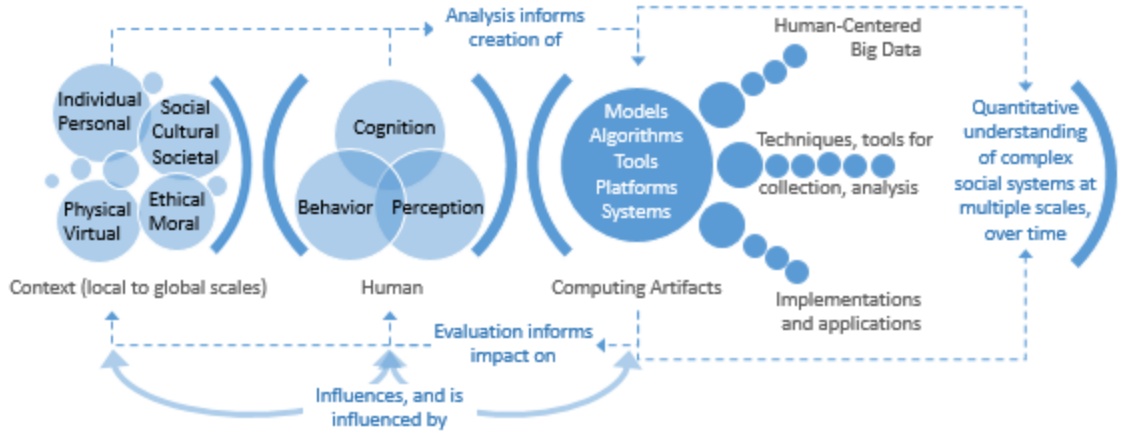


Figure 3: Human-Centered Computational Social Science

1.3.4 Connecting to the Bigger Picture

In this dissertation, the computational artifacts being created are directly in the service of larger-scale social science research—i.e., computational models to predict persistent digital social ties, sentiment and bias, as well as a crowdsourced (human computation) method to support the design, development, evaluation, and validation of the computer models (see Figure 4). In every case, these computational artifacts are:

a) informed by human-centered methods and established social science theories,

b) leveraging human-centered data from Technology-Mediated Communications (TMC),

c) for the purpose of aiding analysis of TMC content at larger scales, over time, and

d) intended to contribute to a better understanding of the broader social implications of TMC use, as well as the co-evolution of societies and technologies.

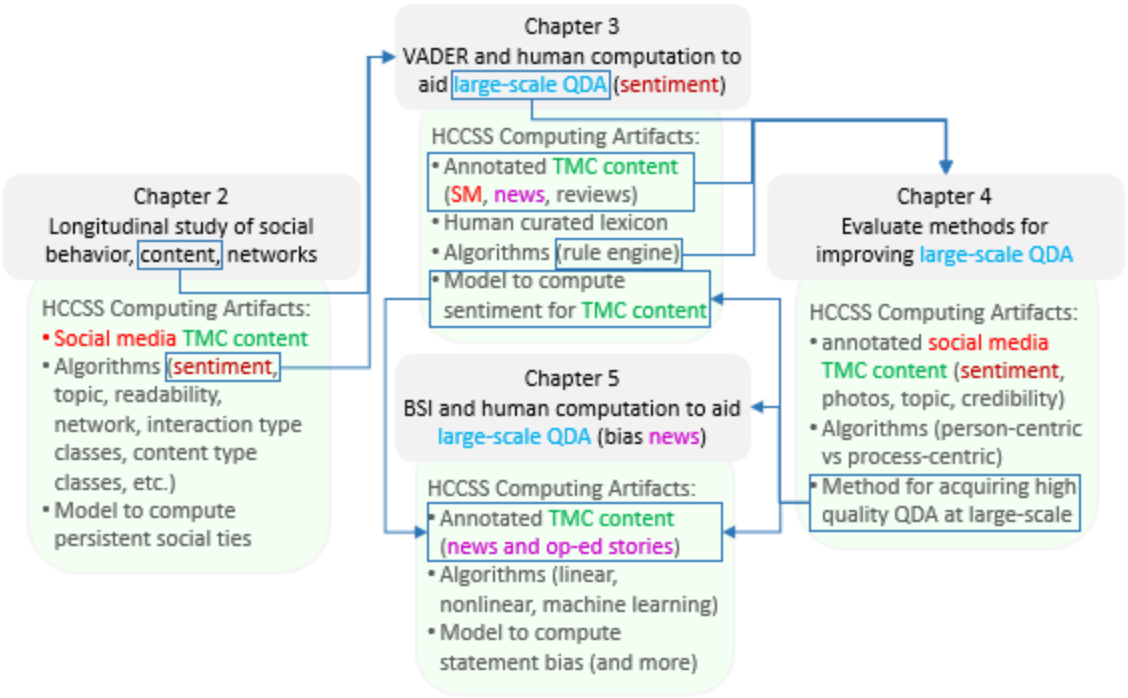


Figure 4: Themes and connections between dissertation chapters.

CHAPTER 2. DIGITAL PREDICTORS OF PERSISTENT SOCIAL TIES

2.1 Chapter Overview

Followers are Twitter's most basic currency. Building an audience of followers can create access to a network of social ties, resources, prestige and influence. Yet, little is understood about how to grow such an audience. This chapter examines multiple factors that affect persistent tie formation and dissolution over time on the social media service Twitter. For this work, I collected behavioral, content, and network data approximately every three months for fifteen months. I examined specific user *social behavior choices* (i.e., communication and interactions) such as: proportions of directed communications versus broadcast communications [21]; the total number of tweets produced; communication burstiness; and profile completeness [132]. I also assessed numerous attributes specific to the *content* of users' social media messages (i.e., tweets), such as: propensity to express positive versus negative sentiment [123,184]; topical focus [215]; proportions of tweets with "meformer" content (i.e., content written by users about themselves) versus informative content [154]; frequency of others "retweeting" a user's content [19]; linguistic sophistication (reading difficulty) of tweets; and hashtag usage. Finally, I evaluated the impact of users' evolving *social network structure*, collecting snapshots of their friends and followers every three months for fifteen months. With this data, I evaluated the effects of network status, reciprocity [75], and common network neighbors.

The above variables were selected from prominent theoretical constructs bridging social science, linguistics, computer mediated communications, and network theory. This chapter compares the relative contributions of factors from each perspective for predicting persistent social tie formations in online social networks. I take a temporal perspective and develop a model that accounts for social behavior, message content, and network elements at several intervals for over a year. I evaluate this longitudinal approach via a negative binomial auto-regression model to explore the changes in users' follower counts over time. I find that *message content significantly impacts follower growth*. For example, [123] observed static snapshots of social networks—rather than a longitudinal view of the evolving networks—and observed that sharing negative emotions correlated with higher numbers of followers. In contrast to [123], I find that expressing negative sentiment has an adverse effect on follower gain, whereas expressing positive sentiment helps to facilitate it. Similarly, I show that informative content attracts new followers with a relative impact that is roughly *thirty times higher* than the impact of "meformer" content, which deters growth. I also find that *behavioral choices can also dramatically affect follower growth*. For example, choosing to complete one's profile and choosing directed communication strategies over broadcast strategies significantly stimulates follower growth over time. Finally, I show that *even simple measures of topology and structure are useful predictors of evolutionary network growth*. I close the chapter with practical and theoretical implications for designing social media technologies.

Comparing across multiple variables related to message content, social behavior, and network structure allows me to interpret their relative effect on follower growth from different theoretical perspectives, helping to build greater understanding of the underlying

social theory and insights for its application. This is the first research to compare the impact of all these factors together within a single longitudinal study of social media users. The temporal nature of the longitudinal method is crucial because it more strongly suggests causal relationships between these factors and persistent social tie formation on Twitter.

2.2 Study Variables Informed by Social Science Theory

In this section, I consider established research showing how social behavior, message content, and network structure relate to follower growth. The current study draws from this prior work in deciding which variables to include in the analysis, and contributes new results to the body of literature by considering these variables temporally, and in conjunction with one another. For convenience and organizational purposes, I group these variables into three categories: social behaviors (e.g., interactional communication choices that a user makes), message content (e.g., linguistic cues from text), and social network structure. These categories are intended to be neither mutually exclusive nor exhaustive. However, I specifically call attention to variables related to *message content* because they seem to be underrepresented in much of the related literature on follower growth dynamics [75,81,123,130,136], and because they help shape the research challenges addressed in subsequent chapters of this dissertation.

2.2.1 Social Behavior and Follower Growth

2.2.1.1 Social Capital and Communication Behavior

Social capital refers to "the actual or potential resources which are linked to a durable network of more or less institutionalized relationships of mutual acquaintance or

recognition" [18]. It is your relative social "worth," resulting from your position in a social network: i.e., the number and kind of the ties you maintain, your relative access to resources desired by those in your network, as well as your level of access to the resources your network ties possess [217]. In prior work, researchers distinguished between three kinds of social behavior that affect social capital on the social networking site, Facebook: (1) directed communications with specific, target individuals; (2) broadcast communications, which are not targeted at anyone in particular; and (3) passive consumption of content [21]. Because personalized messages are more likely to contain content that strengthens social relationships (such as self-disclosure and general supportiveness), it has been suggested that directed communications are useful for maintaining existing ties and for encouraging the growth of new ones. Indeed, previous research found that, when compared to broadcast communications and passive consumption, personalized one-on-one communication strategies have a measurably greater impact on self-reported social capital of Facebook users [21]. Other research suggests that informal personal conversation is a major reason for using a social media like Twitter [92,109], even for work and enterprise purposes [231,232]. However, the volume of messages and the rate at which they are transmitted (i.e., their "burstiness") are both correlated with unfollowing on Twitter [130]. In the current research effort, I test whether these behaviors help to grow persistent social ties on Twitter.

2.2.1.2 Profile Elements as Social Signals

Because there is some effort incurred with producing it, user-generated profile content is an important signal for conveying a trustworthy identity [49,50,132]. The shared context of social networking sites like Facebook and Twitter helps facilitate explicit and

implicit verification of identity claims, and users are motivated to present their "ideal self" [74] in order to attract new connections. In [132], the authors explore the relationship between profile structure (namely, which fields are completed) and number of friends on Facebook. Based on a static snapshot of the social network at a large university, the authors found that the act of populating profile fields was strongly correlated with the number of friendship links. Compared to users without profile elements, users who had entered profile content had about two to three times as many friends. Based on this established literature as well as my own intuition, I anticipate similar effects in the longitudinal data regarding network growth on Twitter. Assuming that people will be more likely to follow those who include identity cues in their profile (such as description, location, and personalized URL), I expect that the more these elements are included, the more successful one will be in growing an audience. The research described in this chapter tests these assumptions.

2.2.2 Message Content and Follower Growth

2.2.2.1 Sentiment and Emotional Language

Sentiment analysis refers to the computational treatment of opinion, sentiment, and subjectivity in text [163]. Previous research found significant correlations between the number of followers of a Twitter user and that user's tendency to express emotions like joy and sadness [123] or positive versus negative sentiments [184] in their tweets. However, the authors in [123] acknowledge that an important limitation of the study was the static nature of the correlation analysis. In particular, note the following passage from the paper:

> With the current analysis we cannot deduce causality; e.g., whether the emotional richness of interactions draws more

> followers or whether people tend to share more emotional content when they have larger audiences. (p. 382)

Although not explicitly stated, this same limitation also applies to [184]. I build on their prior work and extend it by studying changes in audiences *over time*. By relying on time-dependent regression analysis of longitudinal data to identify the relative effects of sentiment expression on follower gain, I mitigate the limitation noted above. This is conceptually similar to the approach used by [85] to characterize the relative effects of various factors on predicting Twitter adoption among young adults. Exploring dynamics over time provides a stronger case for causality.

I also build on the approach in both [123] and [184] by improving upon the LIWC2007 text analysis package to automatically classify positive and negative sentiment. LIWC [175] is a widely used and validated dictionary-based coding system often used to characterize texts by counting the frequency of more than 4,400 words in over 70 categories. However, LIWC does not include many features that are important for sentiment analysis of tweets. For example, the work in this chapter incorporates the 905 words in LIWC categories for P*ositive Emotion* and *Negative Emotion,* plus an additional ~2,200 words with positive or negative sentiment[2], as well as additional considerations for sentiment-laden acronyms/initialisms, emoticons, slang, and the impact of negations. These supplementary characteristics are known to be important features of sentiment analysis for microblogs like Twitter [42]. Also, some words connote more extreme sentiment than others (e.g., "good" versus "exceptional"). Thus, in addition to simply

[2] http://fnielsen.posterous.com/afinn-a-new-word-list-for-sentiment-analysis

counting occurrences of positive or negative words (i.e., the LIWC method), I also assess the directional magnitude (i.e., *intensity*) of the sentiment for each word, associating human coded valence scores ranging from -5 to +5 for each word in the dictionary. The above summary explanation provides sufficient context needed for the current chapter; I further explore the human-centered development and validation of my computational tool for sentiment analysis in much greater detail in Chapters 3, with accompanying methodological framework described in Chapter 4.

2.2.2.2 Topical Focus

The principle of *homophily* asserts that similarity engenders stronger potential for interpersonal connections. In the selection of social relationships, people tend to form ties to others who are like them – a finding that has been one of the most pervasive empirical regularities of modern social science [149]. Sharing interests with another person is one form of similarity [60]. A Twitter user who discusses a wide range of topics may appeal to a broader audience, therefore attracting more followers – a notion that, according to [215], is supported by the economic theory of network externalities [116,187]. In [215], the authors describe how initial topical focus affected users' ability to attract followers. However, the users in [215] self-identified as providers of *politically oriented* tweets, and it is unknown whether the findings from [215] will hold for a more heterogeneous sample of Twitter users. The research described in this chapter also addresses this uncertainty.

2.2.2.3 Informativeness: Information Brokering and "Meformers"

In [131], the authors highlight the dual nature of Twitter as both a social network and as a news/information medium. Also, [154] suggests two basic categorizations of

Twitter users as Informers (those who share informative content) versus "Meformers" (those who share content about themselves). Meformers were reported to have almost three times fewer followers than Informers [154]; but, the authors note that "the direction of the causal relationship between information sharing behavior and extended social activity is not clear". My work here explores whether this type of message content affects a person's ability to attract, acquire, and retain the persistent social ties needed for growing a social media audience over time.

2.2.3 Network Structure and Follower Growth

2.2.3.1 Network Size, Reciprocity and Mutuality

Preferential attachment, or the phenomenon whereby new network members prefer to make a connection to popular existing members, is a common property of real life social networks [12] and is useful for predicting the formation of new connections [136]. The number of followers a person maintains has been shown to reduce the likelihood that the person will be unfollowed in the future [122], meaning popular people often remain popular. Additionally, one can calculate the "attention status" of an individual within their own Twitter network by taking the ratio of followers (those who pay attention to the user) to following (those among whom the user divides their attention). Such measures reflect ego-level network attributes that affect the decision of others to follow the user. On the other hand, [75] shows that follower counts alone do not fully explain interest in following. In other words, popularity, in and of itself, does not beget popularity. Dyadic properties such as *reciprocity* and *mutuality* also play key roles in the process of tie formation and dissolution [75,122].

2.2.3.2 Common Neighbors: Structural Balance and Triadic Closure

In addition to dyadic structural network properties, I also consider *triads* (structures of three individuals). Specifically, I am interested in the concepts of *structural balance* and *triadic closure*. For example, consider the case where three people form an undirected network. If A is friends with X, and X is friends with B, then according to Heider's theory of cognitive balance, the triad is "balanced" when A is also friends with B, but "unbalanced" when A is not friends with B [88]. As the number of common neighbors (occurrences of "X") between A and B increases, the likelihood of the A-B tie being formed also increases [28]. This principle of structural proximity is known as *triadic closure* [55]. Measuring the occurrences of common network neighbors is useful for link predictions in real life social networks [136] as well as online social networks [75,81,122]. I explore the extent to which such network structures impact persistent social tie formation, and compare the impact of network structural features to the impacts of features related to message content and social behavior.

2.2.4 Limitations (and Benefits) of Longitudinal Observations

Making causal claims with observational data can be problematic. It is impossible to absolutely rule out every possible "third factor" that might account for some portion of an association between an independent variable and its effect on the dependent variable. I attempt to mitigate this problem by accounting for as many "third factors" as is feasible, and considering them all in conjunction with one another. Longitudinal studies are still correlational research, but such correlations have greater power because of time-dependent, repeated observations. In other words, when input A is consistently and reliably observed

preceding outcome B for the exact same group of individuals time after time, then one has greater confidence in suggesting a causal relationship between A and B.

2.3 Dataset and Theory-Motivated Operational Definitions

2.3.1 Data Collection and Reduction

I collected data from 507 active Twitter users who collectively provided a corpus of 522,368 tweets spanning 15 months. In addition to the tweets, I also have snapshots of friends and followers taken at periodic intervals (a total of five periods, each approximately three months in duration). I am interested in discovering the relationship between the factors discussed above within each three-month period and the subsequent changes in follower counts at the end of that period. To build the dataset, Twitter accounts were obtained by recording unique account IDs that appeared on the public timeline during a two-week period preceding full data collection, and then screened for certain attributes. The subset selected for inclusion in this study consisted of those accounts that met the following four criteria when sampled approximately every three months:

1. Tweet in English, as determined by inspecting the users' profiles for the designated language via Tweepy[3], a Twitter API library for Python, as well as Python's Natural Language Tool Kit[4] (NLTK) for language detection on the users' 20 most recent tweets. This filter is necessary for the linguistic predictors (described later), although it may restrict the generalizability of the results.

[3] http://code.google.com/p/tweepy
[4] http://www.nltk.org

2. Have Twitter accounts that are at least 30 days old at the time of the first collection period, and are therefore not new to the service. This was done to avoid the potential confounding effects of users who have just joined and are likely building up their followership based on existing friends and acquaintances (rather than attracting followers based on the variables I track).

3. Follow at least fifteen other "friends" and have at least five followers. This removes a large portion of unengaged or novice users, and is close to Twitter's own definition of an "active user"[5,6] at that time.

4. Tweet at least twenty times within each time period (a *time period* is the approximately three-month interval between snapshots of users' social networks; twenty tweets in three months is not quite two tweets each week). This removes the confounding effects of inactive accounts, and ensures data is available for this analysis.

2.3.2 Response Variable (Dependent Measure) Operational Definition

Follower growth: change in follower counts for users at the end of a given three-month time period, as compared to the follower counts at the end of the previous period.

2.3.3 Predictor Variable Operational Definitions

2.3.3.1 Behavioral and Social Interaction Variables

[5] http://www.businessinsider.com/chart-of-the-day-how-many-users-does-twitter-really-have-2011-3

[6] http://techland.time.com/2011/09/09/twitter-reveals-active-user-number-how-many-actually-say-something

Tweets in period: the total number of tweets produced by a user in a three-month time period.

Peak tweets per hour ("burstiness"): for a given three-month time period, the maximum rate of tweets per hour.

Directed communications index: captures replies and mentions, as well as consideration for the social signal sent when the person "favorites" someone else's tweet, calculated as "@" count plus favorites count divided by the total number of tweets in a period.

Broadcast communication index: the ratio of tweets with no "@" at all in the tweet to total number of tweets in a period.

Profile cues of "trustworthiness" of Twitter identity: (1) the length, in characters, of the user's self-defined profile description, (2) whether the user has indicated a personal URL in their profile, and (3) whether the user has indicated their location. I collected data about whether a user had a personal profile image or the default image, but there was insufficient variation in the data to use it (all users in the sample had non-default images).

2.3.3.2 Message Content Variables

Positive (Negative) sentiment intensity rate: ratio of the sum of the valence intensity of positive (negative) language used in tweets to the total number of tweets in a period. In a separate formative evaluation involving a small subset of tweets from the corpus (n=300), my custom sentiment analysis engine performed quite well. The correlation coefficient between my sentiment analysis engine and ratings from three human judges was high (r =

0.702); better than the Pattern.en sentiment analysis engine[7] (r = 0.568). The correlation among human judges was r = 0.851. I further refine, enhance, and improve upon this initial sentiment analysis engine in a subsequent research effort (see Chapter 3).

Informative content index: the ratio of tweets containing either a URL, "RT", "MT", "HT" or "via" to total number of tweets in the period.

Meformative content index: the ratio of tweets containing any of the 24 self-referencing pronouns identified in LIWC (e.g., words like "I", "me", "my", "we", "us") to total number of tweets in the period.

Topic focus: following [215], this is the average cosine similarity (ranging between 0 and 1) for every unique paired combination of a user's tweets in a given time period.

User tweets retweeted ratio: the total number of times a user's tweets were retweeted, relative to the total number of tweets produced by the user in the period.

Hashtag usage ratio: the total number of hashtags used in a period relative to the total number of tweets in the period.

TReDIX: the "Tweet Reading Difficulty Index" is a measure I developed to capture the linguistic sophistication of a set of tweets. It is inspired by the Readability Index (RIX, c.f. [6]) and is based on the frequency of real English words with 7 or more letters. TReDIX is a ratio of the total count of long words appearing in tweets within a time period relative to the number of tweets in the period.

[7]http://www.clips.ua.ac.be/pages/pattern-en#sentiment

2.3.3.3 Network Topology/Structural Variables

In-link reciprocity rate: the number of followers that the user is also following relative to the total number of followers in the user's social network for each time period.

Attention-status ratio: ratio of followers (those who pay attention to the user) to following (those among whom the user divides their attention), calculated based on the user's existing social network at the end of each period.

Network overlap: where A is the user of interest and B is either a follower or a friend of A, this is the raw network overlap (count of common neighbors) between A and B. The final measure is the sum for user A's entire network.

2.3.3.4 Other (Control) Variables

Age of account: the age of a user's Twitter account (in days) at the end of a time period, to control for the likely differences between older, more established accounts and newer, developing accounts.

No. of followers: The total number of followers at the end of a given period, a plausible criterion used by other potential followers when evaluating whether or not to follow the user. I include the number of followers as a control to account for popularity-based preferential attachments.

No. of friends ("followees"): The number of accounts the user is following at the end of a given period, also a plausible criterion used by potential followers when deciding whether to follow a user.

Change in followers (previous period): change in follower count at the end of time period t_{-1} (the previous time period), is a lagged variable used to control for second order follower growth dynamics for the dependent variable in the time-dependent auto-regressive model. This addresses the issue of possible preferential (de)attachment for rising or falling "stars" [12], and helps mitigate concerns related to lack of independence among repeated observations.

I test the predictive power of these variables by incorporating auto-regression into a negative binomial regression model. Negative binomial regression is used for modeling count variables, and is well-suited to modeling dependent variables of count data which are ill-dispersed (either under- or over- dispersed) and do not have an excessive number of zeros [25], as is the case with this dataset. Auto-regressive models attempt to predict an output of a system based on previous observations [161], which mitigates concerns associated with lack of independence for repeated measures by incorporating a lagged variable into the statistical model. In the present study, I use auto-regression to account for the overall slope of follower gain heading into a given time period. Change in follower growth at the end of time period t_0 is therefore conditioned upon the change in follower growth at the end of t_{-1} (the previous time period). After removing tweets from the first time period interval (it only provides the initial baseline of counts from which I derive changes in follower growth for subsequent periods) and the second time period (in order to incorporate dependency on change in growth for the auto-regressive model), I have 507 unique active Twitter users who collectively provided 1,836 instances of follower growth across the remaining four time periods of the longitudinal analysis. Figure 5 graphically summarizes the computational modeling and analysis pipeline:

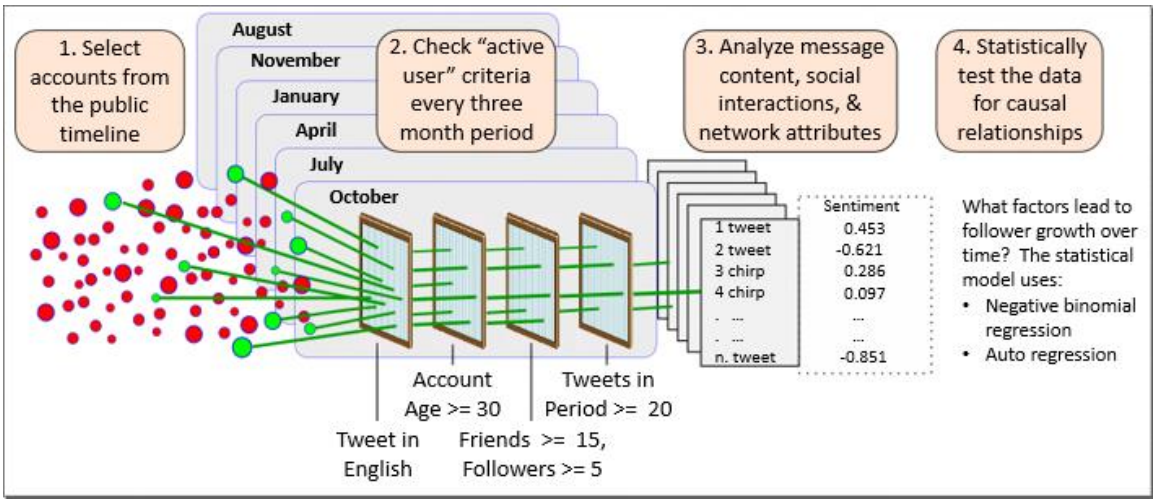


Figure 5: Graphical summary of analysis pipeline for longitudinal study.

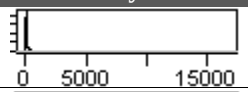
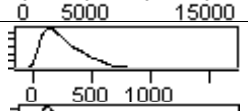
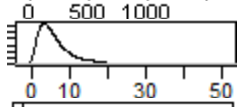
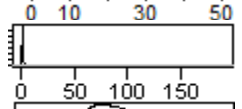
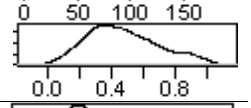
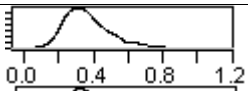
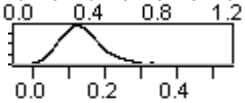
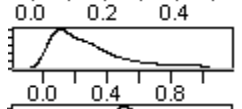
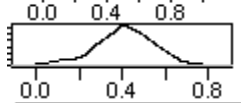
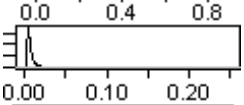
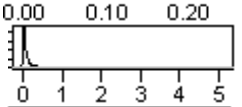
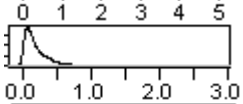
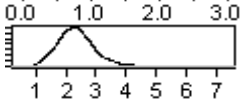
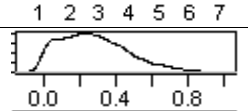
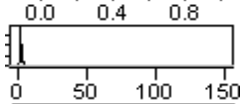
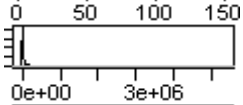
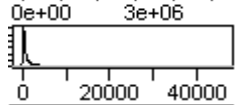
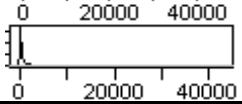
2.4 Analysis and Discussion

I first present descriptive statistics for the dependent measure (follower growth) and the twenty-two predictor and control variables. I organize these variables into three convenience categories: behavioral/social interaction, message content, and network topology/structure.

2.4.1 Descriptive Statistical Characteristics

Table 1 shows descriptive statistics (mean, standard deviation, minimum, first quartile, median, third quartile, maximum, and density plots) for the response variable (follower growth) as well as seventeen of the twenty-two predictor and control variables. The x-axes of the density plots represent the measured value of the variable, and the y-axis indicates the density of users observed at a particular value. For example, one can interpret the table to indicate that most users grew their Twitter audience at a rate of about 12 to 106 new followers (median=36) every 3 months. The density plot indicates that most users fell within this range. For space reasons, I omit user profile data from the table, and instead

Table 1: Descriptive statistics for the dependent variable (follower growth) and seventeen of the twenty-two predictor and control variables (details in Section 2.3.3).

	Variable	Mean	Std Dev	Min	1st Q	Median	3rd Q	Max	Density Plot
D.V.	Follower Growth	194.2	832.7	0	12	36	106	16,623	
Behavioral and Social Interaction Variables	Number of Tweets in period (a control)	262.6	176.3	21	131	222	364	1,552	
	Peak tweets per hour ("Burstiness")	6.39	5.78	0.15	2.79	4.79	7.9	48.9	
	Directed communications	1.91	7.4	0	0.58	0.83	1.22	190.25	
	Broadcast communications	0.48	0.22	0	0.31	0.45	0.62	1	
Message Content Variables	Positive Sentiment Intensity Rate	0.37	0.14	0.05	0.27	0.35	0.44	1.08	
	Negative Sentiment Intensity Rate	0.14	0.06	0	0.095	0.13	0.17	0.5	
	Informative content index	0.3	0.23	0	0.12	0.24	0.41	1	
	"Meformative" content index	0.41	0.14	0	0.33	0.41	0.50	0.79	
	Topic focus	0.008	0.01	0.002	0.005	0.008	0.01	0.25	
	User RT ratio	0.15	0.4	0	0.02	0.05	0.12	5.1	
	Hashtag usage ratio	0.2	0.24	0	0.057	0.13	0.26	2.82	
	Tweet Reading Difficulty Index (TReDIX)	2.36	0.64	0.84	1.94	2.31	2.696	6.95	
Network Topology / Structural Variables	Reciprocity rate	0.28	0.19	0	0.125	0.25	0.4	0.9	
	Attention-status ratio	2.18	7.06	0	0.895	1.19	1.90	149.25	
	Network overlap	94,730	351,388	0	2,070	10,472	50,263	5,308,200	
	No. of followers at end of period (a control)	1,145.42	3391.93	15	175.8	391.5	948.8	45,932	
	No. of friends at end of period (a control)	830.63	2879.43	18	135	289.5	661.2	42,797	

provide the following summary: the majority of users (86%) had URLs listed in their profile, most (97%) also listed their location, and the average profile description was 85 characters long. I also omit the lagged variable *change in followers (previous period)* (mean=106.96, SD=551.84, median=25). The density plots in show the distributions for each variable, which reveals some skewness (lack of symmetry) and generally high kurtosis (peaked, rather than flat, distributions) for many of the variables. This makes the median a better measure of central tendency than the mean for such variables.

2.4.1.1 Behavioral and Social Interaction Variables

Most users tweeted between 131-364 times in three months (median=222), usually with bursts of no more than eight tweets within a single hour. The Broadcast Communication Index shows the proportion of tweets that are not directed to any specific person; people typically use broadcast communication strategies for about 30%-60% of their messages (median=45%).

2.4.1.2 Message Content Variables

Proportionally, most people tweet about twice as much positive and neutral content as negative content, with an average of 106 tweets per user identified as positive (about 40% of their average number of tweets for a given three month interval; roughly the same proportions were neutral tweets), and 51 tweets (about 20% of their average total for a period) were labeled as negative. (Note: this data did not fit in Table 1, but is presented graphically in Figure 6).

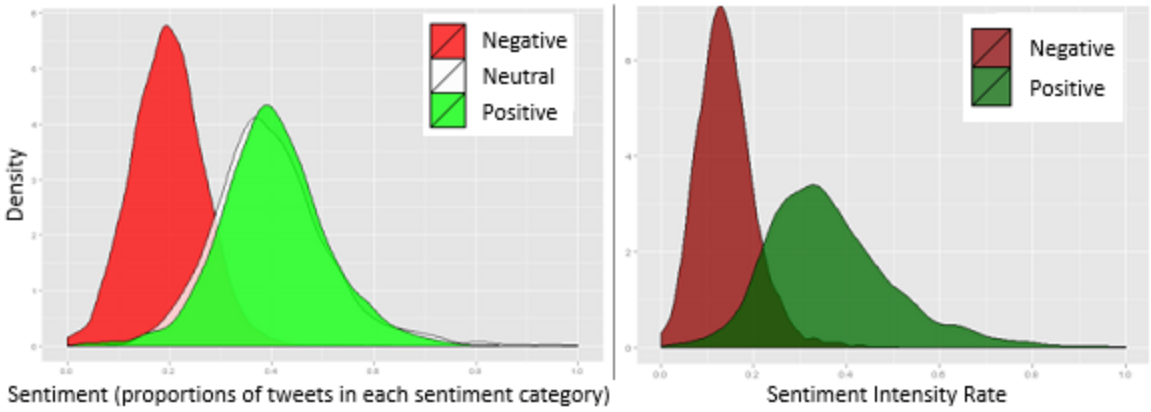


Figure 6: For most people, negative content makes up about 20% of all tweets, while positive and neutral content each make up about 40% of tweets for most people (left). When people tweet sentiment-laden content, the intensity of positive sentiment is about three times higher than negative sentiment (right).

In terms of *intensity* of positive or negative language, most people are generally about three times more positive than they are negative in their tweets (see Table 1 and Figure 6). In subsection 2.4.2, I will assess the extent to which these attributes of message content influence social tie formations that lead to audience growth over time.

The proportion of users' tweets identified as "meformative" content was nearly normally distributed – users talk about themselves in 41% of their messages, on average. Informative content accounted for 24% of messages. This closely resembles the results from [154]. The mean and median of topical focus (average cosine similarity of one's own tweets) indicate that in general, people post a fairly diverse range of content. The ratios of retweets (0.02-0.12, median=0.05) and hashtag usage (0.06-0.26, median=0.13) to total number of tweets in a period are moderate for the majority of users – retweets generally comprised between 5-12% of users' messages, and hashtags were used in about 13-26% of tweets for most users. The Tweet Reading Difficulty Index (TReDIX) is evenly distributed, with most people using moderately sophisticated language – about 2.36 long words per

tweet, on average. On the original RIX scale, an index of 2.4 is equivalent to a seventh grade reading level [6].

2.4.1.3 Network Topology / Structural Variables

The majority of users have 176-949 followers, and 135-661 friends (medians are 391.5 and 289.5, respectively). The density plots indicate that few users fell outside these ranges, but those that exceeded the range did so by a large margin. In general, users reciprocally follow-back about a quarter of their followers (mean=28%, median=25%). The density plot for attention-status ratio (that is, followers to following) shows a very tight distribution around the range 0.895 to 1.9, indicating that many people have similar numbers of in-degree connections (followers) as out-degree connections (friends). About 2K-50K overlapping network neighbors are typical, though some users with very large networks have over two orders of magnitude more.

2.4.2 Relative Prominence of the Factors Predicting Persistent Social Ties

I now turn to the core of the results: how well do these variables predict persistent social tie formation (follower growth over time) and by how much? The overall significance of the negative binomial auto-regressive model is very high (p < 2e-16), meaning the model is well-suited to characterizing the effects of the described variables on social tie formation over time. Significance is judged by the reduction in deviance from a null model, χ^2 (22, N=1,836) = 5943.9 – 2111.9 = 3832.0, p < 2e-16. This is important in order to have confidence when interpreting the regression coefficients of the model components (b and β), which are depicted in Table 2.

Table 2: Negative Binomial Auto-Regressive Model Coefficients.

	b	Std. Err.	Std. *β*	*p*-value
NumTweetsPd	2.63e-04	1.62e-04	5.57e-05	0.104
PeakTPH	2.35e-02	4.94e-03	1.63e-04	1.96e-06***
DirectedComms	4.24e-03	3.37e-03	3.77e-05	0.208
BroadcastComms	–1.02	1.28e-01	-2.67e-04	1.89e-15***
ProfDescLen	3.09e-03	5.57e-04	1.72e-04	2.94e-08***
ProfHasURL	3.91e-01	7.14e-02	1.65e-04	4.27e-08***
ProfHasLocation	3.29e-01	1.52e-01	6.30e-05	0.03995 *
PosSentiRate	8.19e-01	1.96e-01	1.37e-04	2.87e-05***
NegSentiRate	–2.38	4.82e-01	-1.75e-04	7.53e-07***
InformContent	1.18	1.41e-01	3.31e-04	< 2e-16 ***
MeformContent	–6.72e-02	1.99e-01	-1.12e-05	0.736
TopicFocus	3.75e-01	2.32	5.13e-06	0.872
UserTweetRT'd	9.53e-01	7.23e-02	4.60e-04	< 2e-16 ***
HashtagUseRate	–4.28e-01	1.12e-01	-1.23e-04	1.33e-04***
TReDIX	1.28e-01	4.22e-02	9.85e-05	2.43e-03 **
Reciprocity	3.52e-01	1.46e-01	7.95e-05	0.01597 *
Attn-Status	1.63e-02	4.48e-03	1.38e-04	2.79e-04***
NetworkOverlap	1.20e-06	1.26e-07	5.06e-04	< 2e-16 ***
NumFriends	–1.73e-04	2.88e-05	-5.98e-04	1.96e-09***
NumFollowers	2.70e-04	2.4e-05	1.10e-03	< 2e-16 ***
ChngFollPrevPd	–2.71e-04	8.82e-05	-1.79e-04	2.17e-03 **
AgeOfAccount	4.10e-03	2.26e-04	5.50e-04	< 2e-16 ***

The unstandardized *b* coefficients in Table 2 are useful in that they can be directly interpreted according to the native units of each predictor: for each one unit change in the predictor variable, the log count of the response variable is expected to change by the respective *b* coefficient (all else being equal). While this is valuable for a broad range of prediction and forecasting purposes, I am also interested in comparing the *relative* impact of each predictor; I therefore also report the standardize beta (*β*) coefficients (see also Figure 7).

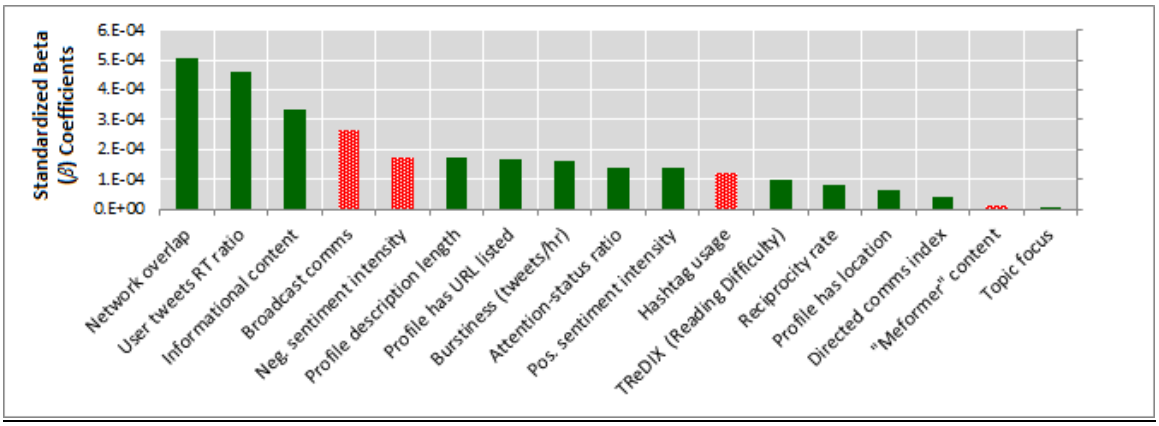


Figure 7: Standardized beta coefficients (β) show the relative effect sizes that each input variable has on follower growth. Green bars indicate positive effects on follower gain, and red bars indicate negative effects (i.e., suppression of follower growth).

Not pictured in Figure 7 are three of the control variables used in this study: extant friends and followers, age of account, and the lagged variable. As expected, these controls absorb comparatively large portions of the variance (see Table 2). Here, I am principally interested in how much the other variables contribute above and beyond the controls.

2.4.2.1 Message Content Influences Social Tie Formation & Retention

Message content variables are evenly distributed along the rank ordered list of predictors (see Figure 7). This leads to the first major finding: *message content significantly impacts audience growth*. Six of the eight content variables (negative and positive sentiment, informational and "retweetable" content, hashtag usage, and linguistic sophistication) were found to be significant predictors of persistent social tie formation and retention. Of the 17 (non-control) variables, expressing negative sentiments in tweets is the second most harmful factor to growing a Twitter audience (see Table 2 and Figure 7). In contrast to [123], where social sharing of negative emotions correlates to higher numbers of followers, I find that expressing negative sentiment has an adverse effect on follower

growth over time. However, [123] studied a static snapshot of existing network ties. The longitudinal data suggest that sentiment expression may have different (indeed opposite) effects on the formation of new ties in the long run. This might be because Twitter is a medium dominated by very weak social ties [70], and negative sentiment from strangers may be unpleasant or uncomfortable for a potential new follower to see. For [123]'s study of *existing* ties, on the other hand, negative expressions such as the sharing of a death, poor health, bad news, or a state of unhappiness, can trigger opportunities to build bonding social capital between stronger ties who want to seek and provide emotional support [217]. Or, as [123] put it, "gift giving where users directly exchange digital 'gifts' in terms of emotional messages".

Producing or passing along informative content is also among the top predictors, having a significant positive effect on follower growth rates (β = 3.31e-04). I also found that informative content attracts followers with an effect that is roughly *thirty times higher* than the effect of "meformer" content, which deters growth. This is possibly due to the prevalence of weak ties on Twitter [70], and that informativeness [81,130] is a more palatable alternative to meforming among such networks. Kollock [125] describes information as a public good that anyone can consume and share. Retweeted content is another such digital public good that provides both attribution—and thus, motivation—to the original author as well as informative content for the community. Having content that is "retweet worthy" is a very good indicator that a user will gain followers (β = 4.60e-04). Retweeted content provides *social proof* [31] that a user may be worth following, enabling the process of triadic closure [55] to unfold, whereby followers of a user's followers complete the triad with the user [75].

The mean and median of topical focus (average cosine similarity of tweets) for our heterogeneous group is roughly an order of magnitude less than those same measures from a more homogenous group of politically-oriented tweeters described in [215]. But while this variable slightly misses being significant in the model (p = 0.872), the positive sign of the regression coefficient (β = 5.13e-06) suggests a trend upward such that a more topically focused users generally tend to attract more followers, congruent with [215]. Twitter users are likely driven by homophily [149], where they seek out content and users who are similar to themselves.

Interestingly, overuse of hashtags in message content ("hashtag abuse") seems to significantly reduce follower gain. This is evidenced by the data in Table 2 that shows the hashtag use rate variable was highly significant (p = 1.33e-04), and the regression coefficients showed relatively strong negative effects (β = –1.23e-04). On one hand, hashtags help signal a broader public conversation. They are valuable for enabling users to discover content, and follow (and potentially engage) in discussions [95]. On the other hand, hashtags are more difficult to read, especially when the tag contains more than a single word, and multiple hashtags are often associated with poorly conceived advertising and marketing campaigns rather than social communications. Prior research demonstrated that, relative to having no hashtags in a tweet at all, having one or two hashtags increases engagement (retweets, mentions, and favorites), but engagement decreased when more than three hashtags were present—and continued to decrease as the number of hashtags grew [110,118]. The data in this study suggests a similar pattern may apply to growing an audience: moderation is key when it comes to using hashtags.

Using more sophisticated language in messages also has a moderately strong relative effect on attracting and retaining followers (β = 9.85e-05), and the Tweet Reading Difficulty Index (TReDIX) has a positive impact on audience growth. Walther's Social Information Processing (SIP) theory suggests that people rely on linguistic cues like spelling and vocabulary to compensate for the lack of traditional contextual cues available in face-to-face settings [209]. Twitter users apparently seek out well-written content over poorly written content when deciding whether to follow another user.

2.4.2.2 Behavioral Choices Matter for Persistent Social Tie Formation and Retention

The second major finding is that *social behavioral choices can dramatically affect network growth*. Similar to previous research that showed positive effects of profile completeness for static Facebook networks [132], I find similar results for evolving Twitter networks: all three of the profile elements (length of description, URL, and location) each emerge as significant predictors of social tie attraction, acquisition, and retention over time. Signaling theory suggests that choosing to complete user profile elements helps persuade other users one's authenticity and trustworthiness, making them more likely to become followers [50]. Profile content provides at minimum *conventional signals* of identity (which are easy to fake), but the nature of profiles on social network sites makes these signals somewhat more reliable due to social accountability [50]. Regardless, users who do take the time to give profile information have the opportunity to emphasize the characteristics that they think will present them in the best light without necessarily being deceptive [74]. Others can use this profile information to form impressions prior to deciding whether to pursue or continue a connection [132].

Likewise, choices about interactions and communication techniques, such as sending directed versus broadcast messages, will also impact the rate at which a user will grow their audience. The Broadcast Communications Index (BroadcastComms) and the burstiness measure (PeakTPH were both significant predictors of persistent social tie formation. The moderately strong negative effect of BroadcastComms (b = -1.02, β = -2.67e-04) suggests that having too many undirected messages will hinder audience growth. In contrast to similar work by [21] for Facebook users, broadcast communication techniques on Twitter have a suppressing effect during the process of network tie formation. Such undirected messages are a relatively novel feature of social media; my results suggest that relying on such communication techniques will significantly subdue follower growth. On the other hand, consistent with [21], I also find that the general trend is for directed communications to have a positive effect on follower growth for Twitter – but interestingly, the Directed Communications Index (DirectedComms) was not significant in the statistical model. Apparently, in the presence of all the other variables, the significance of social interactions using @replies and @mentions is muted, at least in terms of its effect on attracting and acquiring *new* followers.

2.4.2.3 Even Simple Measures of Network Structure Are Useful

Network oriented variables are also evenly distributed along the ranked list in Figure 7. Reciprocity, status, and network overlap were each significant in the model, even in the presence of the variables controlling for network size and user popularity. Thus, the third finding is that *variables related to network structure are useful predictors of audience growth*. This finding is not necessarily surprising, given the emphasis on such factors in much of the related literature [12,75,81,122,136]. Indeed, while the results indicate that

even simplistic calculations of network structure can prove to be quite powerful, I highlight the point such factors should not necessarily be privileged over message content or social behavior measures.

2.4.3 Practical Implications

A vital prerequisite to building social capital of any kind (bonding or bridging) is that a connecting tie must exist between individuals. The practical implication of this fundamental antecedent to social capital motivates the selection of the dependent variable in this study. The number of followers you have is arguably the most important status symbol on Twitter. Rapid follower growth may be an early indication of a rising influencer, or an emerging thought leader, within the network. A rapid gain in followers intuitively implies that people like what you're posting and want more of the same. Thus, social capital is a necessary (though not sufficient) precursor to the notion of *interpersonal influence* in social networks [11] – an attribute of interest to strategic communicators, marketers, advertisers, job seekers, activist groups and any entity or organization wishing to disseminate specific messages in a timely manner. Additionally, many users are simply interested in knowing their own relative degrees of popularity or social networking "clout". Sites like HootSuite.com and SocialFlow.com offer web services oriented towards helping its users capture and retain the attention of social media audiences. Companies like these can directly leverage our results to build tools that that make recruiting and retaining network members easier and more effective. For example, in conjunction with a validated tie-strength model (e.g., [72] or [70]), the results of this study suggest that social media technology developers can help users retain existing followers by actively promoting negative sentiment content only for strong ties, and possibly de-emphasize negative content

for weak ties. Similarly, to attract the attention of new audience members, developers can consider implementing user interface components which a) facilitate the sharing of informative content through positive reinforcement, b) encourage directed communications and group discussions, c) provide feedback regarding behavioral patterns (e.g., burstiness), and so on.

In addition to the practical implications for social computing technology developers, individual users can also benefit from understanding the empirical evidence documented in this research. For example, over the long run, the data from this research can be encapsulated into the following nine guidelines for successful Twitter users:

1. **Don't whine online**. This means tweeting content that is more positive in nature, rather than negative (including swear words). Negative-oriented content will often be a turn-off to a potential new follower who is assessing whether to make a connection with you (exceptions for when negative-oriented content is used in conjunction with humor, inspiration/education, or controversy), but consistent positive-oriented content will help boost follower growth rates over time.

2. **Talk *to* people, rather than talking *at* people**. Employing directed communication strategies (e.g., mentioning other users in your tweets, retweeting others, and replying to or favoriting others' tweets), rather than broadcast communication strategies (which do not target anyone in particular) will help make you more visible and more personable – both of which will help to attract and retain followers. Having engaging interactions with your existing followers also helps you leverage your extended social network in order to become visible to (and hopefully

appealing to) the followers of your followers, as well as those people who have friends or followers in common with you.

3. **Be *informative*, rather than "*meformative*"**. The overwhelming majority of connections on Twitter comprise very weak social ties. In other words, for many Twitter users, Twitter is a social network made up mostly of connections between virtual strangers and weak acquaintances rather than very good friends. For these kinds of ties, details about the mundane minutiae of your everyday personal life (like what you ate for breakfast, the outcome of your daughter's soccer game, etc.) are much less attractive than timely or novel bits of news. In this way, Twitter slants more towards an *information* network rather than strictly a *social* network, per se.

4. **Don't abuse hashtags**. Hashtags serve a very useful function; when used as intended, hashtags help to signal keywords within tweets that are related to a broader public topic, conversation, or group. They are also useful for expressing humor, excitement, sarcasm or other contextual content, for example, "Just found out my mom is my health teacher. #awkward" or "It's Monday!! #excitedsarcasm". On the other hand, hashtags are more difficult to read, especially when the tag contains more than a single word (e.g., #multiwordhashtagsarehardtoparse, #keepitsimplesilly). So when you combine the readability issue with the fact that some users are tempted to #spam #with #hashtags #in #short #tweets (i.e., over-tagging a single Tweet), then it is no wonder that many micro-bloggers feel that excessive #hastagscanbeannoying.

5. **Use more sophisticated writing**. People rely on linguistic cues like spelling and vocabulary to compensate for the lack of traditional contextual cues available in face-to-face settings. When deciding whether or not to follow a virtual stranger, Twitter users seek out well-written content over poorly written content.

6. **Be clear about who you are, and what you're about**. Completely fill in all the parts of your user profile. Again, in the absence of face-to-face interactions, it's about sending the signals that indicate you are a real person with real interests. Having a personalized photo, something about your geographic location, and listing a website are helpful. Your profile description should also indicate what it is you will likely be tweeting about – the richer the details in your description, the better the results for attracting new followers.

7. **Tweet more, and don't go too long between updates**. It's all about visibility and engagement! The more you tweet, the more visible you are. Most of the users in our dataset tweeted less than 8 times per hour, but some went days and weeks between tweets. Accounts with long periods of stagnation are less attractive than those with up-to-date content.

8. **Follow-back**. Paying back a new follower by following them in kind (i.e., the principle of reciprocity) is a useful strategy because it reduces the likelihood that new followers will un-follow you, leading to sustained/persistent audience growth.

9. **Stay on topic**. When faced with the choice between tweeting about numerous different topics (to appeal to a broader audience) and choosing to tweet about a select set of topics, the data was inconclusive ('topic focus' was not significant in

the presence of all the other variables in the model). However, the directional trend agreed with previous research suggesting it is better to build a reputation for interest in specific topics.

2.4.4 Theoretical and Methodological Implications

The findings from this study also have theoretical (and, by extension, methodological) implications. The variables were selected from prominent theoretical perspectives bridging social science theory (e.g., social capital, signaling theory, presentation of self, homophily, social proof, status/power/attention, social information processing theory), and network theory (size and preferential attachment, tie strength, reciprocity and mutuality, structural balance and triadic closure). I also consider behavioral aspects of computer mediated communications (profile completeness, directed versus broadcast communication strategies) and message content (sentiment, informative versus meformative content, topical focus, linguistic sophistication). Few social media studies have attempted to report on relative impacts of such diverse variables. Compared to how much is known about each theory, very little is known about how they relate to one another. This research compares their *relative* contributions to predicting link formations in online social networks. This was a significant undertaking, but more work should be done to understand the relative effects of different variables—as well as different theoretical perspectives and methodological approaches—on study outcomes.

2.4.5 Study Limitations

I have attempted to be reasonably thorough and inclusive; but this is still merely a single study, and other variables could explain some of the results. For example, a person's

real-world celebrity status, or other exogenous factors like being publicly mentioned in mass communications (news media, printed press, commercials and advertisements, etc.) may contribute to audience growth. Secondly, I do not segment the Twitter data sample into types of users or types of uses, although [21], [154], and [184] suggest ways in which categories for specific user and uses may illuminate the processes of attracting network members. Thirdly, this is a quantitative study based on observations with calculated latent measures from those observations. This approach is useful for describing *what* happens, but without a corresponding qualitative approach, I can only speculate on *why*. Future work could explore why certain variables predict follower growth more than others. Finally, Twitter is one site. I don't know how well the results presented here translate into other social media technologies, or how durable they will remain as the platform matures.

2.5 Chapter Summary

I believe this is the first longitudinal study of audience growth on Twitter to combine such a diverse set of theoretically-grounded variables [99]. I explore the relative effects of social behavior, message content, and network structure on persistent social tie formation and show which of variables have more explanatory power than the others. Though these results are specific to Twitter and a particular dataset, they are important for the following reasons. First, multiple snapshots from the longitudinal method helps begin to offer casual explanations for audience growth. Second, comparisons across many variables inspired by different theoretical perspectives allows researchers to interpret, compare, and contrast the relative effects of each. Third, the impact of message content and social behavior are comparative to network structure, which suggests future work should take caution in privileging any one perspective over another.

It is this third point that prompts much of the effort described in the remaining chapters of this dissertation. Rigorous study of *message content* seems to be relatively underrepresented in scientific literature when compared to the body of works which investigate behavioral interactions and network characteristics. Perhaps this is because technology more readily facilitates observation and measurement of quantitative oriented features. For example, it is fairly simple to count the number of Likes and Shares or Favorites and Retweets. Likewise, technological implementations specifically suited to social network (structural) analysis are also prevalent, making characteristics of the network similarly straightforward to extract. Contrariwise, details of social communications contained within the message content itself poses substantial challenges that make it more effortful to study, typically involving advancements in techniques related to computational Natural Language Processing (NLP). One way to help facilitate researchers giving equal prominence to studying attributes of message content is to develop, implement, and validate new methods and tools to facilitate large scale analysis of those social phenomena of interest within text-based digital content. This is the research challenge I subsequently address.

CHAPTER 3. SENTIMENT ANALYSIS FOR SOCIAL TEXT

3.1 Chapter Overview

Sentiment analysis is useful to a wide range of problems that are of interest to human-computer interaction practitioners and researchers, as well as those from fields such as sociology, marketing and advertising, psychology, economics, and political science. The inherent social nature of microblog content - such as those observed on Twitter and Facebook - poses serious challenges to practical applications of sentiment analysis. Some of these challenges stem from the sheer rate and volume of user generated content, combined with the contextual sparseness resulting from shortness of the text and a tendency to use abbreviated language conventions to express sentiments.

A comprehensive, high quality lexicon is often essential for fast, accurate sentiment analysis on such large scales. An example of such a lexicon that has been widely used in the social media domain is the Linguistic Inquiry and Word Count (LIWC, pronounced "Luke") [174,175]. Sociologists, psychologists, linguists, and computer scientists find LIWC appealing because it has been extensively validated. Also, its straightforward dictionary and simple word lists are easily inspected, understood, and extended if desired. Such attributes make LIWC an attractive option to researchers looking for a reliable lexicon to extract emotional or sentiment polarity from text. Despite their pervasive use for gaging sentiment in social media contexts, these lexicons are often used with little regard for their actual suitability to the domain.

This chapter describes the development, validation, and evaluation of VADER (Valence Aware Dictionary and sEntiment Reasoner), a system comprised of both a newly developed social media oriented sentiment lexicon and a rule-based algorithm for analyzing textual content using that (or another preferred) lexicon. More specifically, I use

a combination of qualitative and quantitative methods to produce, and then empirically validate, a gold-standard[8] sentiment lexicon that is especially attuned to text-based content in microblog-like social communications (but is also suitable for other online content like news articles or product and movie reviews). I next derive five generalizable rules that embody grammatical and syntactical conventions that humans use when expressing or emphasizing sentiment *intensity*. I find that incorporating these heuristics improves the accuracy of the sentiment analysis engine across several domain contexts (social media text, New York Times news editorials, online movie reviews, and online product reviews). Interestingly, the VADER lexicon performs exceptionally well in the social media domain. The correlation coefficient shows that VADER (r = 0.881) performs as well as individual human judgments (r = 0.888) at matching ground truth (aggregated group mean from 20 human raters for sentiment intensity of each tweet). Surprisingly, when I further inspect the classification accuracy of VADER ($F1$ = 0.96), it actually even outperforms individual human raters ($F1$ = 0.84) at correctly classifying the sentiment of tweets into positive, neutral, or negative classes. VADER also generalized well into the other text domains.

VADER preserves (and improves on) the benefits of traditional sentiment lexicons like LIWC: it is bigger, yet just as simply inspected, understood, quickly applied (without a need for extensive learning/training) and easily extended. Like LIWC (but unlike some other lexicons or machine learning models), the VADER sentiment lexicon is *gold-standard* quality and has been validated by humans. VADER distinguishes itself from LIWC in that it is more sensitive to sentiment expressions in social media contexts while

[8] Gold standard is a historical term (borrowed from economists) signifying a monetary standard, under which basic units of currency were defined by a stated quantity of gold. The value of any country's currency was stated in terms of the gold standard, making it possible to compare different currencies for international trading. The analogy should be clear in this context: the aggregated judgements from numerous (appropriately screened, trained) human judges denotes the best standard available by which to compare the results of any other sentiment analyses (including those by individual humans themselves).

also generalizing more favorably to other domains. I also make VADER freely available for download and use[9].

3.2 Sentiment Analysis in Computer and Social Science Scholarship

Sentiment analysis, or opinion mining, is an active area of study in the field of natural language processing that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text. It is not my intention to review the entire body of literature concerning sentiment analysis (indeed, such treatments are already available in [142] and [163]. I do, however, provide a brief overview of canonical works and techniques which help situate the current research effort.

3.2.1 Sentiment Lexicons

A substantial number of sentiment analysis approaches rely greatly on an underlying sentiment (or opinion) lexicon. A *sentiment lexicon* is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative [141]. Manually creating and validating such lists of opinion-bearing features via detailed qualitative data analysis (QDA), while being among the most robust methods for generating reliable sentiment lexicons, is also one of the most time-consuming. For this reason, much of the applied research leveraging sentiment analysis relies heavily on preexisting manually constructed lexicons. Because lexicons are so useful for sentiment analysis, I briefly provide an overview of several appropriate benchmarks. I first review

[9] https://github.com/cjhutto/vaderSentiment

three widely used lexicons (LIWC[10], GI[11], Hu-Liu04[12]) in which words are categorized into binary classes (i.e., either *positive* or *negative*) according to their context free semantic orientation. I then describe three other lexicons (ANEW[13], SentiWordNet[14], and SenticNet[15]) in which words are associated with valence scores for sentiment *intensity*.

3.2.1.1 Semantic Orientation (Polarity-based) Lexicons

LIWC is text analysis software designed for studying the various emotional, cognitive, structural, and process components present in text samples. LIWC uses a proprietary dictionary of almost 4,500 words organized into one (or more) of 76 categories, including 905 words in two categories especially related to sentiment analysis (Table 3):

Table 3: Example words from two of LIWC's 76 categories. These two categories can be leveraged to construct a semantic orientation-based lexicon for sentiment analysis.

LIWC Category	Examples	No. of Words
Positive Emotion	Love, nice, good, great	406
Negative Emotion	Hurt, ugly, sad, bad, worse	499

LIWC is well-established and has been both internally and externally validated in a process spanning more than a decade of work by psychologists, sociologists, and linguists [174,175]. Its pedigree and validation make LIWC an attractive option to researchers looking for a reliable lexicon to extract emotional or sentiment polarity from social media text. For example, LIWC's lexicon has been used to extract indications of political

[10] www.liwc.net

[11] http://www.wjh.harvard.edu/~inquirer

[12] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[13] http://csea.phhp.ufl.edu/media/anewmessage.html

[14] http://sentiwordnet.isti.cnr.it/

[15] http://sentic.net/

sentiment from tweets [210], predict the onset of depression in individuals based on text from social media [44], characterize the emotional variability of pregnant mothers from Twitter posts [43], unobtrusively measure national happiness based on Facebook status updates [126], and differentiating happy romantic couples from unhappy ones based on their instant message communications [84]. However, as I point out in Chapter 2, despite its widespread use for assessing sentiment in social media text, LIWC does not include consideration for sentiment-bearing lexical items such as acronyms, initialisms, emoticons, or slang, which are known to be important for sentiment analysis of social text [42]. Also, LIWC is unable to account for differences in the sentiment *intensity* of words. For example, "The food here is *exceptional*" conveys more positive intensity than "The food here is *okay*". A sentiment analysis tool using LIWC would score them equally (they each contain one positive term). Such distinctions are intuitively valuable for fine-grained sentiment analysis.

The General Inquirer (GI) is a text analysis application with one of the oldest manually constructed lexicons still in widespread use. The GI has been in development and refinement since 1966, and is designed as a tool for *content analysis*, a technique used by social scientists, political scientists, and psychologists for objectively identifying specified characteristics of messages [205]. The lexicon contains more than 11K words classified into one or more of 183 categories. For my purposes, I focus on the 1,915 words labeled *Positive* and the 2,291 words labeled as *Negative*. Like LIWC, the Harvard GI lexicon has been widely used in several works to automatically determine sentiment properties of text [58,114,211]. However, as with LIWC, the GI suffers from a lack of coverage of sentiment-

relevant lexical features common to social text, and it is ignorant of *intensity* differences among sentiment-bearing words.

Hu and Liu [94,143] maintain a publicly available lexicon of nearly 6,800 words (2,006 with positive semantic orientation, and 4,783 negative). Their opinion lexicon was initially constructed through an automated bootstrapping process [94] using WordNet [62], a well-known English lexical database in which words are clustered into groups of synonyms known as *synsets*. The Hu-Liu04 opinion lexicon has evolved over the past decade, and (unlike LIWC or the GI lexicons) is more attuned to sentiment expressions in social text and product reviews – though it still does not capture sentiment from emoticons or acronyms/initialisms.

3.2.1.2 Sentiment Intensity (Valence-based) Lexicons

Many applications would benefit from being able to determine not just the binary polarity (positive versus negative), but also the *strength* of the sentiment expressed in text. Just how favorably or unfavorably do people feel about a new product, movie, or legislation bill? Analysts and researchers want (and need) to be able to recognize changes in sentiment *intensity* over time in order to detect when rhetoric is heating up or cooling down [228]. It stands to reason that having a general lexicon with strength valences would be beneficial.

The Affective Norms for English Words (ANEW) lexicon provides a set of normative emotional ratings for 1,034 English words [20]. Unlike LIWC or GI, the words in ANEW have been ranked in terms of their pleasure, arousal, and dominance. ANEW words have an associated sentiment valence ranging from 1-9 (with a neutral midpoint at five), such that words with valence scores less than five are considered

unpleasant/negative, and those with scores greater than five are considered pleasant/positive. For example, the valence for *betray* is 1.68, *bland* is 4.01, *dream* is 6.73, and *delight* is 8.26. These valences help researchers measure the intensity of expressed sentiment in microblogs [43,44,159] – an important dimension beyond simple binary orientations of positive and negative. Nevertheless, as with LIWC and GI, the ANEW lexicon is also insensitive to common sentiment-relevant lexical features in social text.

SentiWordNet is an extension of WordNet [62] in which 147,306 synsets[16] are annotated with three numerical scores relating to positivity, negativity, and objectivity (neutrality) [10]. Each score ranges from 0.0 to 1.0, and their sum is 1.0 for each synset. The scores were calculated using a multi-step process involving eight semi-supervised learning algorithms and then a random walk algorithm. It is thus not a *gold standard* resource like WordNet, LIWC, GI, or ANEW (which were all 100% curated by humans), but it is useful for a wide range of tasks. I interface with SentiWordNet via Python's Natural Language Toolkit[17] (NLTK), and use the difference of each sysnset's positive and negative scores as its sentiment *valence* to distinguish differences in the sentiment intensity of words. The SentiWordNet lexicon is very noisy; a large majority of synsets have no positive or negative polarity. It also fails to account for sentiment-bearing lexical features relevant to text in microblogs.

SenticNet is a publicly available semantic and affective resource for concept-level opinion and sentiment analysis [23]. SenticNet is constructed by means of *sentic*

[16] WordNet is a lexical database for the English language that groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and describes several types of relationships across synsets and among synset members [62].

[17] http://www.nltk.org

computing, a paradigm that exploits both AI and Semantic Web techniques to process natural language opinions via an ensemble of graph-mining and dimensionality-reduction techniques [24]. The SenticNet lexicon consists of 14,244 common sense concepts such as *wrath*, *adoration*, *woe*, and *admiration* with information associated with (among other things) the concept's sentiment *polarity*, a numeric value on a continuous scale ranging from –1 to 1. I access the SenticNet polarity score using the online SenticNet API and a publicly available Python package[18].

3.2.1.3 Lexicons and Context-Awareness

Whether one is using binary polarity-based lexicons or more nuanced valence-based lexicons, it is possible to improve sentiment analysis performance by understanding deeper lexical properties (e.g., parts-of-speech) for more context awareness. For example, a lexicon may be used in conjunction word-sense disambiguation (WSD) [3]. Word-sense disambiguation refers to the process of identifying which sense of a word is used in a sentence when the word has multiple meanings (i.e., its contextual meaning). For example, using WSD, we can distinguish that the word *catch* has negative sentiment in "At first glance the contract looks good, but there's a *catch*", but is neutral in "The fisherman plans to sell his *catch* at the market". I use a publicly available Python package[19] that performs sentiment classification with word-sense disambiguation.

Despite their ubiquity for evaluating sentiment in social media contexts, there are generally three shortcomings of lexicon-based sentiment analysis approaches: 1) they have

[18] senticnet 0.3.2 (https://pypi.python.org/pypi/senticnet)

[19] https://pypi.python.org/pypi/sentiment_classifier/0.5

trouble with coverage, often ignoring important lexical features which are especially relevant to social text in microblogs, 2) some lexicons ignore general sentiment intensity differentials for features within the lexicon, and 3) acquiring a new set of (human validated gold-standard) lexical features – along with their associated sentiment valence scores – can be a very time consuming and labor intensive process. The research effort described in this chapter is an opportunity not only to address this gap by constructing just such a lexicon and providing it to the broader research community, but also a chance to compare its efficacy against other well-established lexicons with regards to sentiment analysis of social media text and other domains.

3.2.2 Machine Learning Approaches

Because manually creating and validating a comprehensive sentiment lexicon is labor and time intensive, much work has explored automated means of identifying sentiment-relevant features in text. Typical state of the art practices incorporate machine learning approaches to "learn" the sentiment-relevant features of text.

The Naive Bayes (NB) classifier is a simple classifier that relies on Bayesian probability and the naive assumption that feature probabilities are independent of one another. Maximum Entropy (MaxEnt, or ME) is a general purpose machine learning technique belonging to the class of exponential models using multinomial logistic regression. Unlike NB, ME makes no conditional independence assumption between features, and thereby accounts for information entropy (feature weightings). Support Vector Machines (SVMs) differ from both NB and ME models in that SVMs are non-probability classifiers which operate by separating data points in space using one or more

hyperplanes (centerlines of the gaps separating different classes). I use the Python-based machine learning algorithms from scikit-learn.org for the NB, ME, SVM-Classification (SVM-C) and SVM-Regression (SVM-R) models.

Machine learning approaches are not without drawbacks. First, they require (often extensive) training data which are, as with validated sentiment lexicons, sometimes troublesome to acquire. Second, they depend on the training set to represent as many features as possible (which often, they do not – especially in the case of the short, sparse text of social media). Third, compared to dictionary-based lexicons, they are often more computationally expensive in terms of CPU processing, memory requirements, and training/classification time (which restricts the ability to assess sentiment on streaming data). Fourth, they often derive features "behind the scenes" inside of a black box that is not (easily) human-interpretable and are therefore more difficult to either generalize, modify, or extend (e.g., to other domains).

3.3 VADER Development, Validation, and Evaluation

My development approach seeks to leverage the advantages of parsimonious rule-based modeling to construct a computational sentiment analysis engine that 1) works well on social media style text, yet readily generalizes to multiple domains, 2) requires no additional training data, but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon 3) is fast enough to be used online with streaming data, and 4) does not severely suffer from a speed-performance tradeoff.

Figure 8 provides an overview of the research process and summarizes the methods used in the study described in this chapter. In essence, this research involves three

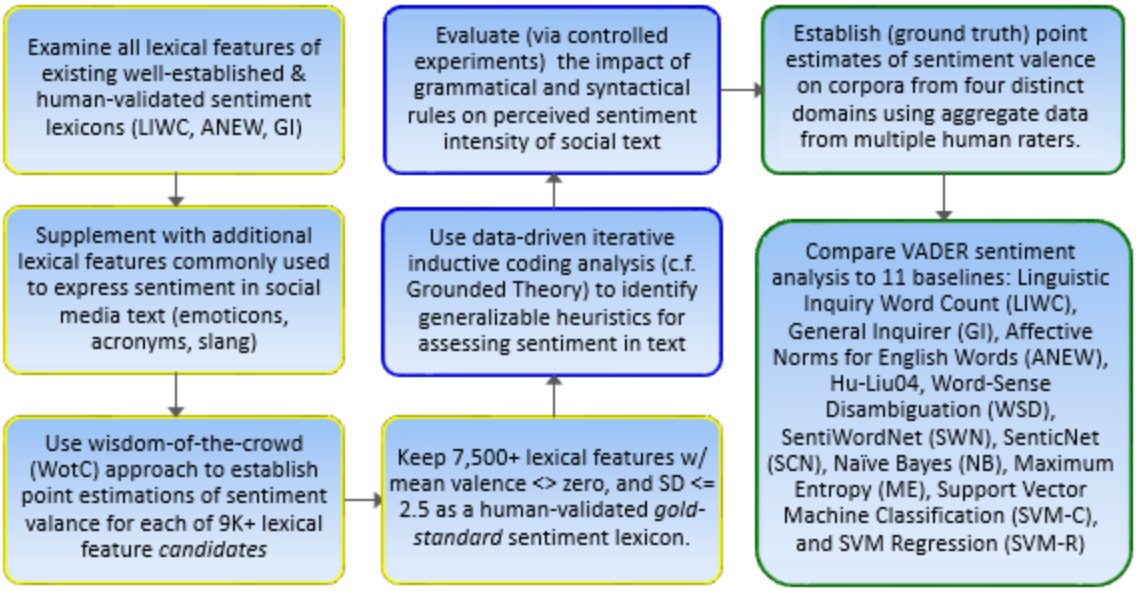


Figure 8: Process for VADER development, validation, and evaluation.

interrelated efforts: 1) the development and validation of a gold standard sentiment lexicon that is sensitive both the *polarity* and the *intensity* of sentiments expressed in social media microblogs (but which is also generally applicable to sentiment analysis in other domains); 2) the identification and subsequent experimental evaluation of generalizable rules regarding conventional uses of grammatical and syntactical aspects of text for assessing sentiment intensity; and 3) comparing the performance of a parsimonious lexicon and rule-based model against other established and/or typical sentiment analysis benchmarks. In each of these three efforts, I incorporate an explicit human-centric approach. Specifically, I combine qualitative analysis with empirical validation and experimental investigations leveraging the wisdom-of-the-crowd [208].

3.3.1 Constructing and Validating a Valence-Aware Sentiment Lexicon: A Human-Centered Approach

Manually creating (much less, validating) a comprehensive sentiment lexicon is a labor intensive and sometimes error prone process; therefore, many opinion mining researchers and practitioners rely heavily on existing lexicons as primary resources. There is, of course, a great deal of overlap in the vocabulary covered by such lexicons; however, there are also numerous items unique to each. For this effort, I begin by constructing a list inspired by examining existing well-established sentiment word-banks (LIWC, ANEW, and GI). I next incorporate numerous lexical features common to sentiment expression in social media and microblogs, including a full list of Western-style emoticons[20] (for example, ":-)" denotes a "smiley face" and generally indicates positive sentiment), sentiment-related acronyms and initialisms[21] (e.g., LOL and ROFL are both sentiment-laden initialisms), and commonly used slang[22] with sentiment value (e.g., "nah", "meh" and "giggly"). This process produces over 9,000 lexical feature *candidates*.

Next, I assessed the general applicability of each feature candidate to sentiment expressions. I used a wisdom-of-the-crowd (WotC) approach [208] to acquire a valid point estimate for the sentiment valence (intensity) of each context-free candidate feature. I collected intensity ratings on each candidate lexical features from ten independent human raters (for a total of 90,000+ ratings). Features were rated on a scale from "[–4] Extremely Negative" to "[4] Extremely Positive", with allowance for "[0] Neutral (or Neither, N/A)". Ratings were obtained using Amazon Mechanical Turk (AMT), a micro-labor website where workers perform minor tasks in exchange for a small amount of money (see

[20] http://en.wikipedia.org/wiki/List_of_emoticons#Western

[21] http://en.wikipedia.org/wiki/List_of_acronyms

[22] http://www.internetslang.com/

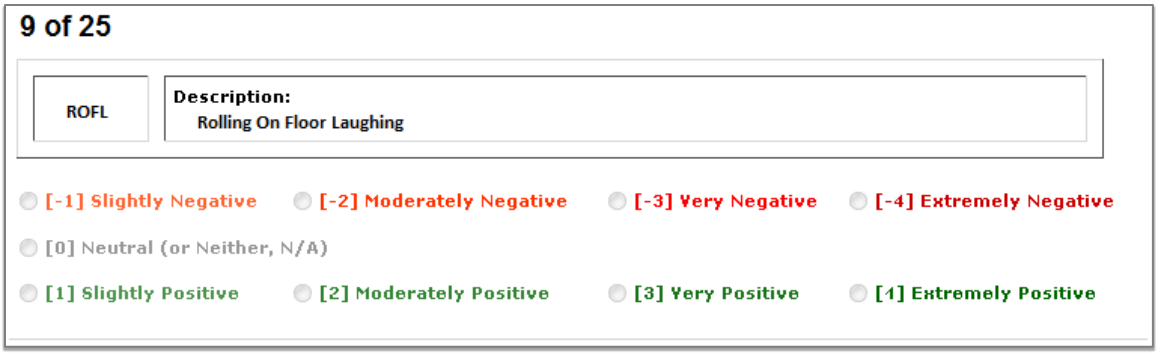


Figure 9: Example of the interface implemented for acquiring valid point estimates of sentiment valence (intensity) for each context-free candidate feature comprising the VADER sentiment lexicon. A similar UI was used for all rating activities described in sections 3.3.1–3.3.4.

subsection 3.3.1.1 for details on how I was able to consistently obtain high quality, generalizable results from AMT workers). Figure 9 illustrates the user interface implemented for acquiring valid point estimates of sentiment intensity for each context-free candidate feature comprising the VADER sentiment lexicon. (A similar UI was leveraged for all of the evaluation and validation activities described in subsections 3.3.1, 3.3.2, 3.3.3, and 3.3.4.) I kept every lexical feature that had a non-zero mean rating, and whose standard deviation was less than 2.5 as determined by the aggregate of ten independent human raters. This left just over 7,500 lexical features with validated valence scores that indicated both the sentiment *polarity* (positive/negative), and the sentiment *intensity* on a scale from –4 to +4. For example, the word "okay" has a positive valence of 0.9, "good" is 1.9, and "great" is 3.1, whereas "horrible" is –2.5, the frowning emoticon ":(" is –2.2, and "sucks" and "sux" are both –1.5. This gold standard list of features, with associated valence for each feature, comprises VADER's sentiment lexicon, and is available for download from our website[23].

[23] http://comp.social.gatech.edu/papers/

3.3.1.1 Screening, Training, Selecting, and Data Quality Checking Crowdsourced Evaluations and Validations

Previous linguistic rating experiments using a WotC approach on AMT have shown to be reliable – sometimes even outperforming expert raters [200]. On the other hand, prior work has also advised on methods to reduce the amount of noise from AMT workers who may produce poor quality work [51,121]. I therefore implemented four quality control processes to help ensure I received meaningful data from our AMT raters (the effectiveness of these methods, and others, are discussed in greater detail in Chapter 4). First, every rater was prescreened for English language reading comprehension – each rater had to individually score an 80% or higher on a standardized college-level reading comprehension test. Second, every prescreened rater then had to complete an online sentiment rating training and orientation session, and score 90% or higher for matching the known (pre-validated) mean sentiment rating of lexical items which included individual words, emoticons, acronyms, sentences, tweets, and text snippets (e.g., sentence segments, or phrases). The user interface employed during the sentiment training (Figure 9) always matched the specific sentiment rating tasks discussed in this chapter. The training helped to ensure consistency in the rating rubric used by each independent rater. Third, every batch of 25 features contained five "golden items" with a known (pre-validated) sentiment rating distribution. If a worker was more than one standard deviation away from the mean of this known distribution on three or more of the five golden items, I discarded all 25 ratings in the batch from this worker. Finally, I implemented a bonus program to incentivize and reward the highest quality work. For example, I asked workers to select the valence score that they thought "*most other people*" would choose for the given lexical feature

(early/iterative pilot testing revealed that wording the instructions in this manner garnered a much tighter standard deviation without significantly affecting the mean sentiment rating, allowing us to achieve higher quality (generalized) results while being more economical).

I compensated AMT workers $0.25 for each batch of 25 items they rated, with an additional $0.25 incentive bonus for all workers who successfully matched the group mean (within 1.5 standard deviations) on at least 20 of 25 responses in each batch. Using these four quality control methods, I achieved remarkable value in the data obtained from AMT workers, issuing bonuses for high quality to at least 90% of raters for most batches.

3.3.2 Generalizable Heuristics Humans Use to Assess Sentiment Intensity in Text

I next analyze a purposeful sample of 400 positive and 400 negative social media text snippets (tweets). I selected this sample from a larger initial set of 10K random tweets pulled from Twitter's public timeline based on their sentiment scores using the Pattern.en sentiment analysis engine[24] (they were the top 400 most positive and negative tweets in the set). Pattern is a web mining module for Python, and the Pattern.en module is a natural language processing (NLP) toolkit [45] that leverages WordNet to score sentiment according to the English adjectives used in the text.

Next, two human experts individually scrutinized all 800 tweets, and independently scored their sentiment intensity on a scale from –4 to +4. Following a data-driven inductive coding technique similar to the Grounded Theory approach [206], I next used qualitative analysis techniques to identify properties and characteristics of the text which affect the

[24] http://www.clips.ua.ac.be/pages/pattern-en#sentiment

perceived sentiment intensity of the text. This deep qualitative analysis resulted in isolating five generalizable heuristics based on grammatical and syntactical cues to convey changes to sentiment intensity. Importantly, these heuristics go beyond what would normally be captured in a typical bag-of-words computational model; some heuristics incorporate *word-order sensitive relationships* between terms:

1. Punctuation, namely the exclamation point (!), increases the magnitude of the intensity without modifying the semantic orientation. For example, "*The food here is good!!!*" is more intense than "*The food here is good.*"

2. Capitalization, specifically using ALL-CAPS to emphasize a sentiment-relevant word in the presence of other non-capitalized words, increases the magnitude of the sentiment intensity without affecting the semantic orientation. For example, "*The food here is GREAT!*" conveys more intensity than "*The food here is great!*"

3. Degree modifiers (also called *intensifiers*, *booster words*, or *degree adverbs*) impact sentiment intensity by either increasing or decreasing the intensity. For example, "*The service here is extremely good*" is more intense than "*The service here is good*", whereas "*The service here is marginally good*" reduces the intensity.

4. The contrastive conjunction "*but*" signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. "*The food here is great, but the service is horrible*" has mixed sentiment, with the latter half dictating the overall rating.

5. Inspired by [91], who showed that the best performing negation strategy was to consider a fixed window length of two words following the negation word. Initial pilot testing revealed that by examining the tri-gram preceding a sentiment-laden lexical feature, I catch nearly 90% of cases where negation flips the polarity of the text. A negated sentence would be "*The food here really isn't all that great*".

3.3.3 Controlled Experiments to Evaluate Impact of Grammatical and Syntactical Heuristics

Using the general heuristics just identified, I next selected 30 baseline tweets and manufactured six to ten variations of the exact same text, controlling the specific grammatical or syntactical feature that is presented as an independent variable in a small experiment. With all such variations, I end up with 200 contrived tweets, which I then include in a new set of 800 tweets similar to those used during the prior qualitative analysis. I next asked 30 independent AMT workers to rate the sentiment intensity of all 1000 tweets to assess the impact of these features on perceived sentiment intensity. (AMT workers were all screened, trained, and data quality checked as described in subsection 3.3.1.1). Table 4 illustrates some examples of contrived variations on a given baseline:

Table 4: Example of baseline text with eight test conditions comprised of grammatical and syntactical variations.

Test Condition	Example Text
Baseline	Yay. Another good phone interview.
Punctuation1	Yay! Another good phone interview!
Punctuation1 + Degree Mod.	Yay! Another extremely good phone interview!
Punctuation2	Yay!! Another good phone interview!!
Capitalization	YAY. Another GOOD phone interview.
Punct1 + Cap.	YAY! Another GOOD phone interview!
Punct2 + Cap.	YAY!! Another GOOD phone interview!!
Punct3 + Cap.	YAY!!! Another GOOD phone interview!!!
Punct3 + Cap. + Degree Mod.	YAY!!! Another EXTREMELY GOOD phone interview!!!

Additionally, Table 5 also depicts the *t*-test statistic, *p*-value, mean of differences for rank ordered data points between each distribution, and 95% confidence intervals. Differences in means were all statistically significant beyond the 0.001 level. I incorporated these mean differences between each distribution into VADER's rule-based model. For example, from Table 5, it is evident that for 95% of the data, using an exclamation point (relative to a period or no punctuation at all) increased the intensity by 0.261 to 0.322, with a mean difference of 0.291 on a rating scale from 1 to 4 (I use absolute value scale here for simplicity, because it did not matter whether the text was positive or negative, using an exclamation made it equally more extreme in either case).

Table 5: Statistics associated with grammatical and syntactical cues for sentiment intensity.

Test Condition	*t*	*p*	Diff.	95% C.I.
Punctuation (. vs !)	19.02	< 2.2e-16	0.291	0.261 - 0.322
Punctuation (! vs !!)	16.53	2.7e-16	0.215	0.188 - 0.241
Punctuation (!! vs !!!)	14.07	1.7e-14	0.208	0.178 - 0.239
All CAPS (w/o vs w)	28.95	< 2.2e-16	0.733	0.682 - 0.784
Deg. Mod. (w/o vs w)	9.01	6.7e-10	0.293	0.227 - 0.360

I incorporated consideration for rule 4 by splitting the text into segments around the contrastive conjunction "*but*", and diminished the total sentiment intensity of the text preceding the conjunction by 50% while increasing the sentiment intensity of the post-conjunction text by 50%.

3.3.4 Ground Truth in Multiple Domain Contexts

I next obtained gold standard (human-validated) ground truth regarding sentiment intensity on corpora representing four distinct domain contexts. For this purpose, I recruited 20 independent human raters from AMT (raters were all screened, trained, and data quality checked consistent with the process described in subsection 3.3.1.1). All four

sentiment-intensity annotated corpora are available for download from the Comp.Social website[23]:

1. Social media text: includes 4,000 tweets pulled from Twitter's public timeline (with varied times and days of posting), plus 200 contrived tweets that specifically test syntactical and grammatical conventions of conveying differences in sentiment intensity.

2. Movie reviews: includes 10,605 sentence-level snippets from rotten.tomatoes.com. The snippets were derived from an original set of 2000 movie reviews (1000 positive and 1000 negative) in Pang & Lee [162]; I used the NLTK tokenizer to segment the reviews into sentence phrases, and added sentiment intensity ratings.

3. Technical product reviews: includes 3,708 sentence-level snippets from 309 customer reviews on 5 different products. The reviews were originally used in Hu & Liu [94]; I added sentiment intensity ratings.

4. Opinion news articles: includes 5,190 sentence-level snippets from 500 New York Times opinion editorials.

3.4 Comparing VADER to Other Sentiment Analysis Benchmarks

In order to evaluate my results directly against the broader body of literature, I assess both a) the correlation of computed raw sentiment intensity rating to gold standard ground truth, i.e., the mean sentiment rating from 20 prescreened and appropriately trained human raters, as well as b) the multiclass classification metrics of *precision*, *recall*, and *F1 score* (ground truth in these cases were the binned positive, negative, and neutral gold

standard sentiment scores with thresholds set at –0.05 and +0.05). In statistical analysis of classifier performance, *precision* is the number of true classifications (i.e. the number of items labeled as a particular class that match the known gold standard classification) divided by the total number of elements labeled as that class (including both correct and incorrect classifications). *Recall* is the number of true classifications divided by the total number of elements that are known to belong to the class; low recall is an indication that known elements of a class were missed. The *F1 score* is the harmonic mean of precision and recall.

I compared the VADER sentiment lexicon to seven other well-established sentiment analysis lexicons: Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), Word-Sense Disambiguation (WSD) using WordNet, and the Hu-Liu04 opinion lexicon. For fairness to each lexicon, *all comparisons utilized VADER's rule-based model for processing syntactical and grammatical cues* – the only difference were the features represented within the actual lexicons themselves. As Figure 10 and Table 6 both show, the VADER lexicon performs exceptionally well in the social media domain, and generalizes favorably to sentence level text from movie reviews, product reviews, and news editorials.

The Pearson Product Moment Correlation Coefficients in Table 6 show that VADER (r = 0.881) performs as well as individual human raters (r = 0.888) at matching ground truth (aggregated group mean from 20 human raters for sentiment intensity of each tweet).

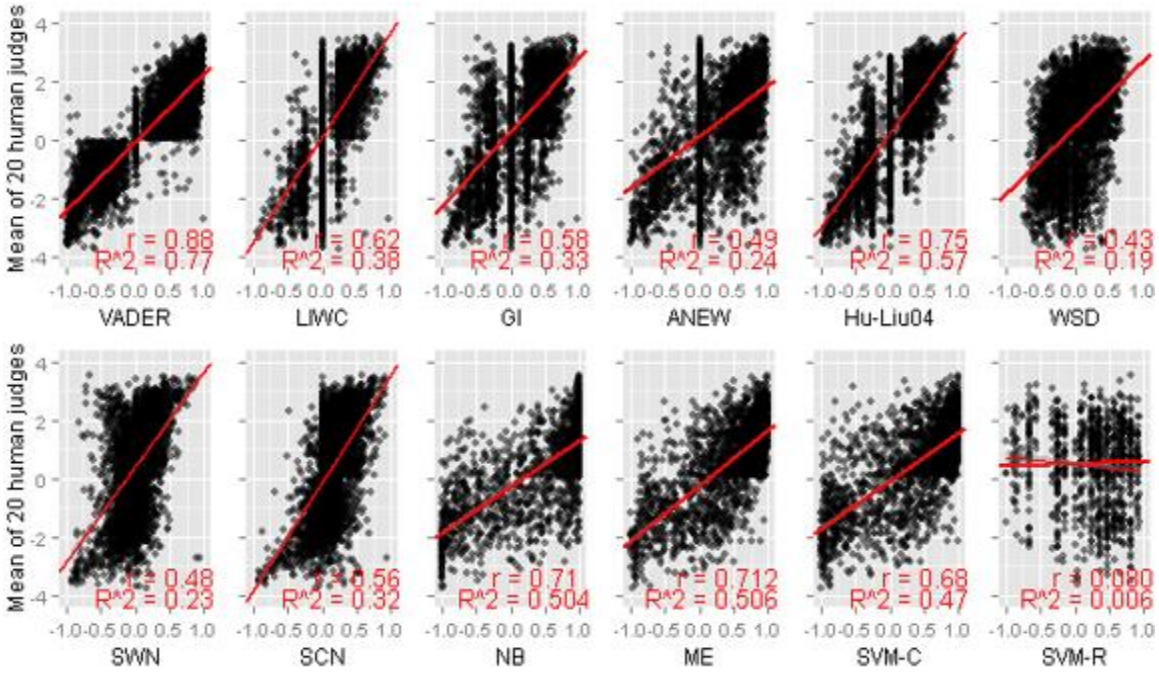


Figure 10: Sentiment scores from VADER and 11 other highly regarded sentiment analysis tools/techniques on a corpus of over 4K tweets. Although this figure specifically portrays correlation, it also helps to visually depict (and contrast) VADER's classification precision, recall, and F1 accuracy within this domain (see Table 6). Each subplot can be roughly considered as having four quadrants: true negatives (lower left), true positives (upper right), false negatives (upper left), and false positives (lower right).

Surprisingly, when I further inspect the classification accuracy (with classification thresholds set at –0.05 and +0.05 for all normalized sentiment scores between -1 and 1), I find that VADER (*F1* = 0.96) actually outperforms individual human raters (*F1* = 0.84) at correctly classifying the sentiment of tweets. Notice how the LIWC, GI, ANEW, and Hu-liu04 results in Figure 10 show a concentration of tweets incorrectly classified as neutral. Presumably, this is due to lack of coverage for the sentiment-oriented language of social media text, which is often expressed using emoticons, slang, or abbreviated text such as acronyms and initialisms.

Table 6: VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, opinion news, product reviews).

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Metrics		
		Overall Precision	Overall Recall	Overall F1				Overall Precision	Overall Recall	Overall F1
Social Media Text (4,200 Tweets)							**Movie Reviews (10,605 sentences)**			
Ind. Humans	**0.888**	0.95	0.76	0.84	**2**	**1**	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	**1***	**2**	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	**3**	**3**	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	**4**	**7**	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	**5**	**5**	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	**6**	**4**	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	**7**	**9**	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	**8**	**8**	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	**9**	**6**	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 sentences)							**NY Times Editorials (5,190 sentences)**			
Ind. Humans	0.911	0.94	0.80	0.85	**1**	**1**	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	**2**	**2**	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	**3**	**3**	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	**7**	**7**	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	**5**	**5**	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	**4**	**4**	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	**9**	**9**	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	**8**	**8**	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	**6**	**6**	0.218	0.55	0.45	0.47

The lexicons for the machine learning algorithms were all constructed by training those models on half the data (again, incorporating all rules), with the other half being held out for testing. While some algorithms performed decently on test data from the specific domain for which it was expressly trained, *they do not significantly outperform the simple model we use*. Indeed, in three out of four cases, VADER performs as well or better *across* domains than the machine learning approaches do in the *same* domain for which they were trained. Table 7 explicitly shows this, and also highlights another advantage of VADER – its simplicity makes it computationally efficient, unlike some SVM models, which were unable to fully process the data from the larger corpora (movie reviews and NYT editorials)

Table 7: Three-class performance (F1 scores) for each machine trained model (and the corpus it was trained on) as tested against every other domain context. (Note: SVM models for the movie and NYT data were too intensive for my multicore CPU with 94GB RAM).

	3-Class Classification Accuracy (F1 scores) Test Sets			
	Tweets	Movie	Amazon	NYT
VADER	**0.96**	0.61	**0.63**	**0.55**
NB (tweets)	0.84	0.53	0.53	0.42
ME (tweets)	0.83	0.56	0.58	0.45
SVM-C (tweets)	0.83	0.56	0.55	0.46
SVM-R (tweets)	0.65	0.49	0.51	0.46
NB (movie)	0.56	**0.75**	0.49	0.44
ME (movie)	0.56	**0.75**	0.51	0.45
NB (amazon)	0.69	0.55	0.61	0.48
ME (amazon)	0.67	0.55	0.60	0.43
SVM-C (amazon)	0.64	0.55	0.58	0.42
SVM-R (amazon)	0.54	0.49	0.48	0.44
NB (nyt)	0.59	0.56	0.51	0.49
ME (nyt)	0.58	0.55	0.51	0.50

even on a multicore system with large RAM (i.e., the system encounter memory errors without completing the machine learning computation):As discussed in subsections 3.3.2 and 3.3.3, I identified and quantified the impact of several generalizable heuristics that humans use when distinguishing between degrees of sentiment intensity. By incorporating these heuristics into VADER's rule-based model, I drastically improved both the correlation to ground truth as well as the classification F1 score of the sentiment analysis engine. Importantly, these improvements are realized *independent of the lexicon or ML model that was used*. That is, when I fairly apply the rules to all lexicons and ML algorithms, I achieve stronger correlation coefficients (mean *r* increase of 5.2%) and better accuracies (mean F1 increase of 2.1%). Consistent with prior work [2,42,196], I find that grammatical features (conventions of use for punctuation and capitalization) and consideration for degree modifiers like "very" or "extremely" prove to be useful cues for distinguishing differences in sentiment intensity. Other word-order sensitive considerations identified via qualitative analysis (e.g., negation, idioms, and contrastive

conjunctions) also help make VADER successful, and is consistent with prior work [2,47,144,201].

Recent work by Socher et. al [201] does an excellent job of summarizing (and pushing) the current state of the art for fine-grained sentence-level sentiment analysis by supervised machine learning models. As part of their work using recursive deep models for assessing semantic compositionality over a sentiment treebank[25], they report that the state of the art regarding accuracy for simple binary (positive/negative) classification on single sentences is around 80%, and that for the more difficult multiclass case that includes a third (neutral) class, accuracies tend to hover in the 60% range for social media text (c.f. [2] and [214]). I find it very encouraging, therefore, to report that the results from VADER's simple rule-based approach are on par with such sophisticated benchmarks. However, when compared to sophisticated machine learning techniques, the simplicity of VADER also carries several advantages. First, it's computationally economy helps to make the analysis faster *without sacrificing F1 score performance*. For example, running directly from a standard modern laptop computer with typical, moderate specifications (e.g., 3GHz processor and 6GB RAM), a corpus that takes a fraction of a second to analyze with VADER can take hours when using more complex models like SVM (if training is required) or tens of minutes if the model has been previously trained. Second, the lexicon and rules used by VADER are directly accessible, not hidden within a machine-access-only black-box. VADER is therefore easily inspected, understood, extended or modified. By exposing both the lexicon and rule-based model, VADER makes the inner workings of

[25] A treebank is a text corpus that annotates linguistic sentence structures of interest such as syntax, semantics, or in this case, sentiment. The resulting "databank" of parsed annotations takes the form of a tree.

the sentiment analysis engine more accessible (and thus, more interpretable) to a broader human audience beyond the computer science community. Sociologists, psychologists, marketing researchers, or linguists who are comfortable using LIWC should also be able to use VADER. Third, by utilizing a *general* (human-validated) sentiment lexicon and *general* rules related to grammar and syntax, VADER is at once both self-contained and domain agnostic – it does not require an extensive set of additional training data, yet it performs well in diverse domains. I stress that in no way do I intend to convey that complex or sophisticated techniques are in any wrong or bad. Instead I show that a simple, human-centric, interpretable, computationally efficient approach can produce high quality results – even outperforming individual human raters.

3.5 Chapter Summary

In this chapter, I report the systematic development, validation, and evaluation of VADER (Valence Aware Dictionary for sEntiment Reasoning). Using a combination of qualitative and quantitative methods, I construct and empirically validate a *gold-standard* list of lexical features (along with their associated sentiment intensity measures) which are specifically attuned to sentiment in social media microblog-like contexts. I then combine these lexical features with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. The results are not only encouraging – they are indeed quite remarkable; VADER performs as well as (and in many cases, *better than*) eleven other highly regarded sentiment analysis tools.

These results highlight the gains to be made in computer science when the human is incorporated as a central part of the development process of computational models. In the next chapter, I formalize and generalize the crowdsourcing methods introduced in this chapter for conducting large scale qualitative data analysis and computational model verification, validation, and evaluations (VV&E). I then empirically assess various strategies for addressing the challenges associated with the crowdsourcing approach. I will use the VADER lexicon and its sentiment analysis rules to help inform a model to investigate bias in news stories in Chapter 5.

CHAPTER 4. LARGE SCALE HUMAN VALIDATION, EVALUATION, AND QDA

4.1 Chapter Overview

The emergence of crowdsourced micro labor markets like Amazon Mechanical Turk (AMT) is attractive for behavioral and empirical researchers who wish to acquire large-scale independent human judgments, without the burden of intensive recruitment effort or administration costs. Yet consistently acquiring well-measured *high quality* judgments using an online workforce is often seen as a challenge [87,98,185,197,204]. This has led to scholarly work suggesting quality control measures to address the problem of noisy data [52,121,148,197]. Many of these studies have investigated the effectiveness of various quality-control measures as stand-alone intervention strategies on one-off tasks. How do these measures affect quality when working in tandem? What are the challenges faced in acquiring quality results when the difficulty of subjective judgments increase? The study presented in this chapter addresses these questions.

Building on some of the most promising strategies identified by prior work (e.g., [197]), we[26] design and conduct a large empirical study to compare the relative impacts and interactions of 34 intervention strategies. Specifically, we collected and analyzed 68,000 human annotations across more than 280 pairwise statistical comparisons for

[26] It is worth noting that while the strategies and tasks associated with Experiment 1 constitute my own principal contributions, this was a cooperative study with my colleague Tanushree Mitra, who was primarily responsible for the strategies and tasks associated with Experiment 2. We compared and contrasted the strategies across experiments, and jointly published the work with equitable division of labor and intellectual input. My references to "we" throughout this chapter reflect the collaborative nature of this research effort.

strategies related to worker screening and selection, interpretive convergence modeling, social motivations, financial incentives, and hybrid combinations. Further, we compare these interactions against a range of representative *subjective judgment-oriented* qualitative coding activities of varying difficulty. Our study makes four principal contributions:

- We reveal several intervention strategies which have a substantial positive effect on the quality of data annotations produced by non-experts, regardless of whether "correctness" is defined by agreement with the most frequent annotation or as agreement with an accepted expert.

- We find that *person-oriented intervention strategies* tend to facilitate high-quality data coding among non-experts. For example, borrowing analogous concepts from the field of Qualitative Data Analysis (QDA) and adapting them for use by a massive, distributed, untrained, transient, anonymous workforce, we find that *prescreening workers for requisite aptitudes, and providing rudimentary training* in collaborative qualitative coding techniques results in improved agreement and interpretive convergence of non-expert workers.

- We find that person-oriented strategies improve the quality of non-expert data coders above and beyond those achieved via process-oriented strategies like the Bayesian Truth Serum (BTS) technique (c.f., [182,197]).

- Finally, of particular importance for contemporary AMT researchers, we note that while our results show significant improvements in the quality of data annotation tasks over control and baseline conditions, the baseline quality has improved in recent years. In short, compared to the control-level accuracies of just a few years

ago [197], the quality of QDA on AMT is improving even before researcher initiated interventions.

4.2 Qualitative Coding, Annotations, and Content Analysis

Qualitative Data Analysis (QDA)—that is, systematically analyzing non-numeric data such as interview transcripts, open-ended survey responses, field notes/observations, and a wide range of text documents, images, video, or audio data—is generally a specialized skill most often acquired through formal education or training. Such skills are costly, both in terms of the financial demand required to obtain the skillset (in undergraduate or graduate school, for example), and in terms of the time, labor, and expense needed to employ the skills. *Qualitative coding*, or the process of interpreting, analyzing, classifying, and labeling qualitative data (e.g., with themes, categories, concepts, observations, attributes or degree anchors, etc.) is a critical step in the larger overall QDA process. As part of qualitative data analysis, many lead researchers employ multiple skilled qualitative *coders* (individuals who perform QDA annotations), each working independently on the same data. Such a strategy makes an explicit trade-off for labor and expense for an increase in accuracy, higher reliability/consistency, and a reduction in potential coding errors. What if we could rapidly, inexpensively, and yet *reliably* obtain *high-quality* content analyses and annotations from a massive, distributed, untrained, anonymous, transient labor force like AMT?

4.2.1 Crowdsourcing Qualitative Coding & Content Analysis

Crowdsourced labor markets are an attractive resource for researchers whose studies are conducive to online (Internet-based) participation. Research study data such as

qualitative content analysis can be obtained relatively cheaply from potentially thousands of human coders in a very short time. For example, researchers have asked workers to: code discussion forum messages for whether they offered information or provided emotional support [216], annotate images to locate people [204], interpret the intensity of sentiments in various textual domains [98], mark the degree of factuality for statements reported by journalists and bloggers [203], and extract thematic categories for messages shared amid Wikipedians [7].

Clearly, crowdsourcing does enable quick, inexpensive content analysis and data coding at large scales (c.f., [7,98,203,204,216]). However, these types of QDA activities are often quite subjective in nature. As such, they are susceptible to conflicting interpretations, dissimilar rubrics used for subjective judgments, different levels of (mis)understanding the instructions for the task, or even opportunistic exploitation/gaming to maximize payouts while minimizing effort. Unfortunately, worker anonymity, lack of accountability, inherent workforce transience, and fast cash disbursements can entice the online labor workforce to trade speed for quality [52]. Consequently, the collected annotations may be noisy and poor in quality. Moreover, quality can be inconsistent across different kinds of coding tasks of varying difficulty [197]. Scholars using AMT must therefore carefully consider strategies for ensuring that the codes and annotations produced by non-experts are consistently of *high quality*—that is, ensuring that the coding produced by anonymous workers is accurate and reliable [87,98,185,197,204]. Previous research suggests several quality control measures to tackle the problem of noisy data [7,52,96,121,133,148,197,199]. Most of these earlier works, in isolation, investigate a select set of specialized interventions, often for a single (or just a few kinds of) coding or

annotation tasks. Many studies also do not address the challenges associated with coding *subjective judgment* oriented tasks of varying difficulty. To address these gaps, we design and conduct a large empirical study to compare the relative impacts and interactions of numerous intervention strategies (including over 280 pairwise statistical comparisons of strategies related to worker screening and selection, interpretive convergence modeling, social motivation, financial incentives, and hybrid combinations – we discuss these strategies in greater detail later). Further, we compare these interactions against a range of qualitative data coding activities that have varying degrees of difficulty for the subjective interpretations required.

4.2.2 Crowdsourcing Data Annotations for Machine Learning

Interest in high-quality human annotation is not limited to qualitative method researchers. Machine learning scholars also benefit from access to large-scale, inexpensive, human intelligence for classifying, labeling, interpreting, or otherwise annotating an assorted variety of "training" datasets. Indeed, human-annotated training data acquisition is a fundamental step towards building many learning and prediction models, albeit an expensive and time-consuming step. Here again, the emergence of micro-labor markets has provided a feasible alternative for acquiring large quantities of manual annotations at relatively low cost and within a short period of time—along with several researchers investigating ways to improve the quality of the annotations from inexpert raters [107,198,200]. For example, Snow and colleagues [200] evaluate non-expert annotations for a natural language processing task; they determined how many AMT worker responses were needed to achieve expert-level accuracy. Similarly, Sheng and colleagues [198] showed that using the most commonly selected annotation category from multiple AMT

workers as training input to a machine learning classifier improved the classifier's accuracy in over a dozen different datasets. Ipeirotis, Provost, & Wang [108] use more sophisticated algorithms, which account for both per-item classification error and per-worker biases, to help manage data quality subsequent to data annotation.

Whereas these studies concentrate heavily on post-hoc techniques for identifying and filtering out low quality judgments from inexpert coders *subsequent* to data collection, we follow in the same vein as Shaw et al. [197] and focus on *a priori* techniques for encouraging workers to provide attentive, carefully considered responses in the first place. Along with the most promising strategies identified by Shaw et al. [197], we add numerous other person-centered and process-centered strategies for facilitating high quality data coding from non-experts across a range of annotation tasks. We describe these strategies in the next section.

4.3 Strategies for Eliciting Consistently High Quality Data

In this section, we consider four challenges that affect the quality of crowd annotated data, and discuss strategies to mitigate issues associated with these challenges.

4.3.1 Challenge 1 – Undisclosed Aptitudes

Certain tasks may require workers to have special knowledge, skills or abilities, the lack of which can result in lower quality work despite spending considerable time and effort on a task [117]. As in offline workforces, some workers are better suited for particular tasks than others. Asking anonymous workers with unidentifiable backgrounds

to perform activities without first verifying that the worker possesses a required aptitude may result in imprecise or speculative responses, which negatively impacts quality.

4.3.2 Strategy 1 – Screen Workers for Targeted Knowledge, Skills, or Abilities

On AMT, requesters often screen workers from performing certain Human Intelligence Tasks (HITs) unless they meet certain criteria. One very common screening tactic is to restrict participation to workers with an established reputation – e.g., by requiring workers to have already completed a minimum number of HITs (to reduce errors from novices who are unfamiliar with the system or process), and have approval ratings above a certain threshold (e.g., 95%) [14,147,171]. This approach has the benefit of being straightforward and easy for requesters to implement, but it is naive in that it does not explicitly attempt to verify or confirm that a worker actually has the requisite aptitude for performing a given task. For example, a more targeted screening activity (that is tailored more to content analysis coding or linguistic labeling tasks) would be to require workers to have a good understanding of the language of interest, or to require workers to reside in certain countries so that they are more likely to be familiar with localized social norms, customs, and colloquial expressions [98,203].

4.3.3 Challenge 2 – Subjective Interpretation Disparity

Qualitative content analysis can often be very subjective in nature, and is therefore vulnerable to differences in interpretations, dissimilar rubrics used for judgments, and different levels of (mis)understanding the instructions for the task by unfamiliar, non-expert workers.

4.3.4 Strategy 2 – Convergence Modeling (Provide Examples and Train Workers)

Providing examples to introduce workers to a particular coding or annotation task, and modeling or demonstrating the preferred coding/annotation behaviors can help workers establish consistent rubrics (criteria and standards) for judgment decisions [225]. This is analogous to qualitative researchers sharing a common "codebook"—the compilation of codes, their content descriptions and definitions, guidelines for when the codes apply and why, and brief data examples for reference [194]. Along with the examples, requesters on AMT can then require workers to obtain a specific qualification which assesses the degree to which the worker understands how to perform the *task-specific* content analysis annotation or labeling activity. Guiding workers through the process of doing the task trains them and calibrates their coding decisions to the nature of desired responses. This strategy helps improve intercoder/interrater agreement, or *interpretive convergence* – i.e., the degree to which coders agree and remain consistent with their assignment of particular codes to particular data [194].

4.3.5 Challenge 3 –Existing Financial Incentive is to Minimize Time-on-Tasks

The micro-labor market environment financially rewards those who work quickly through as many micro-tasks as possible. Consequently, there is little incentive to spend time and effort in providing thoughtfully considered quality responses. If unconsidered judgments and random, arbitrary clicking will pay just as well as thoughtful, carefully considered responses, then some people may attempt to maximize their earnings while minimizing their effort.

4.3.6 Strategy 3 – Financially Incentivize Workers to Focus on Quality

In an effort to incentivize carefully considered responses, rewarding high quality responses has shown to improve annotation accuracy [197]. For every intervention strategy we examine, we include both a non-incentivized and an incentivized group, and we confirm whether financial incentives continue to have significant impacts above and beyond those of a particular intervention strategy.

4.3.7 Challenge 4 – Low Independent (Individual) Agreement

There are several ways to measure the accuracy of any individual coder. A simple approach is to calculate a percent correct for codes produced by a given coder against an accepted "ground truth." Other useful metrics are Cohen's kappa statistics for nominal coding data and Pearson's correlation for ordinal or interval scales. Regardless of how accuracy is measured, the correctness of any individual coder is often less than perfect due to differences in subjective interpretations.

4.3.8 Strategy 4 – Aggregate, Iteratively Filter, or Both

One way to mitigate the problem is to use aggregated data, or by searching for congruent responses by taking advantage of the wisdom-of-the-crowd[27] and accepting only the majority agreement from multiple independent workers [208]. However, it is often still difficult to obtain meaningful (or at least interpretable) results when aggregated responses are noisy, or when large variance among worker judgments challenge the notion of majority agreement [207]. Prior research has addressed this challenge by adding iterative

[27] *Wisdom-of-the-crowd* is the process of incorporating aggregated opinions from a collection of individuals to answer a question. The process has been found to be as good as (often better than) estimates from lone individuals, even experts.

steps to the basic parallel process of collecting multiple judgments [11, 22]. In other words, use crowd-workers to scrutinize the responses of other workers, thereby allowing human judges (as opposed to statistical or computational processes) to identify the best quality annotations [140,147].

4.4 Qualitative Data Analysis and Annotation Tasks

In order to establish a framework of strategies for obtaining high quality labeled data, we administered a combination of the above described strategies across four sets of labeling tasks: identifying the approximate number of people in a picture, sentiment analysis, word intrusion, and credibility assessments (we describe these in more depth in a moment). Each of the analysis/annotation tasks is intended to vary in its level of difficulty for subjective judgments by individual coders. To verify this, we deployed four HITs on AMT (one HIT for each type of task), and used a modified version of the NASA-TLX workload inventory scale to assess difficulty [86]. Response options ranged from "Very Low" to "Very High" on a seven-point scale. Figure 11 shows an example of the subjective difficulty data collection user interface:

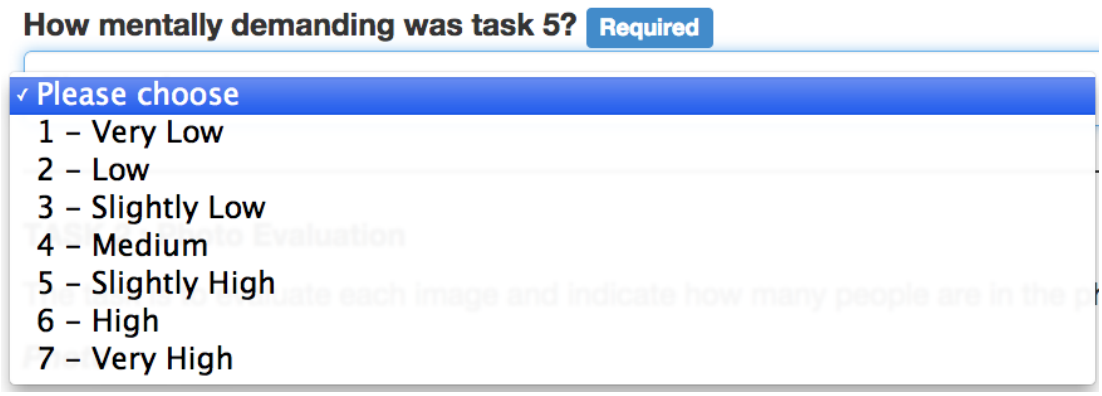


Figure 11: Example of the subjective judgment difficulty user interface.

Each HIT asked 20 workers to perform the four data coding tasks described below, and paid $0.75 per HIT. To account for item effects, we used different content for each annotation task in each of the four HITs. Also, to account for ordering effects, we randomized the order in which the task types were presented. Thus, we collected a total of 80 responses regarding the difficulty of each type of subjective judgment task, providing us with a verified set of tasks with a range of underlying subjective judgment difficulty. Table 8 shows the median subjective judgment difficulty for each task type:

Table 8: Median subjective judgment difficulty for each task type.

Task Name	Abbreviation	Subjective Judgement Difficulty (median)
People in Pictures	PP	1
Sentiment Analysis	SA	2
Word Intrusion	WI	2
Credibility Assessments	CA	3

4.4.1 Task 1: People in Pictures (PP), Median Difficulty = 1

In this task, we presented workers with an image and asked them to estimate the number of people shown in the picture. This is a well-known data annotation activity in the computer vision research area [38,165]. We selected 50 images from the Creative

Figure 12: Example pictures for three of the five possible data annotation categories.

Commons on Flickr[28]. The number of people in each image differed by orders of magnitude, and corresponded to one of five levels: None, About 2 – 7 people, About 20 – 70 people, About 200 – 700 people, and More than 2,000 people.

Expert Annotation / Ground Truth – We determined ground truth at the time we selected the image from Flickr. We purposefully selected images based on a stratified sampling technique such that exactly ten pictures were chosen for each coding/annotation category.

4.4.2 Task 2: Sentiment Analysis (SA), Median Difficulty = 2

In this task, we mimic a sentiment intensity rating annotation task similar to the one presented in [98] whereby we presented workers with short social media texts (tweets) and asked them to annotate the degree of positive or negative sentiment intensity of the text. We selected 50 random tweets from the public dataset provided by [98]; however, we reduced the range of rating options from nine (a scale from –4 to +4) down to five (a scale from –2 to +2), so that we maintain consistent levels of chance for coding the correct annotations across all our subjective judgment tasks.

[28] https://www.flickr.com/creativecommons/by-2.0/

Figure 13: Example of the sentiment analysis annotation task.

Expert Annotation / Ground Truth – We derived ground truth from the validated "gold standard" public dataset provided by [98], and adjusted by simple binning into a five point annotation scale (rather than the original nine point scale). We then manually verified each transformed sentiment rating's categorization into one of the five coding/annotation category options.

4.4.3 Task 3: Word Intrusion (WI), Median Difficulty = 2

In this task, we mimic a human data annotation task that is devised to measure the semantic cohesiveness of computational topic models [30]. We presented workers with 50 "topics" (lists of words produced by a computational Latent Dirichlet Allocation (LDA) process [17]) created from a collection of 20,000 randomly selected English Wikipedia articles. LDA is a popular unsupervised probabilistic topic modeling technique which originated from the machine learning community. The topics generated by LDA are a set of related words that tend to co-occur in related documents. Following the same procedure described in [30], we inserted an "intruder word" into each of the 50 LDA topics, and asked workers to identify the word that did not belong.

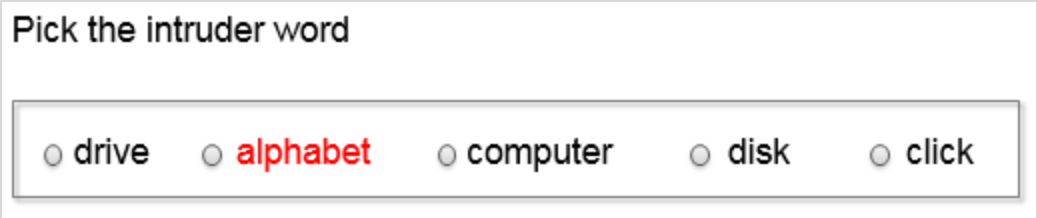


Figure 14: Example of a topic list (with the intruder word highlighted with red text for illustration purposes).

Expert Annotation / Ground Truth – A computational process (rather than a human) selected the intruder word for each topic, making this data annotation task unique among the others in that coders are asked to help establish "ground truth" for the word that least belongs. As such, there was no "expert" other than the LDA computational topic model. In this case, human AMT workers are performing verification, validation, and evaluation (VV&E) of the computational topic model output.

4.4.4 Task 4: Credibility Assessment (CA), Median Difficulty = 3

In this task, we asked workers to read a tweet, rate its credibility level and provide a reason for their rating. This task aligns with scholarly work done on credibility annotations in social media [29,152,183]. To build a dataset of annotation items that closely resembles real-world information credibility needs, we have to make sure that the dataset contains information sharing tweets, specifically those mentioning real world event occurrences [155]. We borrowed existing computational approaches to filter event specific tweets from the continuous 1% sample of tweets provided by the Twitter Streaming API [29,134,233]. Next, we recruited independent human annotators to decide whether a tweet is truly about an event, filtering out false positives in the process. After training the annotators to perform the task, if 8 out of 10 workers agree that a tweet is an event, we add the tweet as a potential candidate for credibility assessment. Next, the first author manually

Figure 15: Example of a tweet along with the five credibility coding/annotation categories modeled according to existing work on credibility annotation categories.

inspected the filtered list to verify the results of the filtering step before sending tweets for credibility assessments on AMT.

Expert Annotation / Ground Truth – Fact-checking services have successfully employed librarians to provide expert information [127]. We recruited three librarians from a large university library as our expert raters. The web interface used to administer the annotation questions to the librarians was similar to the one shown to AMT workers.

4.5 Empirical Evaluation of Intervention Strategies by Tasks

A full factorial design to evaluate all strategies across all coding/annotation tasks results in combinatorial explosion, making a full factorial experiment intractable. We therefore evaluate the strategies across tasks in stages. A total of 34 combinations were explored. We recruited non-expert content analysis / qualitative data coders from Amazon Mechanical Turk, and employed a between-group experimental design to ensure we had 40 unique workers in each intervention strategy test condition (i.e., workers were prevented from performing the same data coding activity under different intervention strategies). In each test condition, we asked workers to make analysis/annotation decisions for 50 different items (i.e., judgments of the number of people in pictures, sentiments of tweets, intruder words, or credibility assessments). Thus a total of 68,000 analysis annotations were collected (50 items * 40 annotations * 34 intervention strategy combinations).

In the design of our HITs, we leverage insights from [7], who find that presenting workers with context (by having them perform multiple classifications at a time) is highly effective. To ensure workers on an average spend equal time (~ 2-5 minutes) on each HIT independent of task type, a pilot test determined the number of items to fix per HIT.

4.5.1 Comparative Measures of "Correctness" for Subjective Judgments

We establish two measures of correctness to judge the quality of annotation in each task: (1) Accuracy compared to crowd (*Worker-to-Crowd*) and (2) Accuracy compared to experts (*Worker-to-Expert*). While the first counts the number of workers who match the most commonly selected response (i.e., the mode) of the crowd, the second counts the number of workers who match the mode of experts. We purposely choose mode over other

measures of central tendency to establish a strictly conservative comparison metric which can be applied consistently across all comparisons.

4.5.2 Statistical Analysis Overview

For all our experimental conditions we calculate the proportion of correct responses using both metrics, and conduct χ^2 tests of independence to determine whether these proportions differ across experimental conditions. Next, as a post-hoc test, we investigate the cell-wise residuals by performing all possible pairwise comparisons. Because simultaneous comparisons are prone to increased probability of Type 1 error, we apply Bonferroni corrections to counteract the problem of multiple comparisons. Pairwise comparison tests with Bonferroni correction allow researchers to do rigorous post hoc tests following a statistically significant Chi-square omnibus test, while at the same time controlling the familywise error rate [145,191].

4.5.3 Experiments

We next present two experiments. At a high level, the first experiment looks at the application of less-complex, *person-centric* a priori strategies on the three easiest subjective judgment tasks. In Experiment 2, we compare the "winner" from Experiment 1 against more complex, *process-oriented* a priori strategies such as Bayesian Truth Serum, social competition, and iterative filtering where subsequent workers judge the quality of prior workers' content analysis annotations.

4.5.3.1 Experiment 1: Design (Strategies 1-3; Tasks 1-3)

The experimental manipulations we introduce in Experiment 1 consist of variations of intervention strategies 1 through 3, described previously in subsection 4.3, as well as a control condition that involves no intervention or incentives beyond the payment offered for completing the HIT. We next describe all control and treatment conditions.

1. **Control condition, no bonus *(Control NB*)**: Workers were presented with simple instructions for completing the qualitative data analysis/annotation task. No workers were screened, trained, or offered a financial incentive for high-quality annotations. "NB" stands for **N**o **B**onus.

2. **Financial incentive only *(Control Bonus - M):*** Workers were shown the same instructions and data items as the control condition, and were also told that if they closely matched the most commonly selected annotation from 39 other workers, they would be given a financial bonus equaling the payment of the HIT (essentially, doubling the pay rate for workers whose deliberated responses matched the wisdom of the crowd majority). "Bonus-M" refers to bonus based on **M**ajority consensus.

3. **Baseline screening *(Baseline NB)***: Screening AMT workers according to their experience and established reputation (e.g., experience with more than 100 HITs and 95% approval ratings) is a common practice among scholars using AMT [7,37,87,164]. We include such a condition as a conservative baseline standard for comparison. Many researchers are concerned with acquiring high quality data coding/annotations, but if intervention strategies like *targeted screening for aptitude* or *task-specific training* do not substantially improve coding quality above

such baseline screening techniques, then implementing the more targeted strategies may not be worth the requester's extra effort.

4. **Baseline w/ financial incentive *(Baseline Bonus - M)***: Workers were screened using the same baseline experience and reputation criteria, and were also offered the financial incentive described above for matching the wisdom of the crowd majority.

5. **Targeted screening for aptitude *(Screen Only NB)***: Prior to working on the data annotation HITs, workers were screened for their ability to pass a short standardized English reading comprehension qualification. The qualification presented the prospective worker with a paragraph of text written at an undergraduate college reading-level, and asked five questions to gauge their reading comprehension. Workers had to get 4 of the 5 questions correct to qualify for the annotation HITs.

6. **Targeted screening with financial incentive *(Screen Bonus - M)***: Workers were screened using the same targeted reading comprehension technique, and they were also offered the financial incentive for matching the majority when they performed the HIT.

7. **Task-specific annotation training *(Train Only - NB)***: In comments on future work, Andre et al. [7] suggest that future research should investigate the value of training workers for specific QDA coding tasks. Lasecki et al. [133] also advocate training workers on QDA coding prior to performing the work. Therefore, prior to working on our data annotation HITs, workers in this intervention condition were required to pass a qualification which demonstrated (via several examples and

descriptions) the *task-specific* content analysis rubrics and heuristics. We then assessed workers for how well they understood the specific analysis/annotation activity; they had to get 8 of 10 annotations correct to qualify.

8. **Task-specific annotation training with financial incentive *(Train Bonus - M)***: Workers were qualified using the same task-specific demonstration and training technique, and they were also offered the financial incentive for matching the majority consensus.

9. **Screening and training *(Screen + Train NB)*** – This intervention strategy combined the targeted screening technique with the task-specific training technique (i.e., workers had to pass both qualifications to qualify for the data analysis and annotation HITs).

10. **Screening, training, and financial incentive based on majority matching *(Screen + Train + Bonus - M)***: Prior to working on the data annotation HITs, workers had to pass both qualifications, and were also offered the financial incentive for matching the majority.

Table 9 summarizes the control and treatment conditions used for Experiment 1 (described above), and previews the test conditions for Experiment 2 (described later).

Table 9: Combinatorial space of experiments - Four task types varying in median subjective judgment difficulty (People in Pictures, Sentiment Analysis, Word Intrusion, Credibility Assessment), two classes of Incentives (NB - No Bonus, Bonus), three types of three types of bonus incentive (M – Majority Consensus, B – BTS, C – Competition), six intervention strategies (Control, Baseline, Screen, Train, Both, Iterative Filtering). A total of 34 combinations were explored (marked ✓).

		Subjective Judgement Tasks															
		People in Pictures (PP)				Sentiment Analysis (SA)				Word Intrusion (WI)				Credibility Assess (CA)			
		Median Difficulty = 1				*Median Difficulty = 2*				*Median Difficulty = 2*				*Median Difficulty = 3*			
	Incentive	**NB**	**Basis of Bonus**			**NB**	**Basis of Bonus**			**NB**	**Basis of Bonus**			**NB**	**Basis of Bonus**		
			M	**B**	**C**		**M**	**B**	**C**		**M**	**B**	**C**		**M**	**B**	**C**
Intervention	Control	✓	✓			✓	✓			✓	✓						
	Baseline	✓	✓			✓	✓			✓	✓						
	Screen	✓	✓			✓	✓			✓	✓						
	Train	✓	✓			✓	✓			✓	✓						
	Both (S+T)	✓	✓			✓	✓			✓	✓				✓	✓	✓
	Iteration													✓			

4.5.3.2 Experiment 1: Results

Table 10 shows that intervention strategies have a significant impact on the number of "correct" data annotations produced by non-experts on AMT, regardless of whether "correct" is defined by worker agreement with the most commonly selected annotation code from the crowd, or as agreement with an accepted expert. For example, from Table 10 we can see that comparing the count of correct annotations to the **crowd**, the χ^2 statistic

Table 10: χ^2 tests of independence for Experiment 1.

Accuracy Metric	Task	df	N	χ^2	p
Worker-to-Crowd	All	9	59,375	388.86	$< 10^{-15}$
Worker-to-Expert	All	9	59,375	149.12	$< 10^{-15}$
Worker-to-Crowd	PP	9	20,000	345.73	$< 10^{-15}$
Worker-to-Expert	PP	9	20,000	46.66	$< 10^{-15}$
Worker-to-Crowd	SA	9	19,675	185.49	$< 10^{-15}$
Worker-to-Expert	SA	9	19,675	160.95	$< 10^{-15}$
Worker-to-Crowd	WI	9	19,700	90.74	$< 10^{-15}$
Worker-to-Expert	WI	9	19,700	59.82	$< 10^{-15}$

is highly significant: χ^2 (*df*=9, *N*= 59,375) = 388.86, $p < 10^{-15}$. The same holds true when comparing worker annotations to an **expert**: χ^2 (*df*=9, *N*= 59,375) = 149.12, $p < 10^{-15}$. Additionally Table 10 shows that the significant differences are robust across three diverse types of data coding/annotation tasks. After seeing a statistically significant omnibus test, we perform post-hoc analyses of all pairwise comparisons using Bonferroni corrections to obtain a more rigorous alpha criterion. Specifically, there are $\binom{10}{2} = 45$ multiple hypothesis tests, so we test statistical significance with respect to $\alpha = \frac{0.05}{45} = 0.001$ for all paired comparisons. In other words, our between-group study design supports 6 sets of 45 comparisons (i.e., $\binom{10}{2} = 45$ pairs) across 3 tasks and across 2 accuracy metrics, for a total of 45x3x2=270 pairwise comparisons. Figure 16 depicts the percentage of correct annotations obtained in each intervention strategy for each type of coding/annotation task, with indicators for those pairs with statistically significant differences and the associated effect sizes.

4.5.3.3 Experiment 2: Design (Strategies 3-4; Task 4)

The experimental manipulations of Experiment 2 are informed by the results from Experiment 1. Referring to the pairwise comparison tests from Experiment 1, we see that targeted screening for task-specific aptitude and training workers to use a standardized, consistent rubric for subjective judgments improves the quality of qualitative data analyses and annotations. Thus we keep targeted screening and task-specific training constant across the conditions of Experiment 2. Also, recall that for the credibility assessment task, the subjective judgment difficulty is even higher than that of the word intrusion task. Based on these observations, we repeat the ***Screen + Train + (Bonus-M)*** as a benchmark condition

for Experiment 2. As test conditions, we compare a range of incentive schemes and iterative filtering:

1. **Screening, training, and financial incentive based on majority matching (*Screen + Train + Bonus - M)***: This condition is same as in Experiment 1 and serves as a benchmark for our second study.

2. **Screening, training, and financial incentive based on Bayesian Truth Serum or BTS (*Screen + Train + Bonus - B*)**: The effectiveness of using financial incentive schemes based on the Bayesian Truth Serum (BTS) technique is reported by Shaw et al. [197]. BTS asks people to prospectively consider other's responses to improve quality. Thus, in this intervention condition, we ask workers for their own individual responses, but we also ask them to predict the responses of their peers. They were told that their probability of getting a bonus would be higher if they submit answers *that are more surprisingly common* (the same wording as [182,197]).

3. **Screening, training, and financial incentive based on Competition (*Screen + Train + Bonus - C*)**: In this condition workers are incentivized based on their performance relative to other workers. Workers were told that their response reason pairs will be evaluated by other workers in a subsequent step to determine whether their response is the most plausible in comparison to their peers' responses. They were rewarded when their response was selected as the most plausible.

4. **Screening, training, and Iteration (*Screen + Train NB – Iteration)***: This strategy presented workers with the original tweets as well as the response-reason pairs

collected in condition 3. Workers were asked to pick the most plausible response-reason pair. Rather than doing credibility assessments directly, workers were acting as judges on the quality of prior assessments, and helping to identify instances where the most commonly selected annotation from the crowd might not be the most accurate/appropriate – that is, they discover whether the crowd has gone astray.

4.5.3.4 Experiment 2: Results

We compare the proportion of correct responses using our two measures of correctness. We find no significant difference when using *Worker-to-Expert* metric. Results are significant for *Worker-to-Crowd*: χ^2 (df=3, N= 7966) = 115.10, $p < 0.008$. To investigate the differences further we again conduct pairwise comparisons with Bonferroni correction. For our four experimental conditions, we conducted a total of $\binom{4}{2} = 6$ comparisons, thus reducing our significance level to $\alpha = \frac{0.05}{6} = 0.008$. For space reasons we omit tabular representation of pairwise comparisons for Experiment 2. We find that across all conditions the winning strategy is the one in which workers are screened for target aptitudes, trained on the task-specific annotation task, and offered incentives for matching the majority consensus from the wisdom of the crowd. Surprisingly, comparing the three incentive conditions (majority-based, BTS-based, and social competition-based incentives) and the iterative filtering strategy, the BTS strategy is the least effective. There is no significant difference between the effectiveness of competition versus iteration treatments. To summarize the relative statistical impact of each strategy:

$$S + T + Bonus\ (M) > [Competition \approx Iteration] > BTS$$

4.6 Analysis and Discussion

In general, we find that screening workers for essential aptitudes, orienting workers to rubrics and heuristics with task-specific training, and financially incentivizing workers to produce well-considered responses are the most successful strategies for improving data analysis and annotation quality. As Table 10 and Figure 16 collectively show, these strategies have a significant impact on the number of "correct" data annotations produced by non-experts, regardless whether "correct" is defined by worker agreement with the most commonly selected annotation from the crowd, or as agreement with the annotation of an accepted expert. Figure 16 (top) also conveys the improvements in intercoder agreement and interpretive convergence among the crowd.

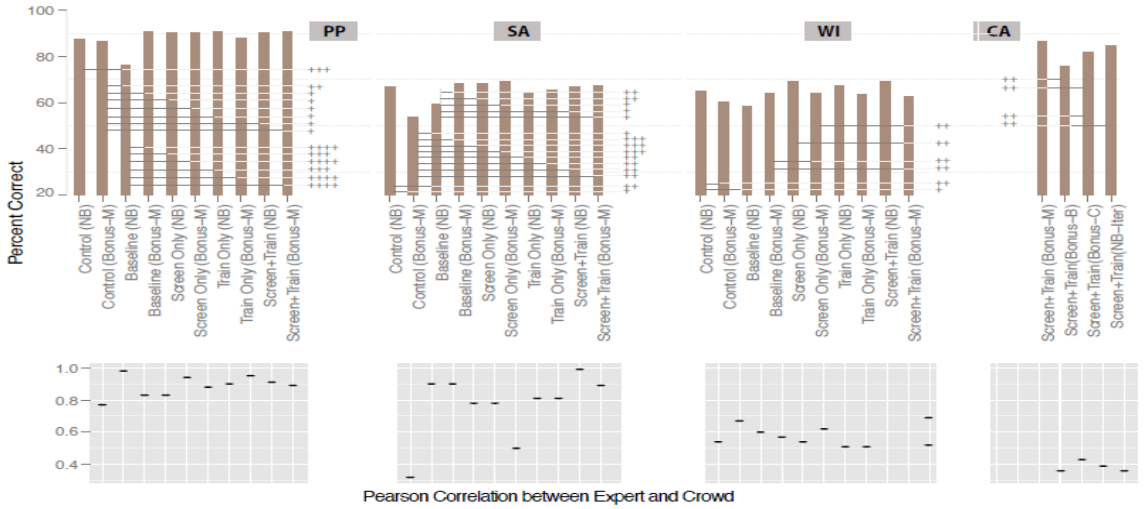


Figure 16: (Top panel) Proportion of correct responses across all tasks with respect to crowd. Pairwise comparisons which are statistically significant are shown with connecting lines (all p-values significant at 0.001 after Bonferroni correction). Effect sizes, as measured by Cramer's V coefficient, are indicated using "+" symbols at four levels: +, ++, +++, and ++++ indicate a very weak effect Cramer's V < 0.15, a weak effect Cramer's V ϵ (0.15, 0.2], a moderate effect (Cramer's V ϵ (0.2, 0.25], and moderately strong Cramer's V ϵ (0.25, 0.3], respectively. (Bottom panel) Pearson correlation between expert and crowd annotations across all tasks.

4.6.1 Effects of Interventions on Annotation Accuracy Increases as Subjective Judgement Difficulty Increases

Based on the overall performance measures, we find that our crowd-generated data analyses and annotations have relatively high data quality (in comparison to prior research, e.g., [197]), even though we use more aggressive criteria for measuring accuracy (that is, *exactly* matching the ratings determined either by wisdom of the crowd or an expert). Further, the effects of interventions are generally robust across a range of representative QDA data annotation tasks of varying judgement difficulty and with varying degrees of subjective interpretation required.

The bottom panel of Figure 16 shows the agreement between the crowd provided annotations and those provided by an accepted expert. In every task, the agreement is well above chance. As data coding tasks become more subjectively difficult for non-experts, it gets harder to achieve interpretive convergence. This is demonstrated by the decreasing correlation trend of the bottom chart in Figure 16. This observation also emphasizes the importance of incorporating strategies for encouraging high-quality annotation accuracy with increased task difficulty. The top of Figure 16 suggests that even modest differences in annotation task difficulty (e.g., a single step increase on the 7 point NASA-TLX scale) produces larger proportional improvements for annotation accuracy over easier coding tasks (e.g., notice that accuracy is already near the ceiling in the people in pictures task).

4.6.2 Person-oriented Strategies Trump Process-oriented Strategies for Encouraging High-Quality Data Analysis and Annotations

A very interesting finding from this study is that – in contrast to commonly employed "process oriented" tactics – when we target intervention strategies towards verifying or changing specific attributes of the individual worker, we see better and more consistent improvements in data analysis and annotation quality. For example, by verifying that the person has the requisite aptitude (knowledge, skill, or ability) necessary to perform a particular data annotation task, we observe significant increases in the number of correct annotations from non-experts when compared to a) control, b) simple experience/reputation baseline, c) BTS, d) competition, or e) iterative filtering strategies.

The insightful work from Shaw and colleagues [197] noted that process-oriented strategies like BTS, which prompt workers to prospectively reason about the responses of their peers, tended to be more effective at promoting better quality annotations. We find that person-oriented strategies such as a prior screening for requisite aptitudes, together with training workers on task-specific data analysis expectations, *can significantly improve effectiveness above and beyond the effects of BTS.* These strategies emulate the methods of sharing a common "codebook" that qualitative scholars have employed for years for adjudicating data coding among collaborative data coders [194]. Non-expert workers who reason about the likely responses of their peers are more likely to achieve greater interpretive convergence – and do so more quickly, with less variation (c.f., [71]) – when they think about the data coding activity in the same ways.

4.6.3 Why Do More to Get Less?

In terms of effort on behalf of both the research-requester and the worker-coder, intervention strategies such as screening and training workers have a one-time up-front cost

associated with their implementation, but their cost quickly becomes amortized for even moderate sized datasets. In contrast, strategies such as BTS, Competition, and Iteration require the same, sustained level of effort for every data item that needs to be coded or annotated. As such, the per-item cost for BTS, Competition, and Iteration are much heavier as the size of the dataset grows. Given that these strategies actually do not perform as well as screening and training, why do more to get less quality?

4.6.4 Amazon is not a neutral observer; AMT is getting better

While our results show significant improvements in the quality of data annotation tasks over control and baseline conditions, we note that the quality of data obtained from AMT workers in those conditions is much higher than we initially expected, given our experience with the platform over the years. We also highlight the finding that in every task across all interventions, the accuracy of crowd-produced annotation is not only well above random chance (20% for all our tasks), but also well above the control condition and even the BTS treatment condition for similar subjective-oriented tasks reported in [197]. For example the "rank content" and "rank users" tasks from [197] are precisely the kind of subjective-oriented tasks that we are targeting with our interventions, but accuracy reported in [197] peaks at ~40% for even the best incentive category (BTS). We also point out that the chance for randomly guessing the correct response for these two subjective judgment tasks was 20% (the same as with our study). Contrast this with our results – even in the more difficult subjective judgment tasks, we find control condition accuracies in the 55-80% range (and our person-oriented treatment conditions are even higher, in the 65-95% range). These performance scores far exceed those reported just a few years ago in [197] for their two subjective judgment tasks. So it seems that compared to just a few years ago,

the quality of QDA on AMT is improving. Interestingly, the kinds of measures Amazon has enacted are also quite person-oriented: e.g., requiring workers to verify their identity by providing their tax information[29], requiring workers to prove their humanity using CAPTCHAs at random intervals before accepting some HITs, or perhaps requesting proof of U.S. residence by providing a utility bill.

These results are not intended necessarily to be prescriptive. Even in a study this size, we still focus on just a subset of potential intervention strategies, subjective judgment tasks, and various financial based incentives. Future work should directly compare the efficiency and effectiveness of a priori person-centric techniques to peer-centric methods (c.f., [96]) and more complex post hoc statistical consensus finding techniques (c.f., [199]). Nonetheless, the person-centric results reported in here help illustrate the value of applying established qualitative data analysis methods to crowdsourced QDA coding by non-experts.

4.7 Chapter Summary

In this chapter, I systematically compared the relative impacts of numerous a priori person- and process-centric strategies for improving the quality of qualitative data analysis and annotations from non-experts, and I (with my colleague) checked their robustness across a variety of different content analysis tasks. I offer several reasons for focusing on *a priori* techniques, as opposed to complex statistical data cleaning techniques performed post-collection:

[29] https://www.mturk.com/mturk/help?helpPage=worker

1. First, the value of a priori strategies are not as well explored, lending novelty to the contributions reported here.
2. Second, for time sensitive judgments (e.g., credibility decisions for rapidly unfolding events), simple a priori methods trump complex post hoc methods.
3. Third, a priori person-oriented strategies emulate the procedures of sharing a common QDA codebook. These results demonstrate the value of applying a well-established social science method for qualitative data analysis to crowdsourced annotations by non-experts.
4. Fourth, targeted screening and task-specific training techniques have a onetime up-front cost which soon amortizes with increases in the size of datasets, so these techniques scale up to large datasets exceptionally well.
5. Fifth, person-oriented strategies are arguably more generalizable; they can be adapted to adjudicate both *objective* and *subjective* judgments. Post hoc data cleaning is suited more for objective tasks and breaks down as data becomes noisy; thus, such procedures are of limited use for subjective oriented judgment tasks.

The third, fourth, and fifth reasons described above are especially relevant to this dissertation, as the research in Chapters 3 and 5 both rely heavily on Amazon Mechanical Turk to provide qualitative data analysis (QDA) on large scales without sacrificing analysis quality, and for human-centered verification, validation, and evaluations (VV&E) of computational models of sentiment analysis (Chapter 3) as well as bias detection and quantification in sentence-level text of journalistic news stories (Chapter 5). In the next

chapter, I demonstrate how the methods and tools described in Chapters 2-4 can be applied towards social science theory building via computationally detecting bias in journalism.

CHAPTER 5. COMPUTING BIAS IN JOURNALISTIC NEWS

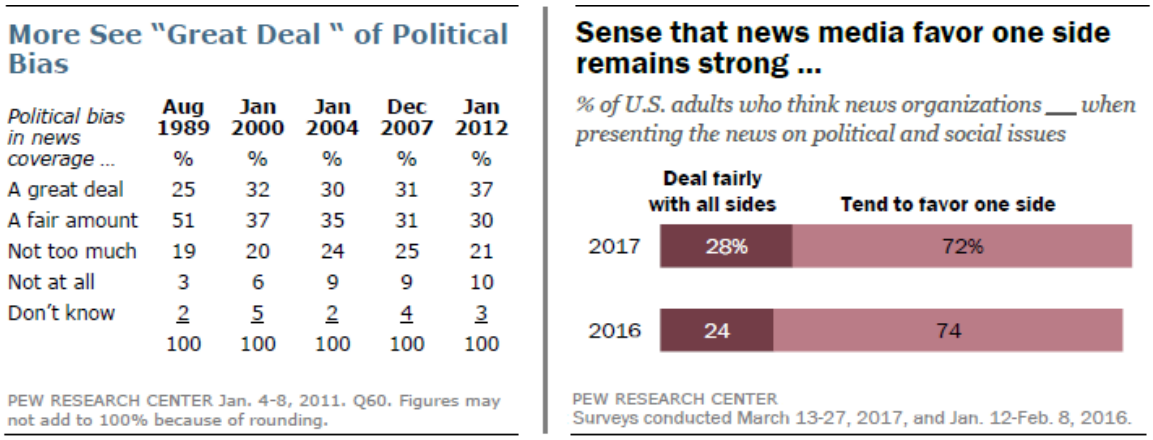
5.1 Chapter Overview

In an effort to maintain standards of journalistic integrity to provide fair, impartial, and balanced presentations of newsworthy stories, most news organizations strictly separate journalistic news and editorial staffs. Perceptions of bias are, unfortunately, nevertheless ubiquitous. For example, perceived credibility of both print and broadcast journalism has been steadily declining for more than a decade [179], and 74% of U.S. adults believe news organizations tend to favor one side when presenting political or social issues [181], reflecting the *hostile media effect* [61]. In this chapter, I first briefly review the evidence for perceptions of hostile media bias (Section 5.2.1), and describe the descriptive framework which forms a basis for discussing manifestations of bias in the news (Section 5.2.2). Next, I synthesize literature from psychology and communications studies regarding the nature of bias, outlining the theoretical underpinnings which help explain the origins of biased perceptions (Section 5.3). I then develop a theory-informed computational model called the Biased Sentence Investigator (BSI, see Section 5.4) that aims to detect and quantify the degree of bias in sentences of news stories. BSI implements a total of 32 sentence-level and lexical-level measures, which I hierarchically organize into 13 higher-order features. These include sentence-level measures such as *sentiment* and *certainty* as well as lexical-level measures such as *presupposition* language markers (which reflect epistemological bias and presupposed truths), and *value-*, *partisan-*, and *figurative-*language markers (which reflect a blend of biases arising from the framing effects associated with certain rhetorical devices) to name a few. After distinguishing BSI from

existing computational approaches for measuring bias (Section 5.5), I next perform a preliminary feasibility assessment of the model's performance against a realistic and representative (albeit synthetic) dataset (Section 5.6). I then build upon the insights gained from this preliminary assessment to conduct an expanded study with a larger dataset and more rigorous feature selection process, and compare 26 different statistical and machine learning models as computational implementations of the BSI conceptual model to predict the perceived bias of sentences in an annotated dataset of news articles (Section 5.7). Implementations range from multiple variations on linear regression models to more complex nonlinear, non-parametric regressions, decision trees, random forests, neural networks, and support vector machines. Extensive feature and model evaluations show that performance of the BSI model and selected features compare quite well to human performance for matching the average perceived bias rating for sentences in news stories (mean Pearson Correlation Coefficient r=0.565 for BSI using Regularized Random Forest machine learning, compared to r=0.661 for human judges). I close the chapter in Section 5.8 by demonstrating BSI model capabilities for investigating statement bias and coverage bias at the sentence and article units of analysis for news stories.

5.2 Perceptions of Bias in Journalism

Fair and impartial reporting is a prerequisite for objective journalism; the public holds faith in the idea that the journalists we look to for insights about the world around us are presenting neutral, unprejudiced facts. Indeed, most news organizations strictly separate *journalistic news* and *editorial* staffs. Bias is, unfortunately, nevertheless ubiquitous in journalism. One area in which bias is particularly prevalent is with political journalism. According to the Pew Research Center, 67% of Americans polled in 2012



More See "Great Deal " of Political Bias

Political bias in news coverage ...	Aug 1989	Jan 2000	Jan 2004	Dec 2007	Jan 2012
	%	%	%	%	%
A great deal	25	32	30	31	37
A fair amount	51	37	35	31	30
Not too much	19	20	24	25	21
Not at all	3	6	9	9	10
Don't know	2	5	2	4	3
	100	100	100	100	100

PEW RESEARCH CENTER Jan. 4-8, 2011. Q60. Figures may not add to 100% because of rounding.

Figure 17: Perceptions of bias in journalism continue to be prevalent in America.

thought that there was either a "fair amount" or a "great deal" of political bias in news coverage [178] (Figure 17, left panel). In 2016 and 2017, 72-74% of U.S. adults thought news organizations tended to favor one side when presenting political or social issues [181] (Figure 17, right panel).

Furthermore, perceived credibility of both print and television journalism has been steadily declining [179] (Figure 18), reflecting an increase in what mass communications researchers have termed the *hostile media effect* [61] or *hostile media bias* [176] (discussed more in the next section). It is therefore at once both intellectually fundamental to understand the nature of bias in the practice of journalism, and pragmatically useful to be able to conduct rapid initial review of news stories for the presence of bias. The practical advantages of systematically exposing indicators of bias and making its nature transparent is that writers, editors, and publishers can self-assess journalistic news from a common contextual viewpoint to diagnose biased stories prior to printing.

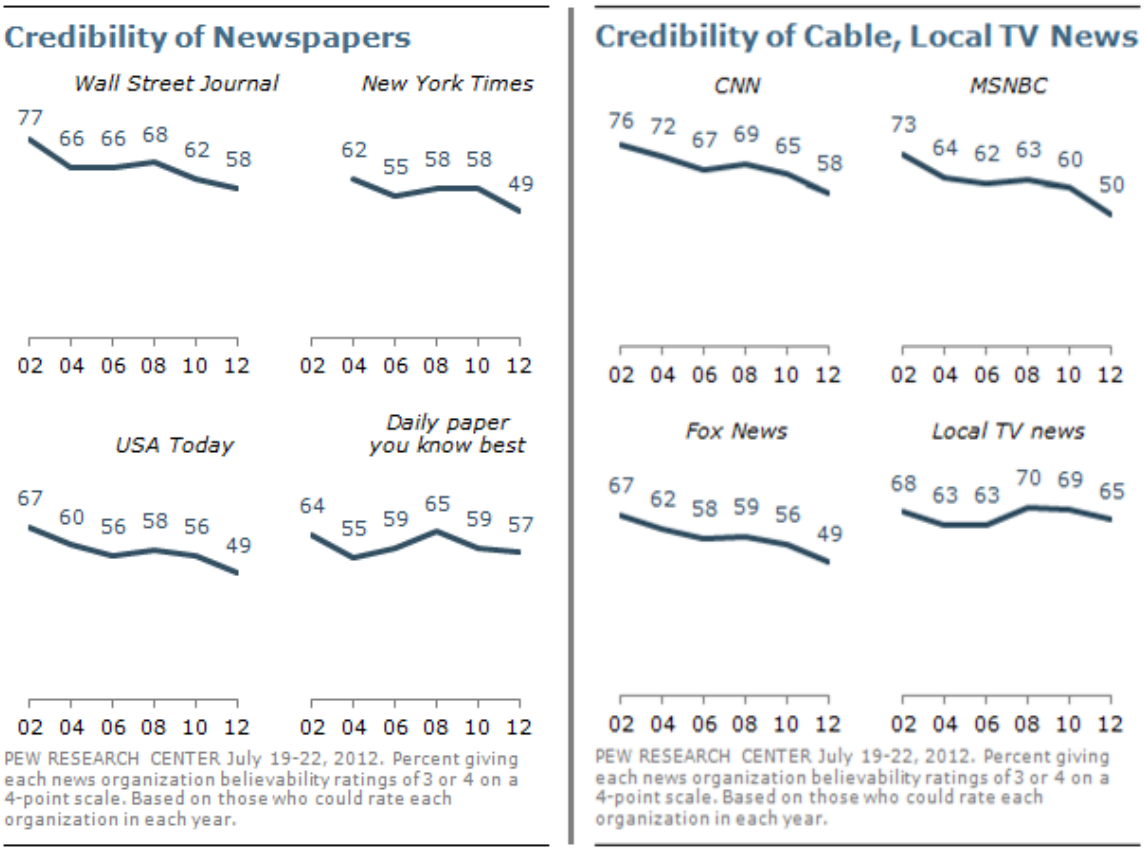


Figure 18: Perceptions of waning credibility in print and television news [179] reflect sensitivities to "hostile media bias" [176].

Similarly, it may not always be apparent to readers that a particular news story is intended to reflect an editorial stance or the writer's opinion. But if sentence structures and language markers of bias were identified and exposed, then readers, curated content providers, and media-monitor groups would have the opportunity to become more aware. Likewise, media analysts, computational journalism researchers, and media studies researchers could then compare the degree of objectivity for news stories over time and across news categories, topics, authors, news organizations/media sources, newspaper corpora, or geographic boundaries.

5.2.1 Hostile Media Bias

The *hostile media phenomenon* [213] – also called *hostile media bias* [176] and *the hostile media effect* [61] – is a well established theory situated within research at the intersection of psychology and mass communications studies. In essence, the theory posits that two people with opposing views will both perceive that the exact same news media coverage is unfairly biased against their point of view in favor of their opponents' position. In other words, regardless of either their own stance or the intention of the reporter, partisans will find news content to be "hostile" to their own point of view (even when such content is judged by nonpartisan individuals to be ostensibly neutral due to balanced or even-handed treatment of the opposing views).

One the one hand, the hostile media phenomenon suggests that perceptions of bias are centered within the individual, rather than the content of news stories. Indeed, a review of the classic research on ego-involvement shows how inextricably connected (and entangled) existing attitude extremity, personal importance of an issue, and other self-relevant attributes are for understanding perceptions of media bias [176]. However, extensive research into mediators of the hostile media effect demonstrate that *selective categorization* – whereby individuals attend more to story content that is unfavorable to their perspective, rather than focusing on favorably oriented content – is among the best constructs for explaining the hostile media effect [195].

With this in mind, it is absolutely conceivable that an in-depth analysis of news story content (alone) might reveal text-based structural and lexical markers for perceptions of bias, *regardless of the position or stance of either the reader or the author*. In computational linguistics and related disciplines there is already a rich literature on stance recognition and argument subjectivity that focuses on identifying which side an article

takes on a two-sided debate (c.f., [137]), casting the research task as a two-way classification of the text as being either for/positive or against/negative (e.g., [5,35,202]) or as one of two opposing views (e.g., [169,230]). Given the existing richness of the literature on stance recognition, the research presented in this chapter focuses instead on detecting and quantifying the degree of media bias, irrespective of stance, by explicitly linking linguistic and text-based structural indicators to specific types of bias informed by well-established social science theory. A comprehensive investigation of the hostile media effect would require an assessment of both the individual – in particular, their preexisting attitudes and sentiments towards topics in the story – as well as an assessment of the presence (and intensity) of biased content within the news story itself. Chapters 2-4 describe unobtrusive computational techniques which lay the foundations for accomplishing the former; this chapter addresses the latter. To this end, in the next few section I begin by describing insights regarding the ways in which bias is manifested in text-based journalism (Section 5.2.2), exploring the forms and nature of bias in certain manifestations (Section 5.3), and then developing computational techniques for quantifying and quantifying aspects of text-based media bias (Section 5.4).

5.2.2 Manifestations of Bias in Journalism

Prior to developing a model for detecting and quantifying bias in news media, it is important to first characterize the ways that bias may be manifested in news media. In a meta-analysis of 59 studies containing quantitative data measuring the partisanship of news media during presidential elections, D'Alessio and Allen [40] organize the literature into three broad categories media bias: *gatekeeping* bias, *coverage* bias, and *statement* bias.

5.2.2.1 Gatekeeping Bias

Gatekeeping bias – also referred to as *agenda bias* in politically oriented media studies research [56], as *selectivity* in political science research [90], or as *selection bias* in communications studies research [77] – is the tendency of writers or editors to select from a body of potential stories those that will be presented to the public (and, by extension, which stories are discarded and for which the mass media audience will hear nothing) [40]. Gatekeeping bias is nearly impossible to detect at the article level because the "population" of all possible stories, or all aspects of stories, is not available. Thus, determining the degree to which particular stories or aspects are *not* selected is unknowable at sentence or article levels of analysis. Gatekeeping bias can be measured at the news outlet level, e.g., by considering a large enough corpus of stories from a particular newspaper; but investigations of gatekeeping bias are most effective when the unit of analysis considers multiple corpora of stories from many media outlets or news organizations. In this way, it becomes straightforward to quantify the degree to which any particular outlet excludes newsworthy stories, or aspects of stories, reported elsewhere [40].

5.2.2.2 Coverage Bias

Coverage bias – also called *visibility bias* [56] – addresses the relative amount of attention given to a particular stance, position, or aspect of an issue in news media [40]. Coverage bias has been operationalized for print media according to the column length of articles in newspapers or newsmagazines (measured as inches or as number of words), the number and size of photographs, the number and size of headlines, or – in the case of audio or visual presentations – the amount of time devoted to certain sides of the issue [40].

Coverage bias can also be influenced by *structural bias* – whereby media routines (e.g., timing and news cycles) and the newsworthiness of either a person (e.g., an incumbent) or an issue (e.g., a controversial or trending/hot topic), rather than ideological positions – lead to more media coverage [39]. Coverage bias lends itself very well to article level units of analysis, and can be extended to larger scales (e.g., assessing coverage bias tendencies at the news organization level of analysis, or even geographic boundaries at local, regional, national, or international scales).

5.2.2.3 Statement Bias

Statement bias – also called *tonality bias* [56] or *presentation bias* [77] – is concerned with linguistic, lexical, and grammatical representations of media content [40]. Statement bias can take many forms, including the degree to which a sentence presupposes some underlying truth, the degree to which a writer's own opinion or values are injected into a statement, the degree of doubt (or certainty) expressed, or the choice of specific non-neutral, partisan-oriented words or phrases to convey subtle preferences for one side over another. For example, consider how the phrase "pro-life" connotes stronger favor when compared to the phrase "anti-choice" in reference to the same ideological side of the abortion issue. Statement bias is most readily assessed using the sentence as the primary unit of analysis, and aggregate/descriptive statistics allow the analytics to extend to article level and corpus level units of analysis.

The research presented in this chapter is principally interested in detecting and quantifying bias at the sentence level, as statement bias appears to be generalizable and generally useful to a broad range of analyses at different scales. To facilitate a well-

informed selection of sentence-level features relevant to quantifying statement bias, I next examine some of the psychological constructs at the root of perceived bias.

5.3 The Nature of Bias: Types and Forms of Biases

Perceptions of media bias occur when news stories violate journalistic standards for objectivity (neutrality and impartiality), balance and even-handedness, or representations of social realities (fatuality/truth and relevance) [80]. In general, the psychological mechanisms underlying such perceptions can be grouped into two broad categories: biases that are a result of the *framing effects* of a news story – whereby people react to information differently to draw disparate conclusions from the same information depending on how that information is presented [26,212], and biases that result from core *epistemological* roots – whereby implicit prejudices or presuppositions form the basis of perspective [76,137,186]. In this section, I consider several specific, well-researched social science constructs related to *framing effects* and *epistemological* underpinnings of bias.

5.3.1 Framing Effects and Biases

The concepts of "frames" and "framing" have numerous connotations within the social sciences, but the conceptualization most relevant to the research presented in this chapter stems from Cappella and Jamieson's definition regarding "the way the story is written or produced," including the cognitive-anchor-setting headlines which orient and signal conclusions, specific word and phrase choices within and accompanying the text, rhetorical devices employed, the narrative form, and so on [26,53]. A number of research established constructs are relevant to deriving valid features for detecting biases resulting from framing effects, including: negativity bias, belief bias, attribution bias, and rhetorical devices

relying on sentiment, subjectivity, and figurative editorializing, as well as general partisan (non-neutral) discourse. I next describe these concepts, and then (in Sections 5.4.2 and 5.4), I operationalize them in order to detect and quantify bias in text.

5.3.1.1 Negativity Bias

In psychology, the *negativity bias* refers to empirical evidence demonstrating that negatively oriented content has greater potency for affecting a person's cognition and behavior than equally intense positive (or neutral) content [13,192]. In other words, people tend to be more sensitive to *criticisms* of a position: selectively attending to and giving greater consideration and cognitive processing to negative presentations which disparage the point of view while minimizing attention to acclaims and affirmations which support the stance. Indeed, the English language itself seems constructed to support more elaborate and more complex cognitive processing for negativity than positivity: negative English vocabulary is much more richly descriptive [172], and there are more terms (and more gradients of connotation) related to negative language compared to positive language [27,98,174,175,205]. Framing certain elements of a news story with a negative orientation can result in selective categorization [195], where news audience members devote greater attention to unfavorable presentations of a position (described previously as an explanation for the hostile media phenomenon).

5.3.1.2 Belief Bias

As evidenced by selective categorization, negativity bias is further exacerbated by the degree to which content either supports or contradicts one's own experience, prior knowledge, existing attitudes, values, or beliefs. *Belief bias* is the tendency to give greater

veracity to arguments when those arguments are presented such that the conclusions are congruent with a person's own beliefs and expectations rather than on the merits or validity of the argument itself [59,151]. Similar to the literature on cognitive dissonance (whereby people tend to reject information that is inconsistent with existing beliefs) [63], belief bias explains why people sometimes accept illogical or invalid arguments, as long as the conclusion supported by the arguments align with existing values and beliefs.

5.3.1.3 Attribution Bias

Attribution bias refers to systematic errors people make when reasoning about the cause of their own or others' behaviors and actions [89,111]. Attribution bias can take the form of the *fundamental attribution error* (FAE), whereby individuals are more likely to overemphasize the role of personality or dispositional factors (internal traits) and underemphasize situational factors when reasoning about others' behaviors [111,112]. For example, a student would be more likely to attribute a teacher's harsh words about class performance on an exam to the teacher's abrasive personality (internal factor), rather than on commentary about the scores of the test (external factor). Another form of attribution bias is the *actor-observer bias* (AOB), which extends FAE to include reasoning about one's own activities by over-valuing situational factors to explain behaviors [113]. When combined with FAE, AOB explains that "actors tend to attribute the causes of their behavior to stimuli inherent in the situation, while observers tend to attribute behavior to stable dispositions of the actor" [113]. For example, a colleague who stays late to finish a project would attribute the behavior to external situational factors (e.g., "I have a client meeting later this week"), whereas coworkers would be more likely to attribute it to dispositional traits (e.g., "She is ambitious and hard-working"). The *ultimate attribution*

error (UAE) further extends FAE and AOB to the group-level, and encompasses tendencies related to belief bias such that in-group and out-group behaviors are evaluated differently depending on whether the behavior reflects positively on or is congruent with existing group beliefs/norms [177]. Thus, UAE is the tendency to attribute negative out-group and positive in-group behaviors to internal/dispositional factors, and attribute positive out-group and negative in-group behaviors to external/situational factors [177].

5.3.1.4 Rhetorical Influences

Rhetoric refers to the art of using writing or speech to persuade, influence, or please [229]. Common rhetorical devices used for persuasive (often biased) writing or speech generally fall into four categories:

1. *Logos* relies on logical arguments and supportive evidence (facts, examples) in conjunction with explicitly stated conclusions to influence the audience.
2. *Pathos* appeals to the emotions of the audience using expressive, value-laden (passionate) discourse, or with figurative language such as metaphors.
3. *Ethos* involves conveying moral competency (good will, virtue, no intent to deceive and without agenda or ulterior motives), expertise, and credibility.
4. *Kairos* refers to the timing, timeliness, or opportune moments for appeals or calls to action.

For biases resulting from framing effects, rhetorical devices relying on *pathos* are particularly prevalent. Using strong value-laden, subjective, or opinion oriented language signal clear attempts to appeal to a reader's emotions. Likewise, explicit statements associated with *logos* oriented arguments might attempt to convey information by using

examples, analogies, metaphors or other figurative language. However, logos is sometimes reflected in more subtle construction of coherent arguments, implying cause/effect, additive, adversarial, or comparative relationships between sentence sub-parts. Similarly, *ethos* oriented arguments might explicitly attend to matters of expertise (of the journalist, of a journalist's source, or of the target of the story); or, sometimes expressions of doubt or questions of credibility are more subtly invoked. When rhetorical devices rely on more oblique techniques, they more closely align with epistemological biases. To address these implicit biases, and to garner further theory-informed insights in addition to considering framing effects, I next turn to epistemological-based considerations for assessing bias.

5.3.2 Epistemological Biases

Epistemological biases occur when implicit prejudices or presuppositions form the basis of perspective [76,137,186]. Epistemological biases are often quite subtle, and are therefore more difficult to detect as they do not explicitly signal a writer's stance or position, but rather reflect either an assumed (presupposed, unchallenged) underlying truth to a proposition [186], or attempts to "shepherd" a reader towards an implied conclusion [124] by implying relationships (such as cause and effect) or by drawing comparisons between subject and predicate clauses in sentences. In discourse analysis, epistemological biases may be invoked using linguistic markers that are not necessarily connected to any particular framing of an argument, but are instead indicative of the writer's own presuppositions. For example, consider two example sentence fragments: (1) "The data in the study revealed that…" versus (2) "The data in the study indicated that…". In the first statement, the verb "reveal" presupposes that there was some underlying truth that the study uncovered, whereas the second statements makes no such presupposition [120].

Similarly, coherence markers such as "…*as a consequence*…", "…*it seems that*…", and "…*and so it follows*…" all reflect sentential complements intended to lead the reader to particular conclusions by implying, for example, cause-effect relationships [124]. Such strategies employ rhetorical logos, but in a more discreet manner than what might be expected with explicit framing techniques.

Consider, as well, two ethos-oriented techniques that subtly either facilitate perceptions of credibility (in the first case), or call credibility in to question: (1) "The economy expert was quoted as saying 'We expect to see significant growth in the number of houses purchased in the coming months'" versus (2) "The economy expert claims the housing market will experience growth next quarter". In the first statement, the writer uses quotations the lend credibility to the story being reported, whereas the word "claims" in the second statement reduces the writer's commitment to the truth of the proposition, calling the statement's credibility into question and implying skepticism [186].

Taken as a whole, the literature and evidence for framing effects and epistemological biases (negativity, belief, attribution, rhetorical influences, presuppositions, and coherence markers) form a compelling foundation of social science theory as motivation for selecting and operationalizing linguistic features of text for detecting and quantifying bias. With these theoretical foundations in mind, I next describe the sentence level and lexical level features I derive in order to develop an initial model for computing bias in news stories. Figure 19, located at the end of Section 5.4, presents a graphical summary of the links between the theoretical foundations described in this section, the computational model, and the individual features used to compute statement

bias described in the next section. (These features are refined based on insights gained from a small preliminary study [104], which is described in more detail in Section 5.6).

5.4 Modeling Bias: Biased Sentence Investigator (BSI)

Informed by well-established theory from psychology, computer-mediated communications [CMC] research, and mass media communications studies described in the previous section, I next derive a total of 32 lexical and sentence level measures of potential bias in news stories. In this section, I described the hierarchical organization of these measures into seven lexical features and six sentence level features.

5.4.1 Lexical Level Indicators of Statement Bias

I implement several lexical level features that I hypothesize have an effect on human perceptions of bias in text. To detect and quantity lexical indicators of statement bias, I count the number of matching words and phrases from the following seven categories:

1. **Presupposition** language markers reflect epistemological bias and presupposed truths; "leading" or suggesting a conclusion. I consider five forms of lexical indicators of presupposition:
    - **Factive verbs** presuppose the truth of an embedded sentence that serves as its complement, as in *realize* in “I didn't realize that Sarah had left”, which presupposes that it is true that Sarah had left [46]. I use a list of 27 factive verbs derived from [120], and draw on inspiration and insights from [186] for using them to detect epistemological bias.

- **Implicative verbs**, like factives, also imply the truth or untruth of their complement, but do so while also implying some additional condition [115], and are therefore another form of epistemological bias [186]. The word implicative is related to *implication*; its root word is *imply*. Consider two examples: (1) "Denise managed to solve the problem", and (2) "Andrew remembered to lock the door". In the first sentence, the verb *manage* implies that the problem was considered to be in some way difficult to accomplish (at least for some people), and that Denise had the skill or ingenuity to succeed. "Denise solved the problem" does not convey the same presupposition. In the second sentence, *remembered* presupposes that Andrew was in some way obligated to lock the door, and that he had the basic willingness and intension to fulfill the commitment. Again, "Andrew locked the door" does not express the same assumptions. I derive a list of 32 implicative verbs from canonical linguistics research on the subject [115].
- **Asserting words**, unlike factives or implicatives, do not presuppose the truth of a proposition, but instead presuppose a degree of conviction for the proposition. Assertive predicates such as *declare*, *certify*, and *testify*, presuppose greater confidence than reportative predicate counterparts such as *state*, *show*, or *tell* [93]. For example, "She *checked* the tire pressure before her road trip" does not presuppose the same degree of conviction compared to "She *verified* the tire pressure before her road trip". This is another form of subtle epistemological bias; it also invokes the ethos

rhetorical device. I derive a list of 66 assertive words from established linguistics research [93].

- **Causation words** such as *create*, *founded*, and *generate* can signal an act or agency which produces an effect. Such word choices can imply subtle presuppositions regarding the truth of an underlying cause-effect relationship, an epistemological bias with logos-based rhetoric. The Linguistic Inquiry and Word Count (LIWC) [173] is text analysis software designed for studying the various emotional, cognitive, structural, and process components present in text samples [174]. LIWC uses a proprietary dictionary of almost 4,500 words organized into one (or more) of 76 categories. I use the LIWC list of validated causation category words [174].
- **Coherence markers** are words (*because*, *therefore*, *so*) or lexical phrases (*as a result*, *for that reason*) that may be used to bias a reader towards a particular conclusion. As stated earlier, such sentential supplements evoke epistemological bias and rhetorical logos by implying, for example, cause-effect relationships in a more discreet manner than what might be expected with explicit framing techniques. I use [124]'s list of coherence markers.

2. **Value-laden** language markers reflect when a writer injects subjective values into the presentation of issues/facts, resulting in a kind framing bias which blends both logos and pathos oriented rhetorical devices. Subjective opinions and positively/negatively loaded emotional language are two examples of value-laden language lexical indicators:

- **Opinion words** signal the expression of positive or negative attitudes or opinions, which may be influenced by emotional (pathos) biases. I use the validated opinion lexicon from my VADER sentiment analysis engine [98,100] to count opinion-matching lexical features detected in news stories.
- **Non-neutral subjective intensifiers** are contextual cues (often adjectives or adverbs such as *extremely*, or *slightly*) that modify the intensity or degree of a verb, adjective, or other adverb in order to add (subjective) force to the meaning of a phrase or proposition. I garner insights from [186] for detecting framing bias in text using [188] and [227]'s lists of both strong and weak non-neutral subjectives, and then I combine and extend these lists by incorporating the lexicon of degree modifiers from VADER [98,100].

3. **Figurative** language can reflect a blend of framing bias and epistemological bias intended to convey a non-neutral perspective; the inherent nature of figures of speech manipulates the opinion of the reader [146]. Unlike literal language, figurative language leverages linguistic devices such as metaphor, analogy, metonymy, idiom, hyperbole, and so on, in order to guide a reader to a conclusion by (for example) emphasizing or deemphasizing views, similarities, differences, or equivalence [189]. Figurative expressions are often employed in discourse involving humor, sarcasm, or irony. I compile a list of figurative expressions containing idioms [218], general metaphors [219], political metaphors [220], as well as figurative expressions from Wikipedia's lists of "puffery" and "peacock" Neutral Point of View (NPOV) watch words [221]. For example, consider the

figurative phrase *defining figure* in the following: “Bob Dylan is the defining figure of the 1960s counterculture and a brilliant songwriter”. An NPOV sentence would instead present facts, such as: “Bob Dylan was included in Time's Top 100: The Most Important People of the Century, and by the mid-1970s his songs had been covered by hundreds of other artists” [221].

4. **Partisan** language reflects framing via ideological bias and/or belief bias, indicating a non-neutral point of view. Partisan language includes, of course, politically charged words and phrases. Compare the ideologically right leaning, Republican phrases *death tax*, *tax-relief*, *personal account*, and *war on terror* to their ideologically left leaning, Democrat counterpart phrases *estate tax*, *tax break*, *private account*, and *war in Iraq*. I incorporate a list of 120 partisan words and phrases compiled by [69], which comprised the top 30 bigrams and top 30 trigrams used by congressional Democrats and Republicans in speeches or sponsored legislation documented in the 2005 Congressional Record. Partisan language can also include contentious labels or one-sided terms that reflect a writer’s subtle belief bias. Consider the writer’s implicit perspective (epistemological bias) when referring to a group of armed individuals as either *freedom fighters*, *rebels*, *insurgents*, *extremists*, or *terrorists*. I therefore also integrate the list of biased one-sided terms derived by [186], as well as Wikipedia’s list of contentious label NPOV watch words [222].
5. **Attribution** language markers based on using third person pronouns to attribute either dispositional or situational traits can reflect potential biases arising from the psychological constructs of fundamental attribution error (FAE) [111,112], actor-

observer bias (AOB) [113], or ultimate attribution error (UAE) [177]. To attempt to capture third party attribution, dispositional, and situational factors, I include LIWC's validated lists of third person pronouns (e.g., *he*, *him*, *she*, *hers*, *they*), achievement words (e.g., *accomplished*, *master*, *prized*), and work words (e.g., *ambitious*, *resourceful*, *hard-work*) [174].

6. **Doubt** related language markers consists of expressions of reservation, uncertainty, or distrust, and may imply inaccuracies which call a statement's credibility into question – an ethos oriented rhetorical device. I incorporate lexical indicators of doubt comprised of LIWC's list of tentative words (e.g., *bets*, *dubious*, *hazy*, *guess*) [174], a list of "hedge" words from literature in a sub-field of linguistic discourse analysis called metadiscourse [105], as well as Wikipedia's lists of doubt and "weasel" NPOV watch words [223,224].
7. **Self-reference** language markers may indicate personal thoughts rather than an objective/unbiased (neutral) point of view, or else potential biases from AOB [113] or UAE [177]. I include LIWC's validated list of self-referencing pronouns (e.g., *me*, *my*, *I*, *we*, *us*, *our*) [174].

5.4.2 Sentence Level Indicators of Statement Bias

At the sentence level analysis of text, I observe characteristics of the statement as a whole, considering syntactical, grammatical, and structural properties captured using the following six features:

1. Sentence length (**word count**). Whereas shorter sentences are generally easier for people to understand and process, longer sentences afford the opportunity to

employ a greater number of (potentially biased) framing effects or epistemological linguistic features. Aside from the potential to influence perceptions of bias, it is also prudent to include the overall number of words in a sentence as a regulator or reference when assessing the impact of more granular features.

2. The Flesch-Kincaid Grade Level (**FKGL**) formula [119] quantifies the readability of a sentence and associates it with a typical requisite grade level of reading comprehension. The higher the grade level, the more difficult the text. I hypothesize that higher FKGL scores have greater potential to influence perceptions of author credibility, an indicator of ethos related rhetoric.
3. **Quote length** (number of words in quotation). Recent work focused on NLP techniques for assessing how quoting practices influence a reader's judgments of factuality and perceptions of credibility (i.e., ethos) [203] as well as research on how quoting patterns characterize the degree to which media outlets exhibit systematic political coverage bias [158] motivates my consideration of quotes to detect and quantify bias in news stories. I use quote length, rather than a binary Boolean for quote use, as a sentence level feature with the hypothesis that, as with sentence length, longer quotes will afford the opportunity to employ more epistemological and framing effects, while simultaneously conveying (at least the appearance of) objectivity by directly quoting sources.
4. **Sentiment** scores can reflect when a sentence contains subjective expressions of attitude or belief (i.e., pathos rhetoric). VADER's sentiment analysis processing engine implements numerous empirically derived sentiment processing rules related to textual syntax, grammar, punctuation, capitalization, negation, and other

word-order sensitive elements of text [98,100]. However, rather than being concerned with directionality or polarity, I hypothesize that sentiment expression in either direction (positive or negative) indicates bias. I therefore use the absolute value of VADER's compound sentiment score to compute the intensity of the sentiment of each sentence (thus, values range continuously from 0 [neutral, or balanced] to +1.0).

5. **Certainty** is the sentence level affirmative counterpart to lexical level doubt-related language markers which indicate logos and ethos oriented rhetoric. I use Pattern.en's [34] *modality* module to compute degree of certainty for sentences. The module returns the degree of certainty as a value between -1.0 and +1.0, where values >+0.5 represent facts. For example, "*I wish it would stop raining*" scores -0.35, whereas "*It will stop raining*" scores +0.75. Accuracy is about 68% for Wikipedia texts [34].
6. **Negative-Perspective** is a sentence level operational quantification of *negativity bias*, which posits that negatively oriented content has greater potency for affecting a person's cognition and behavior than equally intense positive (or neutral) content [13,192]. The Negative-Perspective Index incorporates consideration for use of *negation*, accounts for the *proportion* of a sentence that is negative (versus positive or neutral), and captures the *intensity* of the negativity within the sentence.

Figure 19 illustrates connections between individual features used in the computational model as well as theoretical underpinnings from psychology, computer-mediated communications [CMC], and mass media communications studies. I refer to this computational model as the *Biased Sentence Investigator (BSI)*.

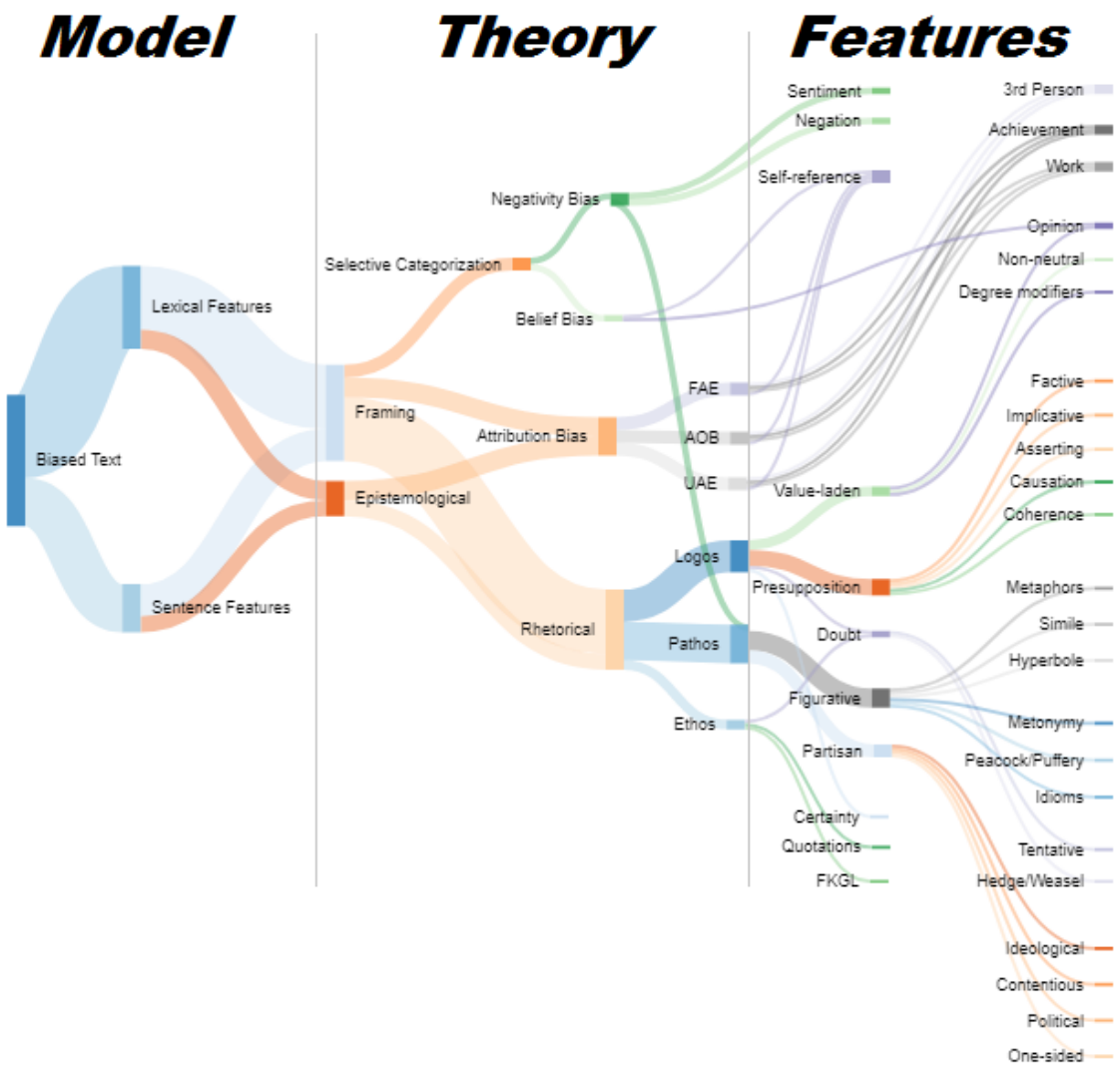


Figure 19: Graphical summary of how individual features used in the Biased Sentence Investigator (BSI) computational model are drawn from theoretical underpinnings from psychology, computer-mediated communications [CMC] research, and mass media communications studies.

5.4.3 Coverage Bias at the Sentence Level: CASTER

To extend the BSI model's capability for assessing bias of sentence level text, I also develop a simplified topic-modeling approach I refer to as Contextual Aspects Summary and Topic-Entity Recognition (CASTER). The CASTER module of BSI extracts important keywords, topics, and entities from text at the sentence level, in contrast to other

computational topic modeling techniques like Latent Dirichlet Allocation (LDA) [17], which relies on a corpus of documents to derive topics. CASTER uses NLP techniques associated with Named Entity Recognition (NER) [16], Part of Speech (POS) identification, and Word Sense Disambiguation (WSD) [3] to obtain word-order sensitive *n*-gram keywords and phrases that summarize the contextual entities and topics related to bias measurements produced by the BSI computational model.

CASTER enables investigations of coverage and gatekeeping bias at the article and corpus units of analysis (as demonstrated in Section 5.8.3). However, prior to demonstrating applications of BSI and CASTER, it is worth describing in more detail (1) how the BSI model compares to, and is distinct from, existing similar efforts along the same lines, as well as (2) an iterative investigation of BSI's feasibility for detecting and quantifying bias in news stories. I address the former of these in the next section, and the latter in Sections 5.6 – 5.7.

5.5 Existing Computational Approaches for Measuring Bias

Readers and broadcast news consumers have some intuition of media bias. For many people, though, it is both cognitively challenging and time-consuming to be aware of the particular biases of all media outlets, let alone be consistent in objectively quantifying them or understanding their relative magnitudes. Computational techniques can address both the issue of scale and the issue of consistent measurement and quantification. A number of relatively recent research studies employ computational techniques to detect and quantify *gatekeeping*, *coverage*, and/or *statement* bias. Many of these earlier investigations rely on natural language processing techniques similar to the

ones I incorporate into the BSI computational model, though most only focus on a single measure or else just a few measures. To address this gap, my approach is to capture the features from across these isolated studies and integrate them into a single measurement model for media bias. This section describes prior work which (holistically, in aggregate) lends further support the features in my BSI computational model, and helps situate the BSI model with respect to relevant existing literature relying on similar computational modeling techniques.

Gentzkow and Shapiro [69] investigate overall coverage bias of U.S. newspapers by constructing an index of media slant that measures the similarity of a news organization's text to the typical language use by Congressional Republicans or Democrats (based on speeches and sponsored legislation recorded in the 2005 Congressional Record). The authors in [69] apply the technique to understand economic drivers of media slant – specifically, whether ideological bias of news outlets were driven by *audience* or *owner* preferences. The study suggests a strong correlation between news content and the political inclinations of the readers, implying that news outlets offer news perspectives which cater to their audience in order to maximize profits, whereas owner influence was not significant. While [69]'s techniques inform my method for measuring lexical indicators of partisanship (I directly include [69]'s lists of Republican and Democrat affiliated words and phrases), it is but one indicator among a variety of features I use to detect and quantify bias. Additionally, my model focuses on finer grained analysis of the article text – i.e., I measure bias at the sentence level, which allows for assessments of coverage bias *within* a single article rather than limiting the coverage bias analysis to comparisons *between* articles or *across* newspapers, though my techniques easily extend to those types of analyses.

Applying [69]'s technique for detecting coverage bias for the two-party U.S. political system, and extending it to the multi-party political climate of Germany, Dallman and colleagues focus on a comparative analysis to identify the relative political biases of four leading online English news sites in Germany [41]. To assess coverage bias, the researchers incorporate two measures of statement bias: (1) the sentiment of the four words both preceding and subsequent to a mentioning of any German political party or party (parliament) member, and (2) the cosine similarity of the vocabulary of the sentence to the various lists of party-affiliated ideological terms. In a similar line of work organizing the political affiliation of Chilean media outlets [57], researchers quantify bias according to an outlet's affinity (i.e., positive/negative polarity associations) towards vocabulary capturing personal and economic issues for a multidimensional characterization of political partisanship. In contrast, while my model also incorporates lexical considerations of partisan language and gives sentence level attention to measuring sentiment, I also account for several other lexical level and sentence level features. My BSI model is therefore useful for measuring a myriad of other types and forms of bias in addition to political slant.

Research by [186] analyzes statement bias by investigating user-generated edits made to Wikipedia pages which had been tagged as violating Wikipedia's Neutral Point of View (NPOV) policy. The research by [186] is both informative and insightful, motivating many of BSI's features for detecting framing bias and epistemological bias, including: factive verbs, implicative verbs, assertive markers, hedges, degree modifiers, and one-sided (partisan) words. In addition to a similar language-based approach, [97] also includes measures of gatekeeping bias and coverage bias for specific *topics* that are either edited out (gatekeeping) or else given differential amounts of attention (coverage) within the

article. However, whereas both [97] and [186] focus on identifying specific words, phrases, or topics that signal bias in encyclopedic reference articles, my work is distinct in that I am interested in quantifying the *degree* (or intensity) of such bias in the context of *news stories*, which – as with encyclopedic reference articles – similarly strive for impartiality.

In [158], the authors present an unsupervised model based on how news outlets quoted presidential speeches. The study shows how quotation patterns can indicate gatekeeping, coverage, and statement biases that capture systematic (biased) perspectives of media outlets. For example, consider the following, as highlighted by [79]:

> The editors in Los Angeles killed the story. They told Witcover that it didn't 'come off' and that is was an 'opinion' story….The solution was simple, they told him. All he had to do was get other people to make the same points and draw the same conclusions and then write the article *in their words* (emphasis in original). — Timothy Crouse, Boys on the Bus [1973, p. 116].

The research by [158] show how the media's quoting behaviors were found to roughly align along a first latent dimension representing the traditional left (liberal) right (conservative) political ideology spectrum and a second latent dimension characterized by a continuum for mainstream versus independent news organizations. The research presented in this chapter is less focused on quantifying left/right ideology or mainstream versus independent news organizations, and instead examines the degree to which quotes might be used as rhetorical framing devices to boost credibility (ethos) and increase perceptions of a writer's own objectivity (pathos), irrespective of whether the quotes actually reflect selection bias (gatekeeping) to convey an implicit narrative (perspective).

Another interesting line of research involves automated support for finding and presenting different perspectives on selected news topics [83,153,156,166–168]. In [83], researchers describe *NewsBird*, a news aggregator that presents international news topics in a format that allows readers to explore various aspects (clusters) of articles matching user-defined search queries. *NewsBird* does not attempt to detect or quantify media bias, per se, but rather aims to mitigate the effects of coverage bias by presenting a broader range of news perspectives for a selected topic. Likewise, Park and colleagues [166–168] designed a system called *NewsCube* that groups articles on similar topics into clusters that reflect different sub-topics (or "aspects") defined by the appearance of co-occurring words. During laboratory experiments, users read more stories and explored more aspects on each topic when using *NewsCube* compared to Google News or an interface that grouped stories randomly [166]. Browser based extensions such as *BS Detector*[30] and *Balancer* [153] extend this concept to users in the wild in order to encourage news consumers towards more diverse political news coverage. Narwal and colleagues [156] incorporate consideration for how visual multimedia is used to convey framing biases. They collect images associated with a news story, use crowd "activists" to find contrasting visual representations, and then present a collage of these images all together to capture the diversity of visual perspectives of news stories. Such (exclusively) *topic*-oriented tools can be further informed by more direct characterizations of other forms of bias associated with news stories. The BSI computational model presented in this chapter can address such gaps.

[30] https://github.com/bs-detector/bs-detector

In [193], researchers employ metrics related to all three forms of media bias (gatekeeping, coverage, and statement bias) in order to characterize the partiality associated with "news communities" derived from online social media networks. They quantify *gatekeeping* bias by determining which media/communities do not cover certain stories, *coverage* bias according to the amount of attention a particular story or person is given, and *statement* bias by computing the sentiments in sentences mentioning different people. These measures of gatekeeping, coverage, and statement bias are straightforward, but perhaps too simplistic. The work by [193], as well as literature discussed previously in this section (e.g., [41,57]), seems to suggest that a model based only on sentiment analysis has viability for being a metric for statement bias. Can a simpler model capture the nuance of framing and epistemological biases in text? If so, then simplicity is preferred. The research presented in this chapter addresses this question.

Another approach is to compute a media bias score based on citation networks. For example, [79] links news outlets that cite think tanks and policy groups with similar citations by Congressional members with known liberal or conservative biases in order to derive a measure of political slant for numerous media outlets. Their research employs a widely accepted measure of political position using the database of liberal/conservative scores obtained from Americans for Democratic Action (ADA) [4]. ADA defines a key set of votes that indicate either strong liberal or conservative positions, and uses a Congressperson's voting record to assign a score ranging from 0.0 (most conservative) to 1.0 (most liberal). News organizations then exhibit bias with regards to whether or not they cite certain think tanks and policy groups (gatekeeping bias), and if so how often (coverage bias). Similarly, a pair of related studies by Lin and colleagues builds a citation network

model based on social media (Twitter) mentions of congressional lawmakers [138,139]. These studies avoid the computationally difficult task of identifying bias in the text of news content itself, and instead focus on quantifying coverage bias according to the attributes of those being cited. These techniques are not without merit, but I posit that the analyses of coverage bias would be strengthened if they were informed, for example, by the bias characterizations produced by a model like the one described in this chapter for not only detecting the presence of biased text, but also for estimating the magnitude of the bias.

5.6 BSI: Preliminary Feasibility Evaluation

As part of a larger overall effort unrelated to this dissertation, my colleagues at GTRI conducted a survey of 91 people to investigate factors that influence the perception of bias in fictitious news stories [65]. During this process, human subjects provided ground-truth gold standard ratings for the degree of perceived bias on a scale from 0 to 3 (representing perceptions of unbiased, slightly biased, moderately biased, or extremely biased) for forty-one sentences across five separate fictitious (but realistic and representative) news articles. I was able to leverage the dataset of sentences and bias-ratings produced by that effort to conduct a preliminary feasibility study of an early version of the Biased Sentence Investigator (BSI). In this section I present a summary of the initial evaluation for the BSI model, and compare its performance against gold standard human judgements of perceived bias in news-like text. (A more detailed account of the preliminary study is reported in [104]).

5.6.1 Dataset of Biased and Unbiased Text from News-Like Stories

Some datasets used previously to quantify bias consisted of texts that typically take an overt stance (such as congressional records, debate transcripts, or editorial news) [69,137,169]. In contrast, I desire the capability to gauge bias even within the much more subtle domain of *journalistic* news, i.e., so-called "objective" news reports. In [65], people rated Presidents Bush and Obama on 25 adjectives and were then randomly assigned to read five realistic (but fictitious) news stories about one of them. Three of the stories described positive outcomes, and two described negative outcomes. In every story, one sentence was randomly manipulated to attribute the outcome to either an *internal* trait of the president or to *external* factors in an effort to observe the effects of moderating and mediating aspects of attribution bias associated with UAE, whereby individuals typically assign greater attribution to internal/personal factors for positive outcomes when the person is someone they like, and to external/situational factors if the outcome is negative.

As part of the study, ninety-one people were surveyed. Participant demographics were skewed somewhat toward male (about 60%) and young adults under age 40 (over 50%). The political attitudes of the participants were of primary interest to [65], though, in particular, attitudes toward Presidents George W. Bush and Barack Obama. About two thirds of the sample had positive opinions about Obama and negative opinions about Bush, and one third exhibiting the opposite pattern. Participants were randomly assigned to provide ratings of one president first (Bush or Obama), followed by ratings of the second. Their responses were then used in a stratified sampling strategy to assign participants to read the five fictional news stories using either the name of the president they viewed most positively or most negatively (and 4 individuals who were neutral to both men were randomly assigned). Across the five stories, the story "target" remained the same once the

participants were assigned to read about either Bush or Obama. The study balanced the presentation order for the five stories to mitigate potential ordering effects. An example news story is presented below:

According to Forrester Research, an estimated 200,000 American jobs are lost annually due to offshore outsourcing. While in the past it was predominantly blue-collar jobs and low-level white-collar jobs that were relocated, the data show even mid- to high-level white-collar jobs are now being outsourced. During ***{Bush/Obama}****'s presidential campaign, he maintained outsourcing is a part of globalization, which will be good for the American people in the long run. High unemployment rates led to growing public condemnation of outsourcing and demand for new regulations to stop or limit outsourcing. In response, corporations increased lobbying efforts to defend their ability to outsource jobs overseas, which they argued is necessary in order to remain competitive with international firms. Ultimately, President* ***{Bush/Obama}*** *rejected the proposal to implement trade protection policies that would discourage outsourcing. The President dismissed the proposal mainly because of…"*

"… ***his unwillingness to stand up to corporate special interests****."(internal attribution)*

OR

"… ***intense pressure from corporations****." (external attribution)*

This first story was about a financial situation where the outcome was negative. The other four stories reported about:

1. The president's decision to eliminate a federal grant program for teachers who would no longer receive incentive grants to work in inner-city school districts due to budget concerns (a negative outcome).
2. The president's promise to seek funding to support better emergency planning efforts, particularly those aimed at assisting with disaster preparedness for individuals with disabilities (a positive outcome).
3. The president's pledge to improve healthcare services to veterans (a positive outcome).
4. A successfully foiled bioterrorism attempt to smuggle aerosolized Ebola virus aboard an airplane in New York City (also a positive outcome).

5.6.2 Human Judgements of Bias in Unattributed News Stories

Participants in [65] first read an entire story in paragraph form, and then were presented each sentence one a time and asked to rate how biased they believed each statement to be. Response options consisted of a 7-point balanced rating scale, with an option for a neutral rating ([–3] *Extremely* biased AGAINST Bush/Obama, [–2] *Moderately* biased AGAINST Bush/Obama, [–1] *Slightly* biased AGAINST Bush/Obama, [0] Fair and Impartial, [+1] *Slightly* biased IN FAVOR of Bush/Obama, [+2] *Moderately* biased IN FAVOR of Bush/Obama, or [+3] *Extremely* biased IN FAVOR of Bush/Obama). In BSI, I aim to quantify the degree of bias (rather than the polarity). I therefore simplify this training dataset by using the absolute value of the coded responses. Thus, bias ratings in this training data are continuous, ranging from 0-3 based on the average of 91 human judgements using the following rating anchors: 0 (unbiased, neutral), 1 (slightly biased), 2 (moderately biased), and 3 (extremely biased biased).

5.6.3 Detecting and Computing Degree of Bias in News Stories

As seen in the example text in Table 11, some sentences of the news story are clearly perceived by human judges as being biased ([65] intended to subtly induce either internal or external attribution bias by manipulating the final two sentence options):

Table 11: Mean (and Standard Deviation) of 91 human-judgments of perceived bias (scale: 0=Unbiased, 1=Slightly, 2=Moderately, and 3=Extremely Biased).

	Sentence Level Text (for sentences from the first news story)	Mean (SD)
1	According to Forrester Research, an estimated 200,000 American jobs are lost annually due to offshore outsourcing.	0.10 (0.42)
2	While in the past it was predominantly blue-collar jobs and low-level white-collar jobs that were relocated, the data show even mid- to high-level white-collar jobs are now being outsourced.	0.11 (0.46)
3	During Bush/Obama's presidential campaign, he maintained outsourcing is a part of globalization, which will be good for the American people in the long run.	**0.71 (1.00)**
4	High unemployment rates led to growing public condemnation of outsourcing and demand for new regulations to stop or limit outsourcing.	0.20 (0.64)
5	In response, corporations increased lobbying efforts to defend their ability to outsource jobs overseas, which they argued is necessary in order to remain competitive with international firms.	0.12 (0.51)
6	Ultimately, President Bush/Obama rejected the proposal to implement trade protection policies that would discourage outsourcing.	**0.70 (1.04)**
7e	The President dismissed the proposal mainly because of intense pressure from corporations.	**1.35 (1.22)**
7i	The President dismissed the proposal mainly because of his unwillingness to stand up to corporate special interests.	**1.90 (1.21)**

At the sentence level unit of analysis of the stories, I observe characteristics of the text statement as a whole, considering a total of 26 initial feature vectors capturing syntactical, grammatical, and lexical properties of sentences, and then iteratively refining the statistical model through variable selection activities.

Feature selection at this phase involved assessing both forward and backwards stepwise Akaike Information Criterion (AIC) [22] to measure the relative quality of each feature for characterizing the degree of bias in text. AIC is founded on information theory: it estimates the relative quality of statistical models for a given set of data by assessing the information lost (or gained) when comparing between models. In forward and backwards stepwise AIC, the criterion is used to judge the information lost between statistical models

that iteratively add or remove features, seeking to balance the trade-off between the goodness of fit and the simplicity of the model. While the AIC is theoretically distinct from the Bayesian Information Criteria (BIC), both address relative model comparisons and model selection. The AIC or BIC for a model is usually written in the form [-2log$L$ + kp], where L is the likelihood function, p is the number of parameters (degrees of freedom) in the model, and k is 2 for AIC and log(N) for BIC (where N is number of observations; i.e., BIC penalizes model complexity more heavily than AIC). Both AIC and BIC differ from Principle Component Analysis (PCA). PCA is a dimensionality reduction method that works by finding the most "meaningful" features in a larger model by assessing the "best" explanations of variance via combinations of features ("principal components") in covariate space (without explicitly considering information loss). All three approaches (AIC, BIC, and PCA) provided qualitatively similar results; in practice, the goals of the analysis drive my choice for AIC: keep as many features as possible while reducing the feature space to those features with the highest impact (the priority being oriented around minimizing information loss rather than strictly reducing to the simplest model). Thus, using step-AIC to reduce the feature space to the most useful and valuable predictors helps in several ways: (1) it helps mitigate the curse of dimensionality, (2) it simplifies the model and makes it easier to interpret, (3) it helps enhance generalization by alleviating the risk of overfitting, and (4) it does all this while also considering the trade-off of information loss. The refined model eventually comprised the following 14 initial features (see Section 5.4 for more detail):

1. VADER Sentiment score
2. Modality (certainty)
8. One-sided (partisan) terms
9. Opinion words

3. Factive verbs
4. Assertive verbs
5. Hedges
6. Strong subjective intensifiers
7. Weak subjective intensifiers
10. Tentative words
11. Third Person Pronouns
12. Achievement words
13. Work words
14. Discrepancy words

5.6.4 Preliminary Evaluation Results

The linear regression analysis for the preliminary 14-feaure model was significant: $F(14,26) = 11.3$, $p = 1.04e\text{-}07$, and accounted for over 85% of the variance in human judgements of bias ($R^2 = 0.859$) in the sample. Figure 20 depicts the proportion of overall R^2 that each feature accounts for, using the mean of three regression techniques (feature added to model first, feature added to model last, and feature beta squared). I find that an initial computational model motivated at first by [186]'s prior work on detecting biased language in reference articles is a useful start for determining the intensity (degree) of bias in news stories, but that additional lexical and sentence level features are also very useful – e.g., notice that five out of the top seven features shown in Figure 20 are features identified in the current research effort. Figure 21 shows the match between observed (measured) bias and the degree of bias predicted by the model using 10-fold cross-validation; the fit is remarkably good.

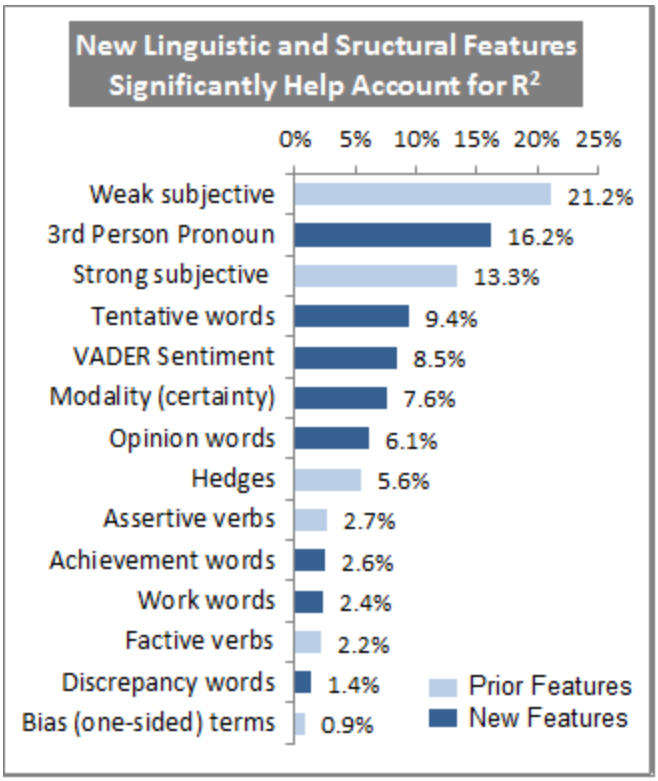


Figure 20: Proportion of variance accounted for by each feature in the improved model using the mean R2 of three regression techniques (feature added to model first, feature added to model last, and feature beta squared).

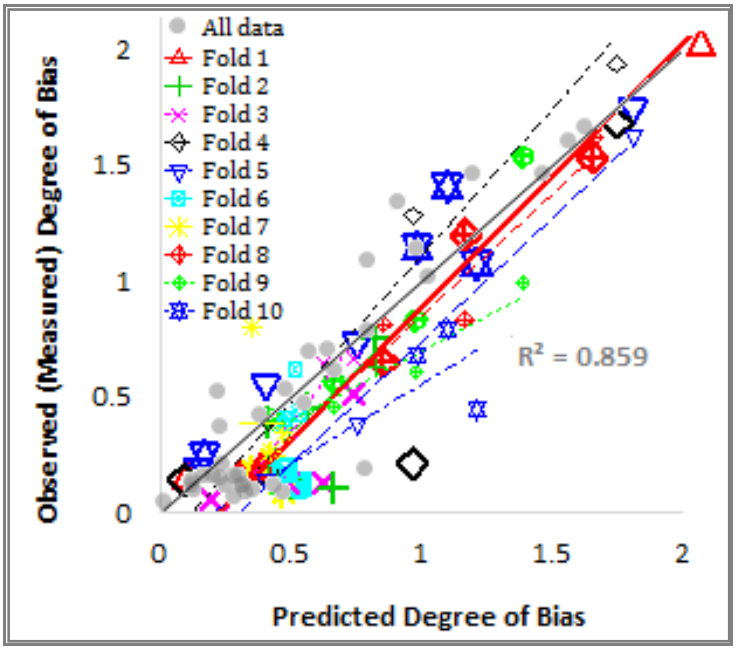


Figure 21: Results of 10-fold cross-validation analysis for fit between observed and predicted values of degree of bias in text.

5.7 BSI: Expanded Study

The preliminary feasibility evaluation described in Section 5.6 empirically demonstrates that the BSI computational model has at least some viability for detecting and quantifying perceived bias in news stories, but there are a few limitations to consider:

1. The stories in the preliminary study are only news-*like*. Although the stories appear realistic, they are nonetheless fictitious (all authored in a laboratory setting by a social science researcher, not a journalist). An expanded study would evaluate BSI against perceptions of bias for sentences from authentic, real-world news articles written by actual journalists.
2. The sample size is extremely small. A total of just 41 sentences are all that get processed during the small feasibility study. BSI needs more sentences from stories reported by a range of news outlets. It is important to obtain a range of (validly labeled) training samples that capture a spectrum of biased expressions.
3. The nature of the independent variable in [65]'s study made it so that the most biased sentence in the story was always specifically crafted to emphasize either dispositional (internal) or situational (external) attributes to test in-group/out-group perceptions of the UAE in assessments of news stories. This artifact of the dataset may not be as prevalent in real-world journalistic news stories.
4. The most biased sentence in all five news-like stories always appears at the end of the story as the last sentence. It is possible that participants in [65]'s study were conditioned first on several unbiased/neutral or just slightly biased sentences, making the final sentence seem comparatively far more biased. While BSI is intended to quantify bias from journalistic news, it should also be evaluated against

perceptions of bias for sentences from opinion-editorial (op-ed) stories, where biased sentences may not contrast as sharply in comparison to the surrounding text.

5.7.1 Expanded Dataset

To address the limitations of the preliminary feasibility evaluation, I obtained an expanded dataset consisting of 100 authentic, real-world news articles. These articles consisted of ten op-ed stories and ten journalistic news articles from across five different news outlets (BBC, CNN, Fox News, MSNBC, and the New York Times). I purposefully selected news outlets based on three criteria: (1) they represent a range along the spectrum of ideological preferences (Figure 22), (2) they have a reasonably large audience (Figure 23), and (3) they are generally more trusted than distrusted (Figure 23) [180]. Twelve human judges evaluated a total of 1,029 sentences, providing 12,348 manually labeled ground truth ratings of perceived bias. For news stories gathered from the wild, the average correlation of each judges' bias rating to the mean of all 12 judges was fair (r = 0.661), and I do not expect my BSI computational model to perform better.

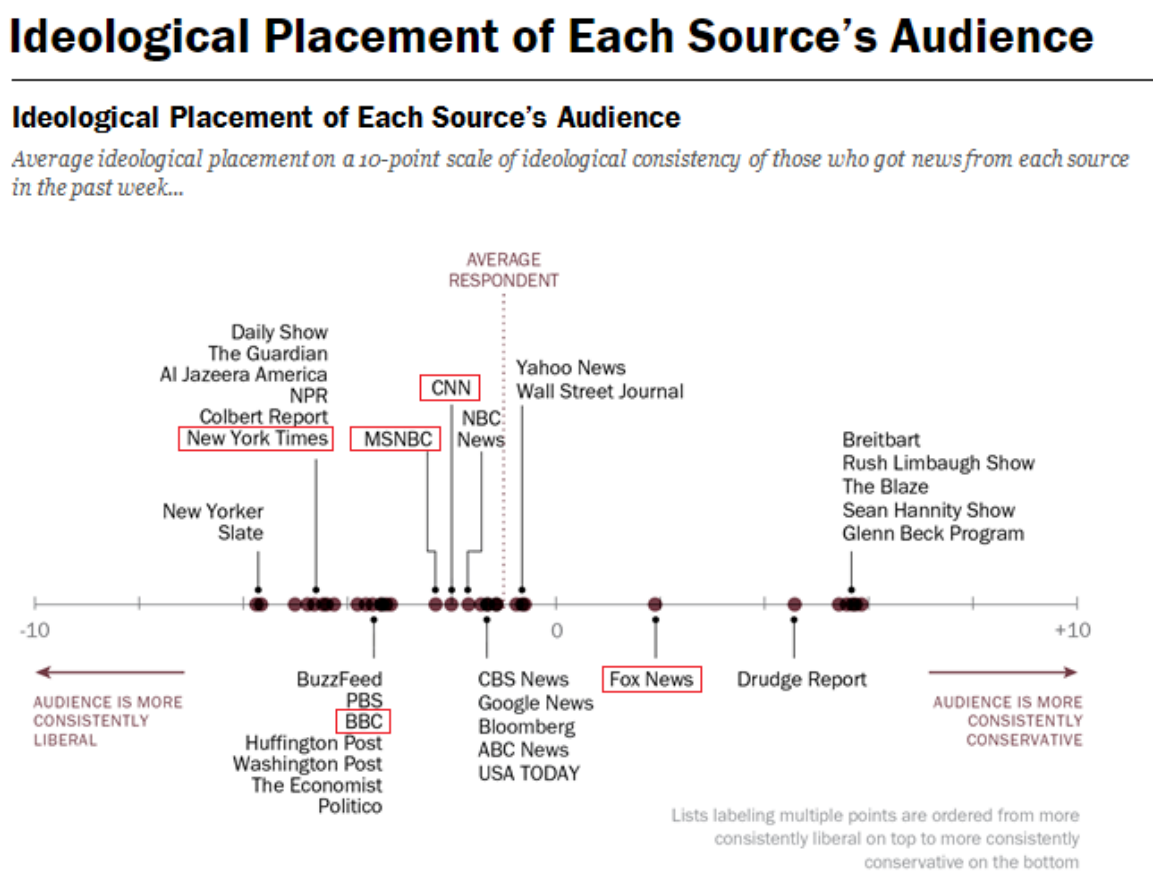


Figure 22: The five news outlets selected for the extended dataset represent a range of political ideological audience preferences [180].

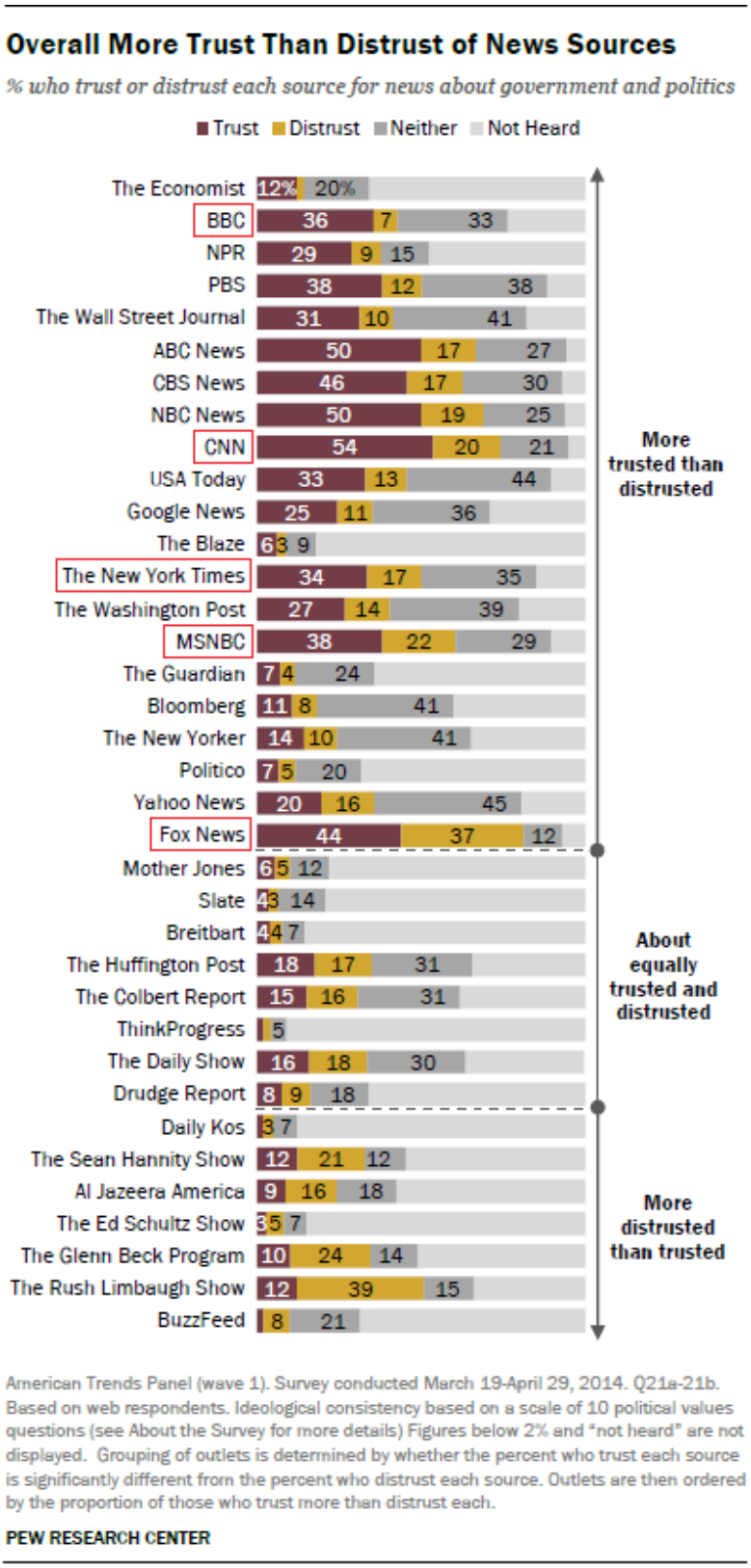


Figure 23: The five news outlets selected for the extended dataset have sizable audiences, and are generally more trusted than untrusted by most Americans [180].

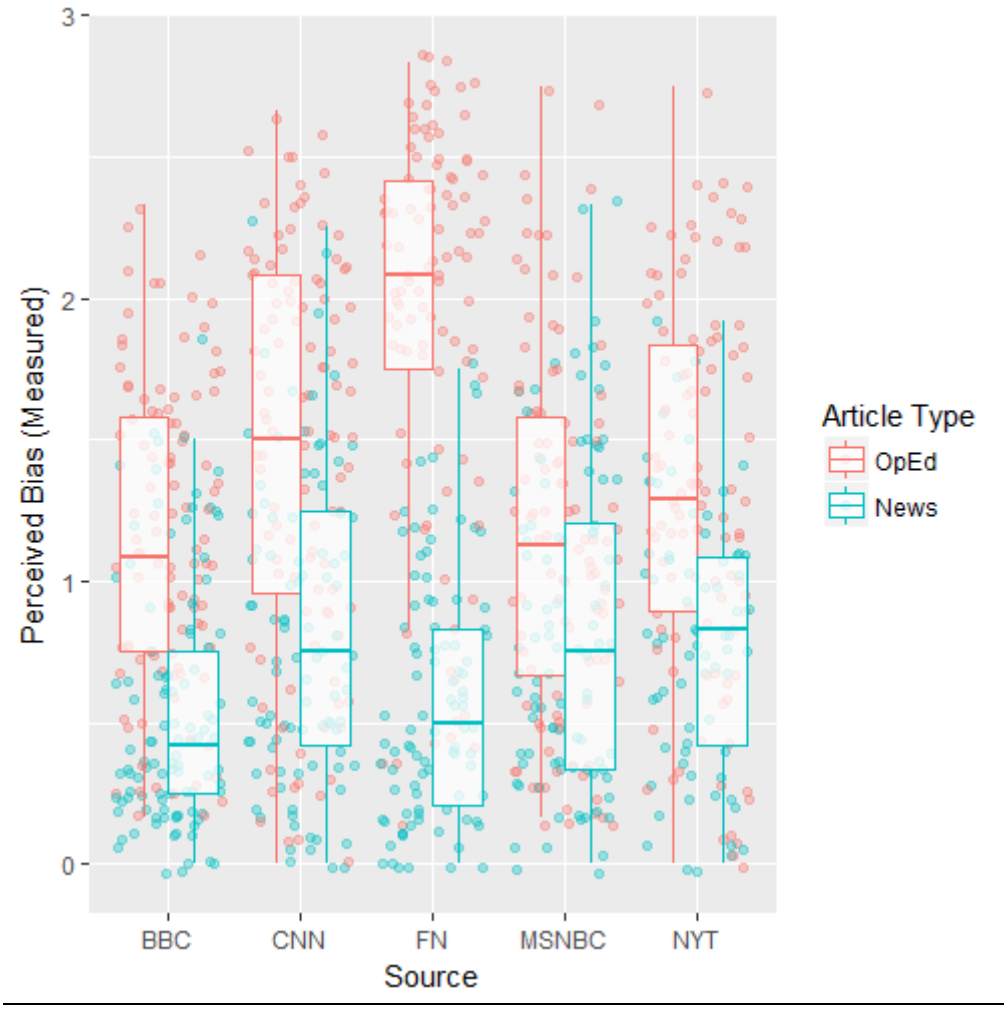


Figure 24: Sentences from opinion-editorial news stories are generally percieved as being more biased than sentences from journalistic news articles.

While the selection of news outlets was purposeful, the selection of current event op-ed and news articles from these sources was generally random. A brief inspection of the measured bias data shows that opinion-editorial stories are generally perceived as being more biased than journalistic news, as one might expect. Figure 24 is a box-and-whiskers plot of perceived bias for each article category (op-ed or news) for each news source. The "box" visually illustrates the middle 50% of the data, the vertical "whiskers" reflect the maximum and minimum points (showing the range), and the horizontal bar within each

box is the mean perceived bias score for each outlet source and news category. While not the focus of the research in this chapter, it is interesting to note a sharp distinction between the perceived bias of op-ed stories versus journalistic news for BBC and Fox News, while the other sources exhibit (sometimes a great deal of) overlap in the perceived bias of op-ed stories and journalistic news. Given the relative amount of distrust for Fox News reported by the survey respondents depicted in Figure 23, it raises the question of whether most news consumers are fully aware of the *type* of news for every article they read; by definition, op-ed stories are intentionally editorialized, and often reflect substantial biases. An application of the BSI model (see Section 5.8) could be to automatically cue readers to such distinctions. With the expanded dataset now described, I next turn to a detailed evaluation of the iterated feature set considered for an improved BSI computational model.

5.7.2 Expanded Feature Set and Feature Evaluation

Insights from the preliminary feasibility investigation presented in Section 5.6, together with the expanded literature review presented in Sections 5.2–5.4 of this dissertation, led to the generation, curation, and organization of an orthogonal set of lexical and sentence level measures. That is, all of the initial 14 features used in the feasibility study were integrated with 18 additional theory and literature-inspired features, and reorganized into the hierarchical factors as described in Section 5.4. In this section, I evaluate this refined feature set using ensemble voting methods for feature selection based on results of correlation matrix inspection for multicollinearity reduction, machine learning algorithms, and seven other measures of relative contribution to regression models.

Multicollinearity refers to the extent to which a variable can be explained by other variables in the analysis (i.e., high correlations between predictors). As multicollinearity increases, it complicates the interpretation of the variate because it is more difficult to ascertain the effect of any single variable, owing to their interrelationships [82]. Many statistical and machine learning models are susceptible to ill effects from multicollinearity. For example, linear regression models and neural networks can have poor performance in situations with multicollinearity [128]. (Other models, such as classification or regression trees, might be resistant to highly correlated predictors, but multicollinearity may negatively affect interpretability and training time for the model) [128]. As Hair and colleagues [82] point out:

> The simplest and most obvious means of identifying collinearity is an examination of the correlation matrix for the independent variables. The presence of high correlations (generally .90 and higher) is the first indication of substantial collinearity. (p. 196).

To check for opportunities for multicollinearity reduction, I use the `corrplot` package[31] from the R software environment for statistical computing and graphics[32] to create a correlogram. A correlogram graphically represents the correlation matrix for a set of variables, and is useful for visually highlighting the most and least correlated variables within a data table. Figure 25 depicts the correlogram for the initial set of features in the BSI computational model, and visually conveys that the model does not have any concerns about the presence of multicollinearity. More specifically, feature pairs with a positive

[31] https://cran.r-project.org/web/packages/corrplot/index.html

[32] https://www.r-project.org/

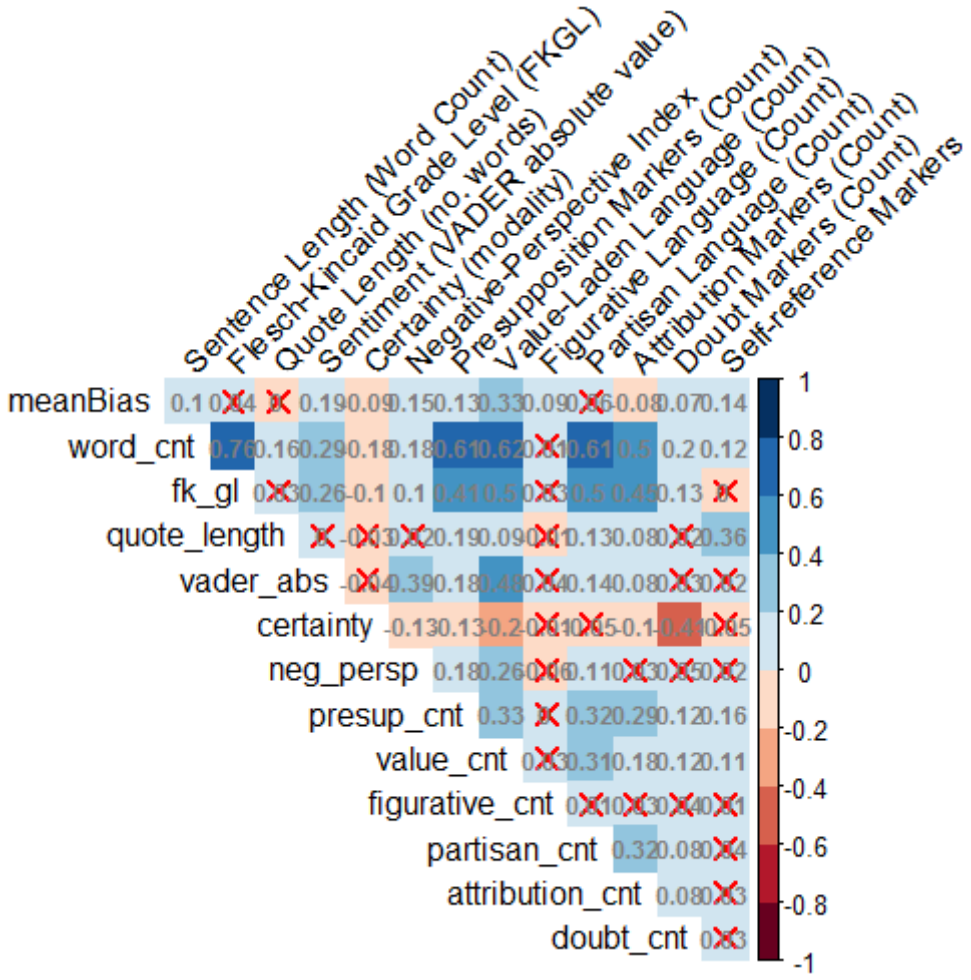


Figure 25: The correlogram graphs the correlation matrix for the initial set of features in the BSI computational model; multicollinearity is not a concern.

correlation coefficient have cells depicted in gradients of blue (darker shades indicate stronger correlations); negative correlations are likewise depicted in gradients of red. Feature pairs where the *p*-value for significance is less than 0.05 are marked with a red X. The ideal situation would be to have a set of features that correlate strongly with the dependent measure (e.g., I desire strong blue or red shaded squares, with few or no X's, for the mean perceived bias score on the top row), but do not correlate strongly with each other (lightly shaded blue or red squares, ideally with X's, on all rows beneath the top row).

While Figure 25 does show that no feature-pairs cause concerns regarding multicollinearity, it also shows that the readability of the text as measured using the Flesch-Kincaid Grade Level (*FKGL*) formula, the number of words in *quotes*, and the lexical markers for *partisan* language do not (on their own) significantly correlate with the mean bias scores. As with the preliminary study in Section 5.6, my goal is to err on the side of inclusion: I want to keep as many features as possible while reducing the feature space to those with the highest impact (the priority being oriented around minimizing information loss rather than strictly reducing to the simplest model). To this end, rather than simply eliminating the *FKGL*, *quote length*, and *partisan language* features out of hand, I next investigate the features in more detail and with more complexity.

In addition to examining the correlation matrix, filtering, reordering, and decomposition methods are also useful for isolating a given feature's contribution towards accounting for the proportion of R^2 in a multiple linear regression model [68,78,234]. The `relaimpo` package[33] in R leverages several such metrics for evaluating the contribution of any given feature, and I use an ensemble approach which averages seven of them. While these metrics are described elsewhere (c.f., [68,78,234]), I briefly introduce and summarize them here for context:

1. LMG: named for the statisticians who conceived the method (Lindeman, Merenda, and Gold) is the R^2 contribution averaged over all permutations of orderings for the features in a linear regression model [78].

[33] https://cran.r-project.org/web/packages/relaimpo/index.html

2. First: is each feature's R^2 contribution when it is included as the first regressor in a linear regression model.
3. Last: is each feature's R^2 contribution when it is included as the last regressor in a linear regression model.
4. β^2: is beta squared, the squared standardized beta coefficient for the feature.
5. Pratt: is the product of the standardized beta coefficient times the correlation of the specific feature with the dependent measure [78].
6. Genizi: is the contribution of a feature according to R^2 decomposition which considers the joint probability distributions of correlated regressors and the dependent measure [68].
7. CAR: is an abbreviation for Correlation-Adjusted (marginal) correlation, the contribution of a feature according to R^2 decomposition which encourages grouping of correlated predictors and down-weights antagonistic variables. CAR is described as an intermediate between marginal correlation and the standardized regression coefficient [234].

The LMG, First, and Last metrics are concerned with changes in R^2 based on the order in which a feature is processed into the regression analysis (variance–order oriented), whereas the Genizi and CAR metrics consider R^2 and various correlation measures within and among the features and depended variables (variance–correlation oriented). The β^2 and Pratt metrics are concerned with beta coefficients, rather than R^2 (i.e., effect size oriented rather than variance oriented). To assess variable importance, effect size, variance, and correlation are all important considerations; so, rather than privileging any single metric, I use an ensemble approach that takes the mean scores from all seven metrics. Figure 26

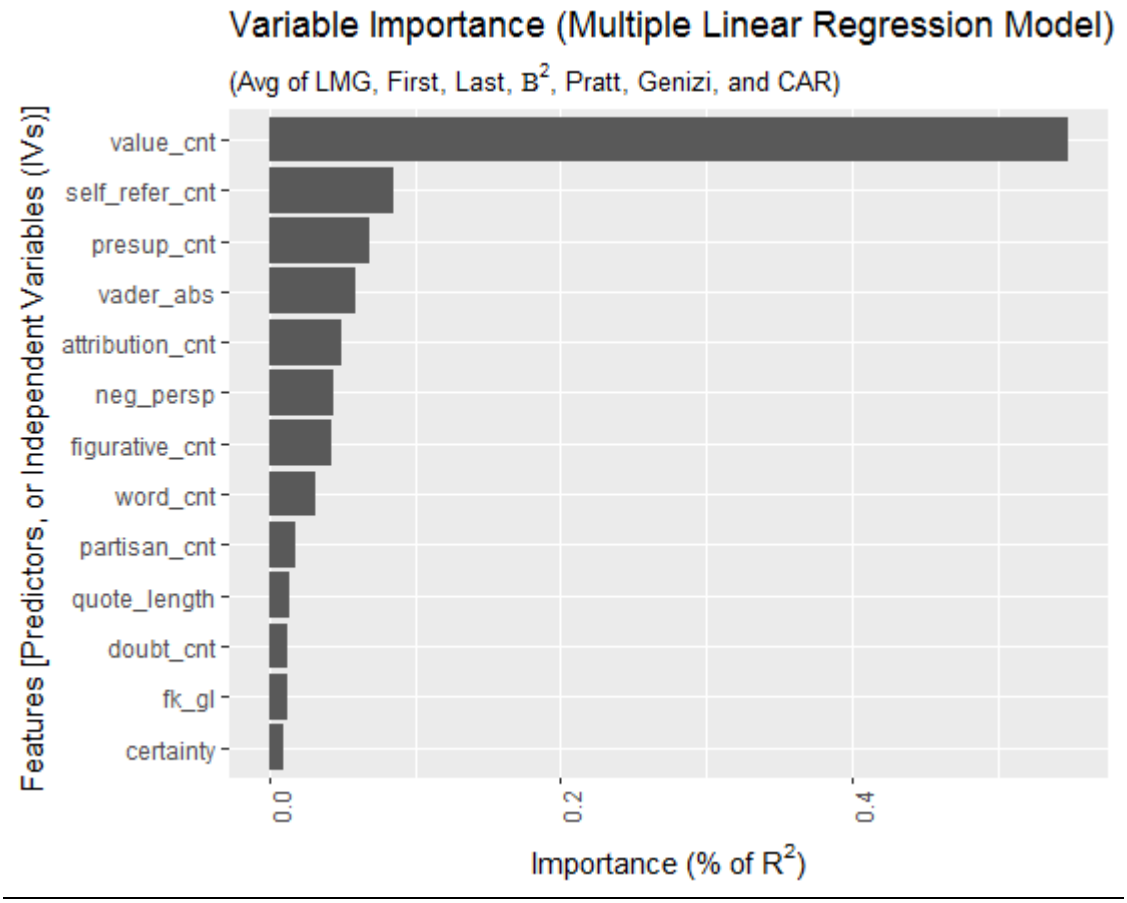


Figure 26: Relative feature importance according to the proportion of variance accounted for by each feature in a multiple linear regression model using the mean of seven regression-based feature evaluation techniques.

graphs the results of my ensemble approach. Using this approach, the sentence level measures for degree of expressed *certainty* and the Flesch-Kincaid Grade Level (*FKGL*), and the lexical measure for *doubt* laden language markers appear in the bottom three positions for variable importance.

Just as increasing the complexity of the feature evaluation from simply assessing the correlation matrix to examining variable importance via a more nuanced filtering and linear decomposition ensemble approach provided useful views into the relative value of each feature in the model, another transition in complexity to nonlinear methods will help

further triangulate on the relative importance of each feature. With this in mind, I next extend the feature evaluation into nonlinear space and evaluate each feature as it interacts in conjunction with other features [129]. I consider two nonlinear machine learning approaches for assessing the relative importance of all features (in conjunction). One uses a Support Vector Machine (SVM) with a polynomial kernel (degree=3) and 10-fold cross-validation, repeated three times. The other uses a random forest (RF) classifier to assess feature importance by comparing the relevance of the given (real) features to that of random decision-tree "shadow" probes.

SVMs are non-probability classifiers which operate by separating data points in space using one or more hyperplanes (centerlines of the gaps separating different classes). The `caret` package[34] for R offers an intuitive programming interface for training and plotting classification and regression models. Caret is short for Classification And REgression Training. Figure 27 graphically illustrates the importance of features according to their ranked *p*-values as computed by the `caret` package SVM model. In this model the least impactful features are the Flesch-Kincaid Grade Level (*FKGL*), *attribution* markers, and *quote* length.

[34] https://cran.r-project.org/web/packages/caret/index.html

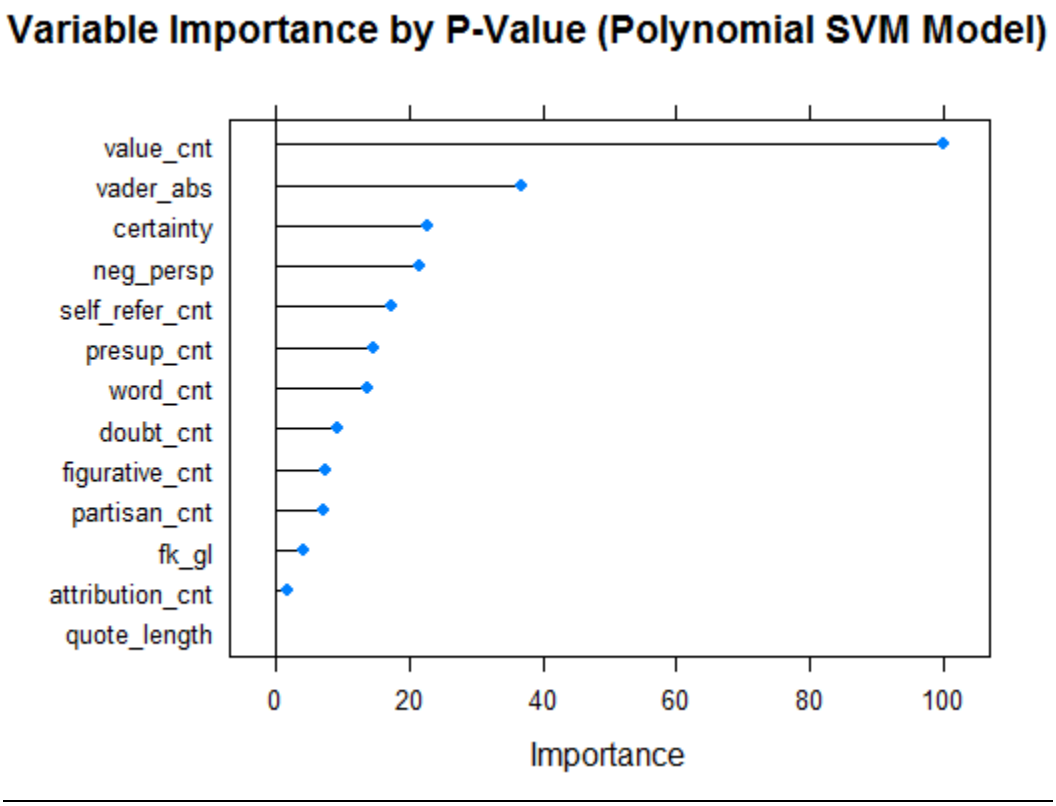


Figure 27: Relative feature importance according to the ranked *p*-values from an SVM model with a polynomial kernel (degree=3).

The RF machine learning algorithm is an ensemble method in which classification is performed by voting of multiple unbiased weak classifiers (i.e., multiple decision trees). The `Boruta` package[35] for R (named for a Slavic mythological god of the forest) implements a RF that is relatively quick, can usually be run without tuning parameters, and gives a numerical estimate of the feature importance when the feature is used in conjunction with other features [129]—all desirable qualities that aid with assessing feature relevance and feature selection. Boruta first works by duplicating the dataset, and then shuffling the values of each column—the values of the columns (feature) data remains the same, they just get randomly reassigned to different rows (observations). Boruta refers to these new

[35] https://cran.r-project.org/web/packages/Boruta/index.html

values as "shadow" features. Next, Boruta trains the Random Forest model on this dataset. In this way, Boruta measure the importance—via the Mean Decrease Accuracy or Mean Decrease Impurity—for each of the features of the data set. The higher the score, the better or more important the feature. Next, Boruta checks whether the original (real) features have higher importance than their corresponding "shadows", that is, whether the feature has a higher Z-score[36] than the minimum, average, or maximum Z-score of its shadow. If the real feature has a higher Z-score, Boruta records this as a "hit" and stores the result in a vector. Boruta continue these steps over several iterations, building up a table of hits for the features. At every iteration, the algorithm compares the Z-scores of the shuffled copies of the features and the original features to see if the latter performed better than the former. If it does, the algorithm will mark the feature as important, in essence, evaluating the importance of the feature by comparing it with numerous random "shadow" copies to increases the robustness. Figure 28 illustrates the relative feature importance according to the random forest ensemble. In this evaluation, all model features performed better than the random shuffle "shadow" probes (of which, the hits/importance for the shadow min, mean, and max are graphed). Lexical features associated with *partisan* and *doubt*-laden language markers were least favored by the RF model.

[36] The Z-score is the number of standard deviations from the mean for a given data point. If a Z-score is 0, then the data point score is identical to the mean score.

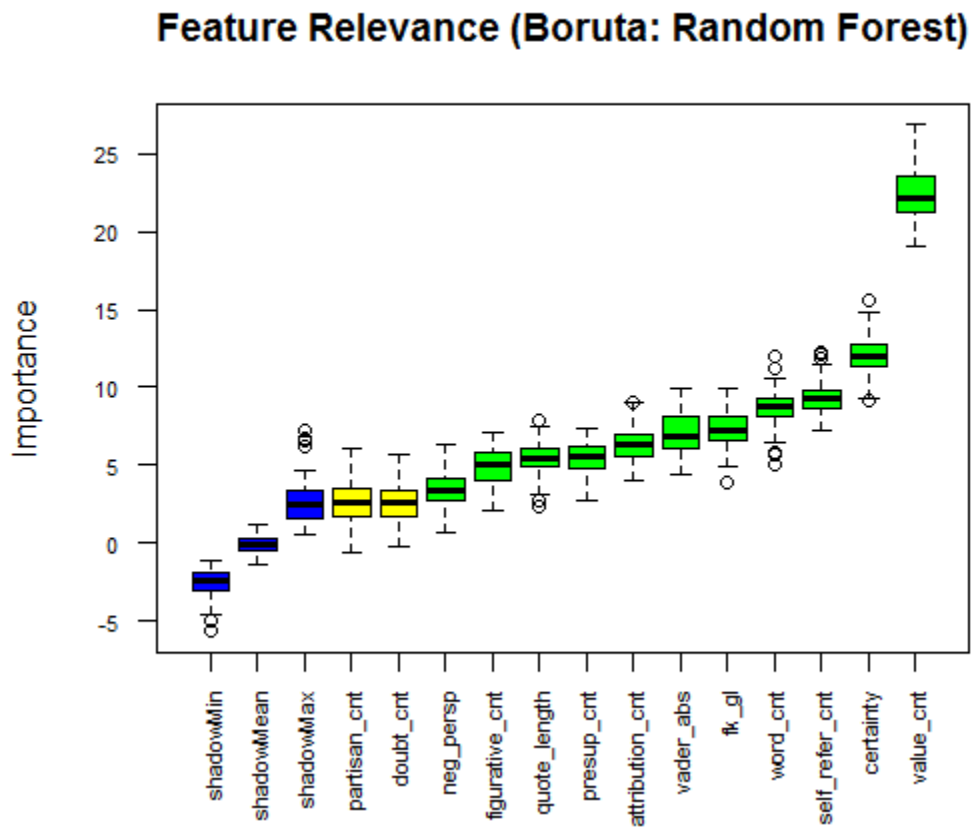


Figure 28: Relative feature importance according to an ensemble of unbiased decision trees (i.e., random forest).

There is definitely evidence that some features—such as value-laden language markers, sentence level sentiment intensity scores, self-reference language markers, presupposition markers, sentence length (word count)—are consistently strong predictors of perceived bias in news stories. Other features—such as sentence level certainty, FKLG, and quote length—ranked low in some evaluations, but higher in others. Given that all features appeared to have at least marginal relevance in at least one importance evaluation approach, and considering that the training sample is still relatively small compared to many other linguistics training sets in the literature, I choose to err on the side of inclusion and keep all of the theory-inspired features in the BSI computational model. In the future,

these features may be re-evaluated using the methods described above once more data are available that provide gold-standard pedigree training samples (i.e., sentences labeled with multiple human judgements regarding the degree of perceived bias). Having confidence that I have a viable set of predictive features, I now turn to evaluating various statistical and machine learning oriented prediction models.

5.7.3 Exploring Prediction Models: Linear, Non-Linear, and Machine Learning

In this section, I explore 26 different statistical and machine learning regression techniques to predict the perceived bias of sentences in news articles using the expanded dataset described in Section 5.7.1 and the feature set described in Section 5.4 (and evaluated in Section 5.7.2). The techniques explored include linear, non-linear, parametric, non-parametric, and machine learning oriented regression models. While mostly included for self-pedagogical purposes, this section will be useful for future efforts related to further development of the BSI computational model by documenting which types of regression techniques work best to predict perceived bias in news stories. Techniques range from multiple variations on linear regression models (LM, ENet, BGLM) to more complex nonlinear, non-parametric regressions (MARS, GAM, ICR, GP, GPRBF, GPPK, KNN), decision trees (CART, BCART, CIT, TGA), random forests (RF, PRF, RFR, RRF, CIRF, EGB), neural networks (NN, MLP, BRNN, ELM), and support vector machines (SVMRBF, SVMPLY). Table 12 captures a very brief description for each of these 26 statistical and machine learning modeling techniques to predict the perceived bias of sentences in news articles:

Table 12: Brief descriptions of 26 statistical and machine learning regression models used to predict perceived bias of news articles at the sentence level.

No.	Model	Abbr.	Description
1	Multiple Linear Regression (**baseline**)	LM	Attempts to model the relationship between two or more explanatory variables and a response variable by fitting a *linear* equation to the observed data.
2	Elastic Net Regression	ENet	A regularized regression method integrating both L_1 and L_2 penalties from LASSO and ridge methods to overcome the "large p, small n" problem - high-dimensional data with few samples.
3	Bayesian Generalized Linear Model	BGLM	A simple alteration of the GLM that uses an approximate EM algorithm to update the betas at each step using an augmented regression to represent the prior information.
4	Multivariate Adaptive Regression Splines	MARS	A non-parametric regression technique and can be seen as an extension of linear models that automatically accounts for nonlinearities and interactions between variables based on a "divide and conquer" strategy, which partitions the input space into segments, each with its own linear regression equation, such that the total model becomes nonlinear
5	Generalized Additive Model using Splines	GAM	Relationships between the individual predictors and the dependent variable follow smooth patterns (non-parametric functions) that can be linear or nonlinear, and these relationships can be added together.
6	Independent Component Regression	ICR	A computational method for separating a multivariate signal into additive subcomponents. This is done by assuming that the subcomponents are non-Gaussian signals and that they are statistically independent from each other.
7	Classification and Regression Tree	CART	A recursive partitioning method, builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The purpose of the analyses via tree-building algorithms is to determine a set of if-then logical (split) conditions that permit accurate prediction.
8	Bagged CART	BCART	Bootstrap Aggregation (or bagging for short), is a simple and powerful ensemble method. At each iteration the base classifier is trained on a fraction subsample of the available training data. The subsample is drawn without replacement.
9	Conditional Inference Tree	CIT	Roughly, the algorithm works as follows: 1) Select the input variable with strongest association to the response (as long as null hypothesis is rejected). Association is measured by a p-value corresponding to a single input variable and the response. 2) Implement a binary split in the selected input variable. 3) Recursively repeat steps 1) and 2).
10	Tree Models from Genetic Algorithms	TGA	Combines the stepwise search procedure of DTs (local optimization for attributes at a particular node, with no global perspective for optimization), with GAs fast global optimization pattern detecting using natural selection and crossover/mutation principles.
11	k-Nearest Neighbors	KNN	A non-parametric method where the input consists of the average of the k closest training examples in the feature space.
12	Random Forest	RF	RF is an ensemble method in which classification is performed by voting of multiple unbiased weak classifiers (i.e., multiple decision trees). RFs are an improvement over bagged decision trees.
13	Parallel Random Forest	PRF	Computationally, allows for parallel processing for RFs. Should be fairly close to the performance of RF (for R2, RMSE, and MAE).
14	Random Forest by Randomization	RFR	Dense randomness creates robustness against over-fitting. In extremely randomized trees, randomness is taken a step further for splits: as in random forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule.

15	Regularized Random Forest	RRF	Regularizing the RF by controlling/limiting the maximum depth parameter helps prevent overfitting.
16	Conditional Inference Random Forest	CIRF	Extends the Conditional Inference Tree (CIT) approach to RFs, creating an ensemble of CITs and aggregating their inputs.
17	Extreme Gradient Boosting	EGB	An implementation of gradient boosted decision trees designed for speed and performance. It builds the model in a stage-wise fashion, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
18	Neural Network	NN	Each neuron receives a number of inputs (either from training data, or from the output of other neurons in the neural network). Each input comes via a connection that has a strength (or weight). Each neuron also has a single threshold value. The weighted sum of the inputs is formed, and the threshold subtracted, to compose the activation of the neuron.
19	Multi-Layer Perceptron	MLP	MLP is one kind of neural network where the activation function is sigmoid, and error term is cross-entropy (logistics) error. A perceptron is always feedforward, that is, all the arrows are going in the direction of the output (no loops or recurring NN).
20	Bayesian Regularized Neural Networks	BRNN	Applies Bayesian inference techniques (EM) to regularize the NN in order to reduce overfitting.
21	Extreme Learning Machine	ELM	Feedforward neural networks with one or more layers of hidden nodes connected to the inputs by (constrained) random weights.
22	Gaussian Process	GP	Whereas BGLMs provide a probabilistic approach to regression by finding a distribution over the parameters that gets updated whenever new data points are observed, the GP approach is an alternative non-parametric approach, in that it finds a distribution over the possible functions f(x) consistent with the observed data.
23	Gaussian Process w/ RBF Kernel	GPRBF	RBF, as the name suggests, is a kernel that is in the form of a radial basis function.
24	Gaussian Process w/ Polynomial Kernel	GPPK	With degree=3, the polynomial kernel takes the form of a cubic.
25	SVM with Radial Basis Function Kernel	SVMRBF	SVMs are non-probability classifiers which operate by separating data points in space using one or more hyperplanes (centerlines of the gaps separating different classes). RBF, as the name suggests, is a kernel that is in the form of a radial basis function.
26	Support Vector Machines w/ Polynomial Kernel	SVMPLY	With degree=3, the polynomial kernel takes the form of a cubic.

To measure the performance of these 26 algorithms, I again use the `caret` package for training the models using 10-fold cross-validation repeated 3 times. Model performance evaluations are compared using three metrics: R^2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The MAE scores the average magnitude of the errors according to the absolute differences between predictions and actual observations, and is a measure of *accuracy* for continuous variables. RMSE, on the other hand, uses a quadratic scoring rule to measure the average magnitude of the error: it is the square root of the average of all the squared differences between predictions and actual observations. Since

the errors are squared before they are averaged, the RMSE penalizes large errors more than small errors. Both MAE and RMSE are indifferent to the direction of the errors, so over estimates and underestimates are indistinguishable. Both are negatively-oriented measures of error, which means lower values are better. Both metrics express average model prediction error in the same units as the dependent variable, and can range from 0 to ∞ (but bound in practice by the range of the dependent variable).

When larger errors are conceptually no more consequential than small errors and do not need to be penalized (as in our case), many researchers prefer the MAE over the RMSE due to its ease of interpretation and robustness [226]. Because MAE is measured in the same units as the dependent variable, it is worth recalling that bias ratings in the training samples are continuous, ranging from 0-3 based on the average of 12 human judgements with the following rating anchors: 0 (unbiased, neutral), 1 (slightly biased), 2 (moderately biased), and 3 (extremely biased). Figure 29 and Table 13 each contain much of the same information—the presentation of performance metrics for the prediction models shown graphically or in tabular format. Table 13 includes an additional measure that is not shown in the graph: the Pearson Product Moment Correlation Coefficient (r) between the predicted bias scores and the observed (measured) bias scores. In general, it appears that random forest oriented machine learning models have the highest R^2 and the lowest MAE, followed by SVM and Gaussian Processes (also with MARS, BRNN and ICR). The linear regression family of models generally occupy the central ordering, trailed by decision tree methods and neural network related techniques.

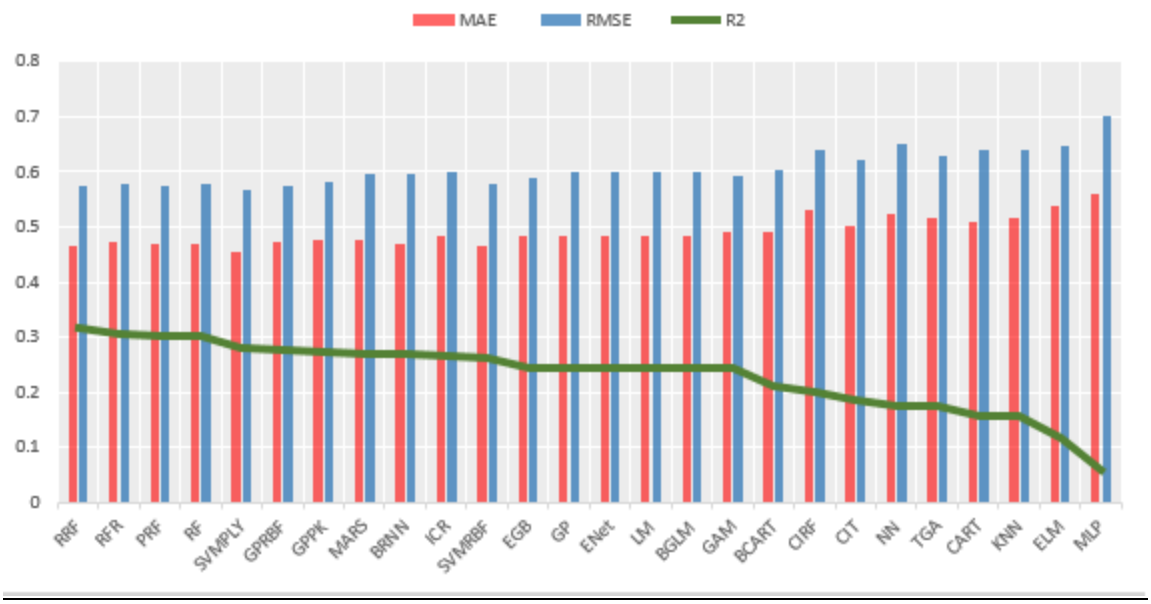


Figure 29: Graphical comparison of Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R^2 for 26 statistical and machine learning prediction algorithms

Recall that the average correlation coefficient of human judgements to the dependent measure in the expanded data set was $r = 0.661$. In comparison, the average correlation for predictions by the Regularized Random Forest (RRF) machine learning model ($r = 0.563$) and even for the simpler multiple linear regression model ($r = 0.495$) appear quite acceptable. Sentences garnered from real news stories in the wild seem to be very challenging (for both humans and machines) to consistently characterize in terms of bias. Compared to the fabricated news-*like* stories created in the lab—where the biased sentence contrasted sharply with the rest of the sentences in the story—ground-truth ratings for human perceptions of biased text was noisier for actual real world news articles. Because of these differences in the data, even the best random forest machine learning models in this expanded comparison did worse at predicting statement bias than the simpler linear regression model used in the small preliminary feasibility study described in Section 5.6.

Table 13: Comparison of 26 prediction models (ordered by R^2)

Model	Abbr.	R^2	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)	Pred. vs Obs. Correlation Coefficient (*r*)
Regularized Random Forest	RRF	0.317	0.574	0.466	0.563
Random Forest by Randomization	RFR	0.306	0.579	0.473	0.553
Parallel Random Forest	PRF	0.303	0.576	0.470	0.551
Random Forest	RF	0.302	0.576	0.468	0.549
Support Vector Machines w/ Polyn. Kernel	SVMPLY	0.281	0.568	0.453	0.531
Gaussian Process w/ RBF Kernel	GPRBF	0.278	0.576	0.474	0.528
Gaussian Process w/ Polynomial Kernel	GPPK	0.274	0.582	0.475	0.524
Multivariate Adaptive Regression Splines	MARS	0.270	0.596	0.475	0.519
Bayesian Regularized Neural Networks	BRNN	0.268	0.595	0.469	0.518
Independent Component Regression	ICR	0.266	0.599	0.485	0.516
SVM w/ Radial Basis Function Kernel	SVMRBF	0.262	0.577	0.466	0.512
Extreme Gradient Boosting	EGB	0.246	0.589	0.482	0.496
Gaussian Process	GP	0.246	0.599	0.484	0.495
Elastic Net Regression	ENet	0.245	0.599	0.484	0.495
Multiple Linear Regression (**baseline**)	LM	0.245	0.599	0.484	0.495
Bayesian Generalized Linear Model	BGLM	0.245	0.599	0.484	0.495
Generalized Additive Model using Splines	GAM	0.244	0.593	0.492	0.494
Bagged CART	BCART	0.211	0.605	0.490	0.460
Conditional Inference Random Forest	CIRF	0.202	0.641	0.532	0.449
Conditional Inference Tree	CIT	0.185	0.622	0.503	0.430
Neural Network	NN	0.177	0.651	0.525	0.420
Tree Models from Genetic Algorithms	TGA	0.175	0.629	0.515	0.418
Classification and Regression Tree	CART	0.159	0.640	0.509	0.399
k-Nearest Neighbors	KNN	0.159	0.639	0.517	0.398
Extreme Learning Machine	ELM	0.119	0.649	0.538	0.345
Multi-Layer Perceptron	MLP	0.059	0.700	0.559	0.242

Also, the average correlation coefficients (*r*) of each individual human judges' bias rating to the mean of all 12 judges was somewhat low ($r = 0.661$) when compared to typical measures of agreement among human judges for similar linguistic rating tasks, which often include correlation coefficients in the mid-to-high 0.80s [98,99,104]. Considering (1) the relatively small sample size of the training data even for the expanded dataset, (2) the range of diversity for linguistic expressions of both obvious and subtle biases that may be exhibited, and (3) the complex cognitive processes involved when humans attempt to estimate the degree of magnitude for such biases in sentences, it is no wonder that detecting

and quantifying bias in real world news stories is such a challenging computational task. The BSI model nevertheless seems to be a viable computational approximation.

5.7.4 Comparing BSI to a Parsimonious (Sentiment-Only) Model

In this section, I address the question of how well the more sophisticated BSI computational model compares to a simpler model based solely on sentiment analysis (for example, in the same vein as [41,57,193]). Can a sentiment-only model adequately capture the nuances of framing and epistemological biases in the text of news articles? If simpler models are not markedly worse than complex models, then parsimony should be the rule.

For ease of interpretation, and for fairness, I use a simple linear regression model for comparing the BSI model to the sentiment-only model for detecting and quantifying statement bias. Table 14 shows the features ordered by importance scores using the same

Table 14: Coefficients, error, *t*-test values, and *p*-values for a multiple linear regression using the BSI full model. $F(13,954) = 15.64$, $p < 2.2e\text{-}16$. (Ranked by feature importance using the same ensemble regression-based metric depicted in Figure 26).

Feature	Importance	*b*	Std. Error	*t*-value	Pr(>|t|)
(Intercept)	n/a	0.8745	0.0704	12.4290	< 2e-16***
Value-Laden Language (Count)	0.5480	0.1195	0.0130	9.2170	< 2e-16***
Self-reference Markers (Count)	0.0850	0.1248	0.0329	3.7930	0.00016***
Presupposition Markers (Count)	0.0695	0.0478	0.0141	3.3940	0.00072***
Sentiment (VADER absolute value)	0.0594	0.0638	0.0872	1.7320	0.04644*
Attribution Markers (Count)	0.0496	–0.0384	0.0145	-2.6490	0.00821**
Negative-Perspective Index	0.0445	0.0601	0.0391	1.5370	0.09125'
Figurative Language (Count)	0.0424	0.2738	0.0950	2.8840	0.00402**
Sentence Length (Word Count)	0.0326	–0.0118	0.0041	-2.8870	0.00398**
Partisan Language (Count)	0.0176	0.0311	0.0157	1.9740	0.04868*
Quote Length (no. words)	0.0138	–0.0092	0.0039	-2.3520	0.01889*
Doubt Markers (Count)	0.0135	0.0473	0.0349	1.3580	0.17481
Flesch-Kincaid Grade Level (FKGL)	0.0134	–0.0105	0.0063	-1.6690	0.09544`
Certainty (modality)	0.0106	–0.0138	0.0618	-1.2230	0.08233`

Significance level codes: $p < 0.001$*** $p < 0.01$** $p < 0.05$* $p < 0.1$`

ensemble-based metric depicted in Figure 21, the regression estimates (*b*), standard error, *t*-values, and *p*-values for a multiple linear regression on the full BSI model. Furthermore, the multiple linear regression statistical model also highlights a negative relationship between the dependent variable for perceived bias in sentences of news articles and the features associated with *attribution markers*, *sentence length*, sentence readability (as indicated by the *Flesch-Kincaid Grade Level* score), and the degree of *certainty* expressed in the sentence. It makes sense that as language reflecting greater certainty increases, perceived bias is reduced. The negative relationship for quote length is interesting, as it indicates that as the number of words for quoted material increases, humans typically lend greater credibility (or at least less bias) to the sentence. Similarly interesting is that attribution markers show a negative relationship: increased presence of operationalized indicators for FAE, AOB, and UAE are not perceived as more bias, but rather less biased. This is counter to the result discovered in the small preliminary feasibility study, where these type of sentence samples were explicitly constructed within the dataset. The sentences from the original five news stories in the feasibility study were not included in the dataset of 100 authentic, real-world news stories; thus, the expanded dataset may not have enough FAE, AOB, or UAE related training samples to detect attribution related bias (the negative relationship might instead reflect writers' attributing quotes to third persons, rather than exhibiting attribution bias based on dispositional or situational factors).

As Figure 30 and Table 15 indicate, there is a sharp decline in the average correlation of predictions to measured bias when reducing from the full BSI model (0.495) to a parsimonious sentiment-only model (0.315), as well as a substantial drop in R^2 when comparing the BSI full-feature model (0.245) to the sentiment-only model (0.099).

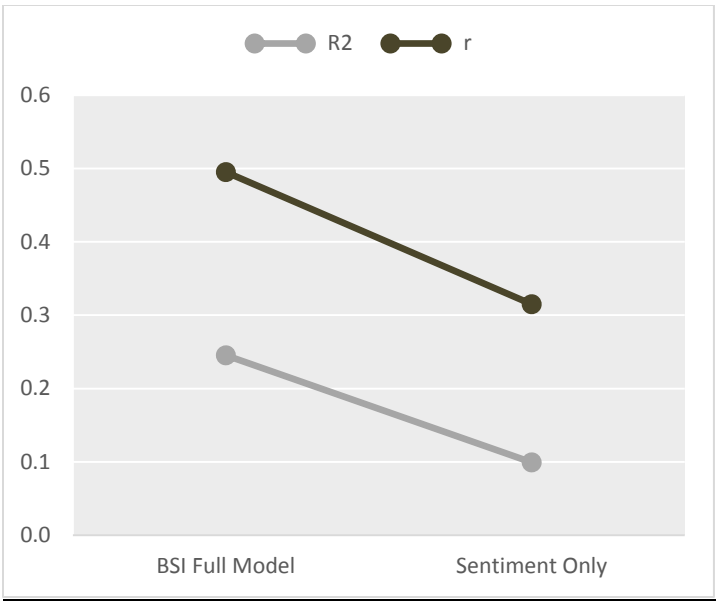


Figure 30: Comparison of BSI prediction model to a sentiment-only model.

Table 15: BSI prediction model compared to a model based solely on sentiment.

Model	R^2	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)	Correlation w/ mean of 12 Judges (*r*)
Biased Statement Investigator (BSI) Model of Statement Bias	0.2453	0.5987	0.4835	0.495
Sentiment-Only Model of Statement Bias	0.0992	0.6438	0.5368	0.315

A common technique for measuring statement bias in news articles used in prior work is to simply compute the sentiment of the sentence (c.f., [41,57,193]). However, the more nuanced full-feature BSI computational model is preferable to a simpler sentiment-only model as evidenced by beneficial decreases in measures of error (RMSE and MAE), as well as favorable increases in R^2 and correlation to human judgements for the BSI model.

5.8 Demonstration: Practical Applications of the BSI model

5.8.1 Statement Bias Computed at the Sentence Level of News Text

To demonstrate the types of analyses that BSI enables, consider the example sentences in Table 16, which were extracted from a news article reported by the Guardian, a news outlet located in the United Kingdom[37]:

Table 16: Example sentences from a Guardian news story with mean bias ratings (and standard deviations). Bias scale is continuous from 0 (neutral) to 3 (extremely biased).

ID	Sentence (source: The Guardian[37])	BSI Computed Bias scores	VADER Sentiment Only Scores
s1	British ministers including Theresa May and Philip Hammond have made hair-raising claims about the dangers of migrants entering the country.	0.996	–0.4939
s2	But do the facts bear them out?	0.875	0.0
s3	When you're facing the world's biggest refugee crisis since the second world war, it helps to have a sober debate about how to respond.	1.422	–0.7506
s4	But to do that, you need facts and data – two things that the British migration debate has lacked this summer.	1.176	0.0
s5	Theresa May got the ball rolling in May, when she claimed on Radio 4 that the vast majority of migrants to Europe are Africans traveling for economic reasons.	0.746	0.0
	Average (Std. Dev)	**1.043 (0.264)**	**–0.2489 (0.353)**

The BSI computed bias scores for these sentences appear to be sensible: sentence *s3* ("When you're facing…") has the highest relative bias, followed by *s4* and *s1*. Since the bias scale is continuous from 0 to 3, with verbal anchors set at each integer as: 0 (unbiased, neutral), 1 (slightly biased), 2 (moderately biased), and 3 (extremely biased), we observe that *s3* is computed to be roughly midway between *slightly* and *moderately* biased. Also reasonable. Note that the sentiment only model fails to capture the nuances of bias in either *s2*, *s4* (which BSI shows as having the second highest bias score in the set), or *s5*.

Although BSI computes bias at the sentence level unit of analysis, the example above demonstrates that it is straightforward to aggregate results and produce descriptive statistics at the article level to show how the tool extends to larger units of analysis. On the

[37] Available at: https://www.theguardian.com/uk-news/2015/aug/10/10-truths-about-europes-refugee-crisis

whole, the 5-sentence news story presented in Table 16 is just slightly biased, according to the average computed statement bias. Traditional (manual) qualitative content analysis to detect and annotate bias at the sentence level is time and labor intensive, and may be subject to differences in rubrics used to quantify the degree of bias. The BSI computational model enables rapid analysis on larger scales, at lower cost, and without the concern for inconsistent annotations.

5.8.2 Diagnostics: Exposing the Nature of Bias in News Stories

BSI also enables the process of systematically exposing indicators of bias and making its nature transparent. For example, I further leverage the individual feature values from a given sentence and combine them with their appropriate regression coefficient (estimated beta, as obtained from the multiple linear regression model) to assess explicitly *which* types and forms of bias have had the most impact on determining the computed bias score for any given sentence. Because features in the BSI model use different measurement units/scales for sentence versus lexical indicators, I first normalize each feature value using a logistic (sigmoidal regularizing) function so that all values are on the same scale (0 to 1, in this case), then multiply the normed feature value by the appropriate regression coefficient (beta), and finally scale the result up by a factor of 100 to improve interpretation ease. Equation [1] describes the function I refer to as the *Feature Impact Index*:

$$\boldsymbol{f(x_v, x_b) = \left(2 * \left(\frac{L}{1+e^{-k\left(abs(x_v)\right)}}\right) - 1\right) * \left(abs(x_b)\right) * 100} \qquad \textbf{Eq. [1]}$$

Where:

x = the selected feature of the BSI computational model

x_v = the computed value for feature x (observed for a given sentence)

x_b = the regression coefficient (beta) for feature x

L = the maximum value for the sigmoid curve (fixed at 1 for current use)

e = the natural logarithm base (Euler's constant, e.g., 2.718281828459…)

k = desired slope for the sigmoid curve (fixed at 1 for current use)

For example, Table 17 shows the Feature Impact Index for each of the five sentences in the example story as calculated based on the regression coefficients and the regularized (normalized) feature value:

Table 17: Feature Impact Index for each of the five sentences in the example story is calculated using a logistic (sigmoidal regularizing) function on a feature's observed value for a sentence, and then multiplying by the regression coefficient (beta).

		s1: "British ministers..."		s2: "But do the facts..."		s3: "When you're..."		s4: "But to do that, you..."		s5: "Theresa May got..."	
Feature	***beta***	**Value**	**Impact**	**Value**	**Impact**	**Value**	**Impact**	**Value**	**Impact**	**Value**	**Impact**
Value-Laden Language (Count)	0.1195	1	5.525	1	5.525	5	11.795	4	11.525	1	5.525
Self-reference Markers (Count)	0.1248	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000
Presupposition Markers (Count)	0.0478	2	3.641	1	2.209	2	3.641	2	3.641	2	3.641
Sentiment (VADER absolute value)	0.0638	0.49	1.545	0.0	0.000	0.75	2.290	0.0	0.000	0.0	0.000
Attribution Markers (Count)	-0.0384	0	0.000	1	1.776	0	0.000	0	0.000	2	2.926
Negative-Perspective Index	0.0601	0.64	0.000	<0.01	1.854	1.01	0.000	0.0	0.000	<0.01	2.805
Figurative Language (Count)	0.2738	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000
Sentence Length (Word Count)	-0.0118	21	1.181	7	1.179	24	1.181	21	1.181	28	1.181
Partisan Language (Count)	0.0311	2	2.367	1	1.436	3	2.813	3	2.813	2	2.367
Quote Length (no. words)	-0.0092	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000
Doubt Markers (Count)	0.0473	3	4.286	0	0.000	1	2.188	0	0.000	3	4.286
Flesch-Kincaid Grade Level (FKGL)	-0.0105	9.6	1.046	2.2	0.837	5.6	1.038	6.8	1.043	7.6	1.045
Certainty (modality)	-0.0138	1.0	0.638	1.0	0.638	-0.5	0.338	0.5	0.338	0.6	0.392

Using *s1* as the example, the presupposition and partisan language features had the same values (both had 2 lexical features counted), but because the regression coefficient for partisan language markers is about two-thirds that of presupposition language markers, the normalized Feature Impact Index of presupposition (epistemological bias) was almost 54% higher than that of partisan (ideological or belief bias) for *s1*. However, even with two lexical features counted, neither presupposition nor partisan language markers (nor doubt markers, with three lexical indicators counted) have as large of an impact on perceived bias as the single value-laden language marker in *s1*, illustrating the considerable influence of pathos-oriented rhetorical framing on perceptions of bias in news stories.

The practical advantages of systematically identifying the types and forms of biased expressions in text is that writers, editors, and publishers can self-assess news using a common tool to diagnose biased sentences in stories before publishing. Similarly for readers and news consumers, it may not always be apparent that a particular news story is intended to reflect an editorial stance or the writer's opinion. But when sentence and language markers of bias are identified and exposed, then readers, curated content providers, and media-monitor groups have the opportunity to become more aware.

5.8.3 Coverage Bias at the Sentence, Article, and Corpus Level

As described in Section 5.4.3, the BSI model also includes a capability for capturing the context of statement bias via a simplified topic-modeling approach I refer to as Contextual Aspects Summary and Topic-Entity Recognition (CASTER). The CASTER module of BSI uses NLP techniques associated with Named Entity Recognition (NER), Part of Speech (POS) identification, and Word Sense Disambiguation (WSD) to obtain

word-order sensitive *n*-gram keywords and phrases which summarize the contextual entities and topics related to the bias measurements produced by the BSI computational model. The keywords are scored and ranked according to their relevance within the text of the sentence (using TF*IDF to prioritize more important aspects). CASTER enables investigations of coverage bias at the article level by aiding analysts with determining which topics or people co-occur with what bias intensity scores and VADER sentiment (favorability) scores. A simple approach is to just compute the average for bias intensity and sentiment intensity for any contextual aspect that appears in more than one sentence in the article. A more complex approach might consider a distribution of bias intensity over contextual aspects using the CASTER aspect relevancy as a weighting factor. To demonstrate these concepts, Table 18 shows an example of using BSI (with VADER and

Table 18: Example of using BSI (with VADER and CASTER) to analyze coverage bias at the article level.

ID: Sentence	BSI Computed Bias Intensity	VADER Sentiment	Contextual Aspects Summary and Topic-Entity Recognition (CASTER) with TF*IDF aspect relevancy
s1: British ministers including Theresa May and Philip Hammond have made hair-raising claims about the dangers of migrants entering the country.	0.996	–0.4939	('philip hammond', 1.566) ('british minister', 0.711) ('theresa may', 0.522) ('hair-raising claim', 0.399) ('country', 0.369)
s2: But do the facts bear them out?	0.875	0.0	[no aspects extracted]
S3: When you're facing the world's biggest refugee crisis since the second world war, it helps to have a sober debate about how to respond.	1.422	–0.7506	('second world war', 1.422) ('sober debate', 0.436) ('refugee crisis', 0.436)
s4: But to do that, you need facts and data – two things that the British migration debate has lacked this summer.	1.176	0.0	('british migration debate', 1.889) ('summer', 0.363)
s5: Theresa May got the ball rolling in May, when she claimed on Radio 4 that the vast majority of migrants to Europe are Africans traveling for economic reasons.	0.746	0.0	('theresa may', 2.716) ('vast majority', 0.453) ('migrants to europe', 0.32) ('radio', 0.32) ('ball', 0.32) ('economic reason', 0.207)
Article summary for any repeated CASTER identified contextual aspect, topic, or entity			'theresa may' total CASTER relevancy: 3.238 'theresa may' average BSI bias: 0.871 'theresa may' average VADER: –0.25

CASTER) to analyze coverage bias at the sentence level, and at the article level using aggregated weighted averages.

For example, *s1* indicates that "Theresa May" has a mediocre relevancy connection (CASTER = 0.522) to a slightly biased sentence (BSI = 0.996) with a negative sentiment score (VADER = –0.4939). On the other hand, *s5* indicates a much stronger relevancy connection (CASTER = 2.716) to a less biased sentence (BSI = 0.746) with a neutral sentiment score (VADER = 0.0). At the article level, it may be interesting to inspect the relative coverage bias of aspects, topics, or entities in the article. For example, when comparing Theresa May to Philip Hammond at the article level, Theresa receives more attention as quantified by a greater total CASTER relevancy (3.238) than Philip (1.566), is associated with sentences that have lower bias (average BSI: 0.871 versus 0.996), and less negative sentiment (average VADER: –0.25 versus –0.4939).

Additionally, this concept can be extended to investigate coverage and gatekeeping bias at the article corpus units of analysis, as analysts using BSI can assess the degree to which CASTER aspects are (or are not) present in the corpus. This may be of use to media analysts, computational journalism researchers, and media studies researchers interested in comparing the degree of bias in news stories over time or across news categories, topics, authors, news organizations/media sources, newspaper corpora, or geographic boundaries.

5.9 Chapter Summary

This chapter first consolidates much of the literature motivating the study of media bias, describing the hostile media phenomena and describing how bias can be manifested as statement bias, coverage bias, or gatekeeping bias. I then explore underlying

psychological theory and related constructs from the social sciences which describe biases in various types of and forms (e.g., those arising from framing effects and epistemological influences). These theories provide a foundation upon which I then develop and operationalize 32 lexical and sentential measures of statement bias in the sentences of news stories, which I hierarchically organize into 13 features of a computational model called the Biased Statement Investigator (BSI). Extensive piloting and then expanded evaluations showed that the performance of the BSI model and selected features compared quite well to human performance for matching the average perceived bias rating for sentences in real world news stories (mean Pearson Correlation Coefficient r=0.565 for BSI using Regularized Random Forest machine learning, compared to r=0.661 for human judges). I compared the BSI model to a sentiment only model using the VADER sentiment analysis model I described in Chapter 3. Finally I demonstrate several applications for BSI for analyzing statement bias and coverage bias at both sentence and article scales of analysis, and posit extensions to analyses of coverage and gatekeeping bias at the article corpus unit of analysis. I argue that these capabilities have both research and practical value.

CHAPTER 6. CONCLUSIONS

This dissertation presents the confluence of social science theory building and application with human-centered development, evaluation, and deployment of computational tools to support the systematic and (unobtrusive) study of human behavior and social phenomena as observed via digital communications that occur at the individual scale, dyadic and personal social network scales, and mass broadcast communication scale. I believe this work makes substantial theoretical, methodological, and technical contributions to the fields of Human-Centered Computing and Computational Social Science. This dissertation elucidates and demonstrates a conceptual framework and process model of Human-Centered Computational Social Science:

(1) ask an interesting question about human behavior or social phenomena that may be challenging or daunting to answer using traditional methods (e.g., due to issues of scale), but could be made easier, faster, more consistent, and perhaps less prone to experimental concerns related to relying solely on self-reports [54], social desirability [170], researcher-induced expectancy bias [190], or the Hawthorne Effect [1], then…

(2) use well-established theory and extant research literature to conceive of and then iteratively develop (leveraging human-centered design methods) a tool which might be useful for helping to answer the question(s) from #1, then…

(3) evaluate the tool (leveraging HCC evaluation methods), then…

(4) use the tool to empirically analyze and learn something about human behavior or social phenomena, then…

(5) repeat steps 1-4 after gaining insights that raise new questions.

This approach demonstrates the value of Computational Social Science together with Human-Centered Computing. I posit that human-centered and social data analytics in the near future will need to embrace more multifaceted representations of human behavior with more complex models. Such models will need to integrate data of disparate forms, using disparate units of measure, collected from disparate sources, at disparate scales. The next generation of computational social scientists will also face issues related to developing methods and tools to help facilitate the collection, processing, analyzing, and visualizing of such multifaceted social data.

This dissertation both illustrates and addresses some of these challenges by advocating and demonstrating *Large-N* and *Multiple-T* psychology-style studies using technology-mediated communications. In order to achieve useful statistical power while incorporating the expanded scope resulting from increased representational complexity, and at the same time preserving broad generalization and application capacities, I argue that social science analysts will need to design and conduct similar studies with larger sample sizes (i.e., "Large-N" studies) collected over multiple instances in time (i.e., "Multiple-T", or longitudinal studies). For example, Chapters 2, and 4 present large scale empirical studies, while Chapters 2 and 5 develop theory-inspired computational approaches that allow me to compare theories, methods, and relative impacts of social science constructs as explanations for empirical observations. These studies are facilitated by the methods and technological tools reported in Chapters 3 and 4 for scaling up the analyses to a level of complexity beyond what was typical for much of social science research of the recent past (or even some today). I expect such study designs to become increasingly more prevalent, and will eventually be the norm for social analytics.

This leads to another conclusion about what this dissertation suggests about the process of studying human behavior at scale: methods that blend qualitative and quantitative techniques are very compelling. Data-driven (bottom-up) and theory-informed (top-down) approaches both have benefits, and each helps mitigate the shortfalls of the other to help address not just the *what*, but also the *how* and the *why*. Developing computational tools to facilitate both qualitative and quantitative research is advantageous for helping to address one's own research questions, and can also be useful for other researchers. For example, the VADER sentiment analysis tool is publically available as an open source Python package, and it is integrated into NLTK. Clearly, VADER is useful, as evidenced by more than 400 citations by sociologists, psychologists, journalists and communications researchers, economists, political scientists, marketing/consumer researchers, business analysts and data scientists, and, not to forget, a host of computer scientists.

BSI might not ever become as popular as VADER (largely because it is a more specialized tool), but I do think it will be quite useful to media studies and communications researchers who are interested in detecting and quantifying bias. I also hope it will be useful to practitioners such as news journalists/writers, editors, readers/consumers, and fact-checkers/watch-dog organizations. Already, an organization called Global Voices[38] is using BSI. Global Voices is an international consortium of more than 1,400 journalists, reporters, writers, editors, analysts, media experts, researchers, and translators representing main stream as well as independent and social media press in 167 countries. Global Voices

[38] https://globalvoices.org/

curates, verifies, translates, studies, and reports on trending news and stories that otherwise may be under-reported by mega-mass media news organizations. They discovered my preliminary model of BSI as an open source project in my GitHub repository, and have been using a prototype version of BSI to detect and quantify the degree of bias in news stories. Being journalists, they wrote an investigative report that leverage BSI to illustrate the ways that news stories have framed Brexit and Immigration in the United Kingdom[39].

[39] https://newsframes.globalvoices.org/2018/03/30/brexit-and-bias-the-framing-of-immigrants-in-the-media/

REFERENCES

1. John G. Adair. 1984. The Hawthorne Effect: A Reconsideration of the Methodological Artifact. *Journal of Applied Psychology* 69, 2: 334–345.

2. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, 30–38.

3. Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, 190–199.

4. Americans for Democratic Action. 2016. ADA Voting Records. *Americans for Democratic Action*. Retrieved January 13, 2018 from https://adaction.org/ada-voting-records/

5. Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 1–9.

6. Jonathan Anderson. 1983. Lix and Rix: Variations on a Little-Known Readability Index. *Journal of Reading* 26, 6: 490–496.

7. Paul Andre, Aniket Kittur, and Steven P. Dow. 2014. Crowd synthesis: extracting categories and clusters from complex data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 989–998.

8. D. Scott Appling, Erica Briscoe, and C.J. Hutto. 2015. Discriminative Models for Predicting Deception Strategies. In *Proceedings of the WWW 2015 Companion for Rumors and Deception in Social Media (RDSM)*, 947–952.

9. Cecilia Aragon, Clayton Hutto, Andy Echenique, Brittany Fiore-Gartland, Yun Huang, Jinyoung Kim, Gina Neff, Wanli Xing, and Joseph Bayer. 2016. Developing a Research Agenda for Human-Centered Data Science. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, 529–535. https://doi.org/10.1145/2818052.2855518

10. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proc. of LREC*.

11. Eytan Bakshy, Jake M. Hofman, Winter Mason, and Duncan Watts. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the Fourth International Conference on Web Search and Data Mining*.

12. A.L. Barabási and Reka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286: 509–512.

13. Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad is stronger than good. *Review of General Psychology* 5, 4: 323–370. https://doi.org/10.1037/1089-2680.5.4.323

14. A. J. Berinsky, G. A. Huber, and G. S. Lenz. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20, 3: 351–368. https://doi.org/10.1093/pan/mpr057

15. Wiebe E Bijker. 1995. *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change*. MIT Press, Cambridge, MA.

16. Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly, Beijing ; Cambridge [Mass.].

17. David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

18. P. Bourdieu. 1985. The forms of capital. In *Handbook of Theory and Research for the Sociology of Education*, J.C. Richardson (ed.). Greenwood, New York, 241–258.

19. danah boyd, Scott A. Golder, and Gilad Lotan. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *43rd Annual Hawaii International Conference on System Sciences*.

20. Margaret M. Bradley and Peter J. Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. NIMH Center for the Study of Emotion and Attention, Center for Research in Psychophysiology, University of Florida.

21. Moira Burke, Robert Kraut, and Cameron Marlow. 2011. Social Capital on Facebook: Differentiating Uses and Users. In *ACM CHI 2011*.

22. Kenneth P. Burnham and David R. Anderson. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* 33, 2: 261–304. https://doi.org/10.1177/0049124104268644

23. Erik Cambria, Catherine Havasi, and Amir Hussain. 2012. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, 202–207.

24. Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. SenticNet: A Publicly Available Semantic Resource for Opinion Mining. In *Proc. of the AAAI Symposium on Commonsense Knowledge*, 14–18.

25. A.C. Cameron and P.K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge University Press, New York.

26. Joseph N. Cappella and Kathleen Hall Jamieson. 1997. *Spiral of Cynicism: The Press and the Public Good*. Oxford University Press, USA, New York, NY.

27. Earl R. Carlson. 1966. The Affective Tone of Psychology. *The Journal of General Psychology* 75, 1: 65–78. https://doi.org/10.1080/00221309.1966.9710350

28. D. Cartwright and F. Harary. 1956. Structural balance: A generalization of Heider's theory. *Psychological Review* 63, 5: 277–293.

29. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *In Proc. CHI*, 675–684, 2011. Retrieved September 13, 2014 from http://dl.acm.org/citation.cfm?id=1963500

30. Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *In Proc. NIPS*, 288–296, 2009.

31. R. B. Cialdini. 2007. *Influence: the psychology of persuasion*. Collins. Retrieved from http://books.google.com/books?id=5dfv0HJ1TEoC

32. Claudio Cioffi-Revilla. 2014. *Introduction to Computational Social Science: Principles and Applications*. Springer London, London. https://doi.org/10.1007/978-1-4471-5661-1

33. Claudio Cioffi-Revilla. 2017. *Introduction to Computational Social Science: Principles and Applications*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-50131-4

34. CLiPS Research Center. Pattern.en Software. Retrieved July 28, 2015 from http://www.clips.ua.ac.be/pages/pattern-en

35. Alexander Conrad, Janyce Wiebe, and and Rebecca Hwa. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, 80–88.

36. R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J. -P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel, and D. Helbing. 2012. Manifesto of computational social science. *The European Physical Journal Special Topics* 214, 1: 325–346. https://doi.org/10.1140/epjst/e2012-01697-8

37. Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8, 3: e57410. https://doi.org/10.1371/journal.pone.0057410

38. Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 886–893. Retrieved September 1, 2014 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1467360

39. Arjen van Dalen. 2012. Structural Bias in Cross-National Perspective: How Political Systems and Journalism Cultures Influence Government Dominance in the News. *The International Journal of Press/Politics* 17, 1: 32–55. https://doi.org/10.1177/1940161211411087

40. Dave D'Alessio and Mike Allen. 2000. Media Bias in Presidential Elections: A Meta-Analysis. *Journal of Communication* 50, 4: 133–156. https://doi.org/10.1111/j.1460-2466.2000.tb02866.x

41. Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. 2015. Media Bias in German Online Newspapers. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 133–137.

42. Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Proceeding of the 23rd international conference on Computational Linguistics*.

43. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

44. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

45. Tom De Smedt and Walter Daelemans. 2012. Pattern for Python. *Journal of Machine Learning Research* 13: 2063–2067.

46. Dictionary.com. 2018. Dictionary.com Unabridged.

47. Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 231–240.

48. Domo. 2018. *Data Never Sleeps 6.0*. Retrieved from www.domo.com

49. Judith S. Donath. forthcoming. *Signals, Truth, and Design*. MIT Press, Cambridge, MA.

50. Judith S. Donath. 2007. Social Signals in Supernets. *Journal of Computer-Mediated Communication* 13, 1: article 12.

51. Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2399–2402.

52. Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. Are your participants gaming the system?: screening mechanical turk workers. In *In Proc. CHI*, 2399–2402, 2010. Retrieved September 14, 2014 from http://dl.acm.org/citation.cfm?id=1753688

53. James N. Druckman. 2001. The Implications of Framing Effects for Citizen Competence. *Political Behavior* 23, 3: 225–256.

54. Nathan Eagle, Alex Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106, 36: 15274–15278.

55. David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, Cambridge, MA.

56. Jakob-Moritz Eberl, Hajo G. Boomgaarden, and Markus Wagner. 2017. One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences. *Communication Research* 44, 8: 1125–1148. https://doi.org/10.1177/0093650215614364

57. Erick Elejalde, Leo Ferres, and Eelco Herder. 2017. The Nature of Real and Perceived Bias in Chilean Media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 95–104.

58. Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 617–624.

59. Jonathan St. B. T. Evans, Julie L. Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition* 11, 3: 295–306. https://doi.org/10.3758/BF03196976

60. Scott Feld. 1981. The Focused Organization of Social Ties. *The American Journal of Sociology* 86, 5: 1015–1035.

61. Lauren Feldman. 2014. The Hostile Media Effect. In *The Oxford Handbook of Political Communication*, Kate Kenski and Kathleen Hall Jamieson (eds.). Oxford University Press, New York, NY.

62. Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

63. Leon Festinger. 1957. *A theory of cognitive dissonance*. Stanford University Press, Stanford, CA.

64. Casey Fiesler, Michaelanne Dye, Jessica L. Feuston, Chaya Hiruncharoenvate, C. J. Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S. Bruckman, Munmun De Choudhury, and Eric Gilbert. 2017. What (or Who) Is Public?: Privacy Settings and Social Media Content Sharing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 567–580.

65. Dennis J. Folds. 2015. Perception of bias in unattributed news stories. In *Procedures of the Annual Meeting of the Association for Psychological Science*.

66. Dennis J. Folds, C.J. Hutto, and Thomas A. McDermott. 2017. Toward Next Generation Social Analytics: A Platform for Analysis of Quantitative, Qualitative, Geospatial, and Temporal Factors of Community Resilience. *International Journal on Advances in Internet Technology* 10, 12.

67. Clifford Geertz. 1973. Thick Description: Toward an Interpretive Theory of Culture. In *The interpretation of cultures: selected essays*. Basic Books, New York, NY, 3–30.

68. Abraham Genizi. 1993. Decomposition of R2 in multiple regression with correlated regressors. *Statistica Sinica* 3, 2: 407–420.

69. Matthew Gentzkow and Jesse M. Shapiro. 2010. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica* 78, 1: 35–71. https://doi.org/10.3982/ECTA7195

70. Eric Gilbert. 2012. Predicting Tie Strength in a New Medium. In *Proceedings of the 2012 ACM conference on Computer supported cooperative work*.

71. Eric Gilbert. What if we ask a different question?: social inferences create product ratings faster. In *In Proc. CHI*, 2759–2762, 2014.

72. Eric Gilbert and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the 27th international conference on Human factors in computing systems*, 211–220.

73. Sharad Goel, Winter Mason, and Duncan J Watts. 2010. Real and Perceived Attitude Agreement in Social Networks. *Journal of Personality and Social Psychology* 99, 4: 611–621.

74. Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Anchor. Retrieved from http://www.amazon.com/Presentation-Self-Everyday-Life/dp/0385094027

75. Scott A. Golder and Sarita Yardi. 2010. Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality. In *Proceedings of the Second IEEE International Conference on Social Computing*.

76. Stephan Greene and Philip Resnik. 2009. More than words: syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

77. Tim Groeling. 2013. Media Bias by the Numbers: Challenges and Opportunities in the Empirical Study of Partisan News. *Annual Review of Political Science* 16, 1: 129–151. https://doi.org/10.1146/annurev-polisci-040811-115123

78. Ulrike Grömping. 2007. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician* 61, 2: 139–147. https://doi.org/10.1198/000313007X188252

79. Tim Groseclose and Jeffrey Milyo. 2005. A Measure of Media Bias. *The Quarterly Journal of Economics* 120, 4: 1191–1237.

80. Robert A. Hackett. 1984. Decline of a paradigm? Bias and objectivity in news media studies. *Critical Studies in Mass Communication* 1, 3: 229–259.

81. Haewoon Kwak, Sue Moon, and Wonjae Lee. 2012. More of a Receiver Than a Giver: Why Do People Unfollow in Twitter? *International AAAI Conference on Weblogs and Social Media; Sixth International AAAI Conference on Weblogs and Social Media*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4598/5042

82. Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson (eds.). 2010. *Multivariate data analysis*. Prentice Hall, Upper Saddle River, NJ.

83. Felix Hamborg, Norman Meuschke, and Bela Gipp. 2017. Matrix-Based News Aggregation: Exploring Different News Perspectives. In *Proceeding of the ACM/IEEE Joint Conference on Digital Libraries*, 1–10. https://doi.org/10.1109/JCDL.2017.7991561

84. Jeffrey T. Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 929–932.

85. Eszter Hargittai and Eden Litt. 2011. The Tweet Smell of Celebrity Success: Explaining Variation in Twitter Adoption among a Diverse Group of Young Adults. *New Media & Society* 13, 5: 824–842.

86. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52: 139–183.

87. Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *In Proc. CHI*, 203–212, 2010.

88. Fritz Heider. 1946. Attitudes and cognitive organization. *The Journal of Psychology* 21, 1: 107–112.

89. Fritz Heider. 1958. *The psychology of interpersonal relations*. Wiley, New York, NY.

90. C. Richard Hofstetter and Terry F. Buss. 1978. Bias in television news coverage of political events: A methodological analysis. *Journal of Broadcasting* 22, 4: 517–530. https://doi.org/10.1080/08838157809363907

91. Alexander Hogenboom, Paul van Iterson, Bas Heerschop, Flavius Frasincar, and Uzay Kaymak. 2011. Determining negation scope and strength in sentiment analysis. In *Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics*, 2589–2594. https://doi.org/10.1109/ICSMC.2011.6084066

92. Courtenay Honeycutt and Susan C. Herring. 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. In *Proceedings of the Forty-Second Hawaii International Conference on System Sciences*.

93. Joan B. Hooper. 1975. On assertive predicates. In *Syntax and Semantics*, J. Kimball (ed.). Academic Press, New York, NY, 91–124.

94. Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177.

95. Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, 173–178. https://doi.org/10.1145/1810617.1810647

96. Shih-Wen Huang and Wai-Tat Fu. 2013. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 621–630.

97. Christoph Hube. 2017. Bias in Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 717–721.

98. C. J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–255.

99. C. J. Hutto, Sarita Yardi, and Eric Gilbert. 2013. A Longitudinal Study of Follow Predictors on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 821–830.

100. C.J. Hutto. 2016. *VADER Sentiment Analysis Software*. Retrieved July 28, 2015 from https://github.com/cjhutto/vaderSentiment

101. C.J. Hutto, Caroline J. Bell, Sarah K. Farmer, Cara Bailey Fausset, Linda R. Harley, and Walter Bradley Fain. 2015. Social media gerontology: Understanding social media usage among older adults. *Web Intelligence Journal* 13, 1: 69–87.

102. Clayton Hutto and Caroline Bell. 2014. Social Media Gerontology: Understanding Social Media Usage among a Unique and Expanding Community of Users. In *Proceedings of the 2014 47th Hawaii International Conference on System Sciences*, 1755–1764. https://doi.org/10.1109/HICSS.2014.223

103. Clayton J. Hutto. 2016. Blending Quantitative, Qualitative, Geospatial, and Temporal Data: Progressing Towards the Next Generation of Human Social Analytics. In *Proceedings of the Second International Conference on Human and Social Analytics (HUSO)*.

104. Clayton J. Hutto, Dennis J. Folds, and D. Scott Appling. 2015. Computationally Detecting and Quantifying the Degree of Bias in Sentence-Level Text of News Stories. In *Proceedings of the First International Conference on Human and Social Analytics (HUSO)*.

105. Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Continuum, New York, NY.

106. InternetLiveStats.com. 2016. Internet Live Stats. *Internet Live Stats - Internet Usage and Social Media Statistics*. Retrieved September 9, 2016 from http://www.internetlivestats.com/

107. Panagiotis G. Ipeirotis. 2010. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2: 16–21.

108. Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on Amazon Mechanical Turk. In *In Proc. SIGKDD Workshop on Human Computation*, 64–67, 2010.

109. Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis*, 56–65.

110. Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. 2013. Analyzing and predicting viral tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, 657–664. https://doi.org/10.1145/2487788.2488017

111. Edward E. Jones (ed.). 1987. *Attribution: perceiving the causes of behavior*. Erlbaum, Hillsdale, N.J.

112. Edward E Jones and Victor A Harris. 1967. The attribution of attitudes. *Journal of Experimental Social Psychology* 3, 1: 1–24. https://doi.org/10.1016/0022-1031(67)90034-0

113. Edward E Jones and Richard E Nisbett. 1971. *The actor and the observer: Divergent perceptions of the causes of behaviors*. General Learning Press, New York, NY.

114. Jaap Kamps, Robert J. Mokken, Maarten Marx, and Maarten de Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004*, 1115–1118.

115. Lauri Karttunen. 1971. Implicative Verbs. *Language* 47, 2: 340–358.

116. Michael L. Katz and Carl Shapiro. 1985. Network Externalities, Competition, and Compatibility. *The American Economic Review* 75, 3: 424–440.

117. Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *In Proc. CIKM*, 1941–1944, 2011. Retrieved September 14, 2014 from http://dl.acm.org/citation.cfm?id=2063860

118. Chris Kerns. 2014. Understanding Brands on Twitter. In *Trendology*. Palgrave Macmillan US, New York, 39–85. https://doi.org/10.1057/9781137479563_3

119. Peter Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* National Technical Information Service, Springfield, Virginia 22151 (AD-A006 655/5GA, MF $2.25, PC $3.75). Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED108134

120. Paul Kiparsky and Carol Kiparsky. 1970. Fact. In *Progress in Linguistics*, Manfred Bierwisch and Karl Erich Heidolph (eds.). Mouton, The Hague, Netherlands, 143–173.

121. Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 453–456.

122. Funda Kivran-Swaine, Priya Govindan, and Mor Naaman. 2011. The impact of Network Structure on Breaking Ties in Online Social Networks: Unfollowing on Twitter. In *CHI 2011*.

123. Funda Kivran-Swaine and Mor Naaman. 2011. Network Properties and Social Sharing of Emotions in Social Awareness Streams. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*.

124. Alistair Knott. 1996. A Data-Driven Methodology for Motivating a Set of Coherence Relations. Department of Artificial Intelligence, University of Edinburgh.

125. Peter Kollock. 1999. The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace. In *Communities in Cyberspace*, Marc Smith and Peter Kollock (eds.). Routledge, New York, 220–239.

126. Adam Kramer. 2010. An unobtrusive behavioral model of "gross national happiness." In *Proceedings of the 28th international conference on Human factors in computing systems*.

127. Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. Integrating on-demand fact-checking with public dialogue. In *In Proc. CSCW*, 1188–1199, 2014. Retrieved September 14, 2014 from http://dl.acm.org/citation.cfm?id=2531677

128. Max Kuhn. 2008. Building Predictive Models in R Using the Caret Package. *Journal of Statistical Software* 28, 5. https://doi.org/10.18637/jss.v028.i05

129. Miron B. Kursa and Witold R. Rudnicki. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software* 36, 11.

130. Haewoon Kwak, Hyunwoo Chun, and Sue Moon. 2011. Fragile Online Relationship: A First Look at Unfollow Dynamics in Twitter. In *Proceedings of the 2011 annual conference on Human factors in computing systems*.

131. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 591–600.

132. Cliff A.C. Lampe, Nicole Ellison, and Charles Steinfield. 2007. A familiar face(book): profile elements as signals in an online social network. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 435–444.

133. Walter S. Lasecki, Mitchell Gordon, Danai Koutra, Malte F. Jung, Steven P. Dow, and Jeffrey P. Bigham. 2014. Glance: rapidly coding behavioral video with the crowd. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 551–562.

134. Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line Trend Analysis with Topic Models:$\backslash$# twitter Trends Detection Topic Model Online. In *COLING*, 1519–1534. Retrieved December 22, 2013 from https://www.aclweb.org/anthology/C/C12/C12-1093.pdf

135. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational Social Science. *Science* 323, 5915: 721–723.

136. David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, 556–559.

137. Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 109–116.

138. Yu-Ru Lin, James P. Bagrow, and David Lazer. 2011. More Voices Than Ever? Quantifying Media Bias in Networks. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*.

139. Yu-Ru Lin, James P. Bagrow, and David Lazer. 2012. Quantifying Bias in Social and Mainstream Media. *SIGWEB Newsletter*, Summer: 1–6.

140. Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. Exploring iterative and parallel human computation processes. In *In Proc. CHI*, 68–76, 2010. Retrieved September 1, 2014 from http://dl.acm.org/citation.cfm?id=1837907

141. Bing Liu. 2010. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing* (2nd ed.), Nitin Indurkhya and Fred Damerau (eds.). Chapman & Hall/CRC Press, Boca Raton, FL.

142. Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, San Rafael, CA.

143. Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th international conference on World Wide Web*, 342–351.

144. Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. 2011. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, 347–356.

145. Paul L. MacDonald and Robert C. Gardner. 2000. Type I error rate comparisons of post hoc procedures for I j Chi-Square tables. *Educational and Psychological Measurement* 60, 5: 735–754.

146. Rashid Mahmood, Misbah Obaid, and Aleem Shakir. 2014. A Critical Discourse Analysis of Figurative Language in Pakistani English Newspapers. *International Journal of Linguistics* 6, 3: 210. https://doi.org/10.5296/ijl.v6i3.5412

147. Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1: 1–23.

148. Winter Mason and Duncan J. Watts. 2010. Financial incentives and the performance of crowds. *ACM SIGKDD Explorations Newsletter* 11, 2: 100–108.

149. Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 1: 415–438.

150. Tanushree Mitra, C.J. Hutto, and Eric Gilbert. 2015. Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1345–1354. https://doi.org/10.1145/2702123.2702553

151. Nicola J. Morley, Jonathan St. B. T. Evans, and Simon J. Handley. 2004. Belief Bias and Figural Bias in Syllogistic Reasoning. *The Quarterly Journal of Experimental Psychology Section A* 57, 4: 666–692. https://doi.org/10.1080/02724980343000440

152. Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *In Proc. CSCW*, 441–450, 2012. Retrieved September 13, 2014 from http://dl.acm.org/citation.cfm?id=2145274

153. Sean Munson and Paul Resnick. 2013. Encouraging Reading of Diverse Political Viewpoints with a Browser Widget. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 419–428.

154. Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it Really About Me? Message Content in Social Awareness Streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 189–192.

155. Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *In Proc. CSCW*, 189–192, 2010. Retrieved September 14, 2014 from http://dl.acm.org/citation.cfm?id=1718953

156. Vishwajeet Narwal, Mohamed Hashim Salih, Jose Angel Lopez, Angel Ortega, John O'Donovan, Tobias Höllerer, and Saiph Savage. 2017. Automated Assistants to Identify and Prompt Action on Visual News Bias. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2796–2801.

157. National Science Foundation. 2012. Human-Centered Computing (HCC) Program Description. Retrieved June 17, 2017 from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503302&org=IIS

158. Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns. In *Proceedings of the 24th International Conference on World Wide Web*, 798–808.

159. Finn Arup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 "Making Sense of Microposts": Big things come in small packages*.

160. Donald Norman. 1990. *The Design of Everyday Things*. Doubleday Business. Retrieved from http://www.amazon.com/Design-Everyday-Things-Donald-Norman/dp/0385267746

161. Sudhaker M. Pandit and Shien-Ming Wu. 2001. *Time Series and System Analysis With Applications*. Krieger Publishing Company, Malabar, FL.

162. Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 271.

163. Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1–2: 1–135.

164. Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5: 411–419.

165. Constantine Papageorgiou and Tomaso Poggio. 2000. A trainable system for object detection. *International Journal of Computer Vision* 38, 1: 15–33.

166. Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 443–452.

167. Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2012. A Computational Framework for Media Bias Mitigation. *ACM Trans. Interact. Intell. Syst.* 2, 2: 1–32.

168. Souneil Park, Seungwoo Kang, Sangjeong Lee, Sangyoung Chung, and Junehwa Song. 2008. Mitigating media bias: a computational approach. In *Proceedings of the hypertext 2008 workshop on Collaboration and collective intelligence*, 47–51.

169. Souneil Park, KyungSoon Lee, and Junehwa Song. 2011. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 340–349.

170. Delroy Paulhus. 2002. Socially Desirable Responding: The Evolution of a Construct. In *The role of constructs in psychological and educational measurement*, H.I. Braun, D.N. Jackson and D.E. Wiley (eds.). Erlbaum, Mahwah, NJ, 49–69.

171. Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2013. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*: 1–9. https://doi.org/10.3758/s13428-013-0434-y

172. Guido Peeters. 1971. The positive-negative asymmetry: On cognitive consistency and positivity bias. *European Journal of Social Psychology* 1, 4: 455–474. https://doi.org/10.1002/ejsp.2420010405

173. Pennebaker Conglomerates, Inc. LIWC Text Analysis Software. Retrieved July 28, 2015 from http://www.liwc.net/

174. James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. *The development and psychometric properties of LIWC2007*. LIWC.net, Austin, TX.

175. J.W. Pennebaker, M Francis, and R Booth. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Erlbaum Publishers, Mahwah, NJ.

176. Richard M. Perloff. 2015. A Three-Decade Retrospective on the Hostile Media Effect. *Mass Communication and Society* 18, 6: 701–729. https://doi.org/10.1080/15205436.2015.1051234

177. Thomas F. Pettigrew. 1979. The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice. *Personality and Social Psychology Bulletin* 5, 4: 461–476. https://doi.org/10.1177/014616727900500407

178. Pew Research Center. 2012. *Cable Leads the Pack as Campaign News Source*. Retrieved October 10, 2015 from http://assets.pewresearch.org/wp-content/uploads/sites/5/legacy-pdf/2012%20Communicating%20Release.pdf

179. Pew Research Center. 2012. *Further Decline in Credibility Ratings for Most News Organizations*. Retrieved October 12, 2015 from http://assets.pewresearch.org/wp-content/uploads/sites/5/2012/08/8-16-2012-Media-Believability1.pdf

180. Pew Research Center. 2014. *Political Polarization and Media Habits*. Retrieved January 8, 2018 from http://assets.pewresearch.org/wp-content/uploads/sites/13/2017/05/09144304/PJ_2017.05.10_Media-Attitudes_FINAL.pdf

181. Pew Research Center. 2017. *Americans' Attitudes About the News Media Deeply Divided Along Partisan Lines*. Retrieved January 8, 2018 from http://assets.pewresearch.org/wp-content/uploads/sites/13/2017/05/09144304/PJ_2017.05.10_Media-Attitudes_FINAL.pdf

182. Dražen Prelec. 2004. A Bayesian truth serum for subjective data. *Science* 306, 5695: 462–466.

183. Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *In Proc. EMNLP*, 1589–1599, 2011. Retrieved September 13, 2014 from http://dl.acm.org/citation.cfm?id=2145602

184. D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. 2011. In the Mood for Being Influential on Twitter. In *IEEE 3rd International Conference on Social Computing*, 307–314. https://doi.org/10.1109/PASSAT/SocialCom.2011.27

185. Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In *In Proc. NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 139–147, 2010.

186. Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Meeting of the Association for Computational Linguistics*, 1650–1659.

187. Paul Resnick, Joseph Konstan, Yan Chen, and Robert Kraut. 2012. Starting New Online Communities. In *Evidence-based social design: Mining the social sciences to build online communities*. MIT Press, Cambridge, MA.

188. Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 105–112.

189. Richard M. Roberts and Roger J. Kreuz. 1994. Why Do People Use Figurative Language? *Psychological Science* 5, 3: 159–163.

190. Robert Rosenthal. 2004. Experimenter Expectancy Effect. In *The Sage encyclopedia of social science research methods*, Michael S. Lewis-Beck, Alan Bryman and Tim Futing Liao (eds.). Sage, Thousand Oaks, Calif.

191. Robert Rosenthal and Donald B. Rubin. 1984. Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology* 76, 6: 1028.

192. Paul Rozin and Edward B. Royzman. 2001. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review* 5, 4: 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

193. Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 1679–1684.

194. J. Saldana. 2009. *The Coding Manual for Qualitative Researchers*. SAGE Publications. Retrieved from http://books.google.com/books?id=msuOE0UfXpUC

195. Kathleen M Schmitt, Albert C Gunther, and Janice L Liebhart. 2004. Why Partisans See Mass Media as Biased. *Communication Research* 31, 6: 623–641.

196. Lokendra Shastri, Anju G Parvathy, Abhishek Kumar, John Wesley, and Rajesh Balakrishnan. 2010. Sentiment Extraction: Integrating Statistical Parsing, Semantic Analysis, and Common Sense Reasoning. In *Proceedings of the Twenty-Second Innovative Applications of Artificial Intelligence Conference*.

197. Aaron D. Shaw, John J. Horton, and Daniel L. Chen. Designing incentives for inexpert human raters. In *In Proc. CHI*, 275–284, 2011. Retrieved September 14, 2014 from http://dl.acm.org/citation.cfm?id=1958865

198. Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 614–622.

199. Aashish Sheshadri and Matthew Lease. 2013. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*, 156–164.

200. R Snow, B O'Connor, D Jurafsky, and A.Y. Ng. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *EMNLP*.

201. Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng, and Chris Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of Empirical Methods in Natural Language Processing*.

202. Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 116–124.

203. Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. Modeling Factuality Judgments in Social Media Text. In *In Proc. ACL*, 415–420, 2014.

204. A Sorokin and D. Forsyth. 2008. Utility data annotation with Amazon Mechanical Turk. *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*: 1–8. https://doi.org/10.1109/CVPRW.2008.4562953

205. Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

206. Anselm L. Strauss and Juliet Corbin. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Thousand Oaks, CA.

207. Yu-An Sun, Shourya Roy, and Greg Little. 2011. Beyond Independent Agreement: A Tournament Selection Approach for Quality Assurance of Human Computation Tasks. In *Human Computation*. Retrieved September 1, 2014 from http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewFile/3904/4257

208. James Surowiecki. 2004. *The Wisdom of Crowds*. Anchor Books, New York, NY.

209. Lisa Collins Tidwell and Joseph B. Walther. 2002. Computer-Mediated Communication Effects on Disclosure, Impressions, and Interpersonal Evaluations: Getting to Know One Another a Bit at a Time. *Human Communication Research* 28, 3: 317–348.

210. Andranik Tumasjan, Timm O. Sprenger, Phillipp G. Sandner, and Isabell M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

211. Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.* 21, 4: 315–346.

212. Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. 211, 4481: 453–458. https://doi.org/10.1126/science.7455683

213. R. P. Vallone, L. Ross, and M. R. Lepper. 1985. The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology* 49, 3: 577–585.

214. Hao Wang, Dogan Can, Abe Kazemzadeh, Francois Bar, and Shrikanth Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, 115–120.

215. Yi-Chia Wang and Robert Kraut. 2012. Twitter and the development of an audience: those who stay on topic thrive! In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 1515–1518.

216. Yi-Chia Wang, Robert Kraut, and John M. Levine. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *In Proc. CSCW*, 833–842, 2012.

217. B Wellman and S Wortley. 1990. Different strokes from different folks: Community ties and social support. *American Journal of Sociology* 96: 558–588.

218. Wikipedia contributors. 2017. English-language idioms. *Wikipedia, The Free Encyclopedia*. Retrieved November 3, 2018 from https://en.wikipedia.org/wiki/English-language_idioms

219. Wikipedia contributors. 2017. List of English-language metaphors. *Wikipedia, The Free Encyclopedia*. Retrieved November 3, 2018 from https://en.wikipedia.org/wiki/List_of_English-language_metaphors

220. Wikipedia contributors. 2017. List of political metaphors. *Wikipedia, The Free Encyclopedia*. Retrieved November 3, 2018 from https://en.wikipedia.org/wiki/List_of_political_metaphors

221. Wikipedia contributors. 2017. Wikipedia Manual of Style/Words to watch: Puffery. *Wikipedia, The Free Encyclopedia*. Retrieved November 3, 2018 from https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch#Puffery

222. Wikipedia contributors. 2017. Wikipedia Manual of Style/Words to watch: Contentious labels. *Wikipedia, The Free Encyclopedia*. Retrieved November 3, 2018 from https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch#Contentious_labels

223. Wikipedia contributors. 2017. Wikipedia Manual of Style/Words to watch: Expressions of doubt. *Wikipedia, The Free Encyclopedia*. Retrieved November 3, 2018 from https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch#Expressions_of_doubt

224. Wikipedia contributors. 2017. Wikipedia Manual of Style/Words to watch: Weasel. *Wikipedia, The Free Encyclopedia*. Retrieved November 3, 2018 from https://en.wikipedia.org/w/index.php?title=Wikipedia:WEASEL

225. Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for crowdsourcing social data analysis. In *In Proc. CHI*, 227–236. Retrieved September 14, 2014 from http://dl.acm.org/citation.cfm?id=2207709

226. Cj Willmott and K Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30: 79–82. https://doi.org/10.3354/cr030079

227. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–354. https://doi.org/10.3115/1220575.1220619

228. Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the 19th national conference on Artificial intelligence*, 761–767.

229. Robin Wooffitt. 2005. *Conversation analysis and discourse analysis: a comparative and critical introduction*. SAGE, Thousand Oaks, CA.

230. Tae Yano, Philip Resnik, and Noah A. Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 152–158.

231. Jun Zhang, Yan Qu, Jane Cody, and Yulingling Wu. 2010. A case study of micro-blogging in the enterprise: use, value, and related issues. In *Proceedings of the 28th international conference on Human factors in computing systems*, 123–132.

232. Dejin Zhao and Mary Beth Rosson. 2009. How and why people Twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, 243–252.

233. Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*. Springer, 338–349. Retrieved September 1, 2014 from http://link.springer.com/chapter/10.1007/978-3-642-20161-5_34

234. Verena Zuber and Korbinian Strimmer. 2011. High-Dimensional Regression and Variable Selection Using CAR Scores. *Statistical Applications in Genetics and Molecular Biology* 10, 1. https://doi.org/10.2202/1544-6115.1730