

UNDERSTANDING SOCIAL MEDIA CREDIBILITY

A Thesis
Presented to
The Academic Faculty

by

Tanushree Mitra

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

Georgia Institute of Technology
August 2017

Copyright © 2017 by Tanushree Mitra

UNDERSTANDING SOCIAL MEDIA CREDIBILITY

Approved by:

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

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

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

Dr. Scott Counts
Nexus Group
Microsoft Research

Dr. James W. Pennebaker
Department of Psychology
University of Texas at Austin

Date Approved: 5 May 2017

To Mom & Dad

ACKNOWLEDGEMENTS

This has been a very long journey that would not have been possible without the continuous support and encouragement of numerous people along the way. Indeed, it takes a village and then some to raise a doctoral graduate.

First and foremost, I would like to thank my advisor Eric Gilbert. It has been an honor to be your first PhD student. Thank you for taking a chance on me, for believing in the work I wanted to pursue, for your brilliant insights that has helped shaped my research direction, for the constant encouragement at times when I felt lost, for trusting me during my times of crisis, especially in my first year of PhD, and for always steering me to think big and take risks. You exemplify everything I want to be as an academic.

I am very grateful to all my committee members, Amy Bruckman, Jacob Eisenstein, Scott Counts and James W. Pennebaker, for taking the time out from their busy schedules to read my dissertation and provide me suggestions to make the work stronger. Amy was the first professor I met when I visited Georgia Tech as a graduate student. Thank you for believing in me when I first joined the program. You have provided me the research mantra that will forever guide my future endeavors—“Think how this study will make the world a better place.” Jacob has been an invaluable sounding board and mentor for my thesis work. Thank you for taking the time to brainstorm with me on some deeply technical topics and for pushing me forward. Scott and Jamie have showed me the excitement in psychology research and have helped me embrace the joy of interdisciplinary work. I would also like to thank my early mentors from Texas A&M University, including Dylan Shell and James Caverlee, who taught me the basics of conducting research.

Along the way I have made several friends and research buddies. Thank you to all the members of the *comp.social* lab for your constant help, support and for making the life at

work exciting and fun. To Catherine Grevet, Chaya Hiruncharoenvate, Saeideh Bakhshi, C.J. Hutto, and Eshwar Chandrasekharan, thank you for being such awesome research buddies. I hope that our paths will cross again. Amish Goyal, Graham Wright, and Sanjana Shankar were the student researchers who contributed to the creation and analysis of CREDBANK. I want to thank them for their dedication and hard work. My friends and roommates have been essential for keeping my life at decent levels of sanity. Udit Brahmachari, Reema Kundu, and Poorna Roy, thank you for sharing the hardships of your respective PhD journeys, and for the several outings, movie nights, and cooking sessions we have had over the years.

My family has been the strongest pillar for me. To my parents Seema and Swapan Mitra, for giving me the strength, for believing in me, for always pushing me to aim higher, and for allowing me to realize my own potential. Thank you for always staying strong for us and for showing us the meaning of a loving family. To my little sister, Anuradha, thank you for being my friend, my patient listener, and for showing me the true meaning of strength. I love you to the moon and back, and I sincerely believe that you will get back to your feet once again and win the world with your love, kindness and strength. To my grandma, Geeta Rani Mitra, and grandpa, late Amulya Chandra Mitra. You have taught me the essence of hard work. As Bangladeshi immigrants in India, both of you had to work really hard to provide us with the opportunity to lead a better life. Thank you for all your efforts so that we could live comfortably and for always inspiring us to strive for the best in anything we chose to do. To my new family, Swagata, Rajkumar, and Shatabdi Roy Chowdhury, thank you for your support and love in this process. Finally, and most importantly, to my partner in life, Shauvik Roy Choudhary. Thank you for your love and support, both intellectual and emotional, for having the absolute confidence in me, for putting up with me when I was stressed, for your always-positive attitude during times of crisis, and for taking care of everything when I couldn't. Thank you for finishing this journey with me.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	x
LIST OF FIGURES	xii
SUMMARY	xv
I INTRODUCTION	1
1.1 Defining Credibility	2
1.2 Existing approaches to studying social media credibility	4
1.3 Contributions of this thesis	5
1.4 Thesis Overview	6
II RELATED WORK	8
2.1 Perspectives on Credibility Perceptions	8
2.2 Social Epistemology	9
2.3 Social Media and Credibility	12
2.4 Event Factuality	13
III BUILDING A LARGE-SCALE CREDIBILITY CORPUS	15
3.1 CREDBANK's Construction	16
3.1.1 Streaming Data and Preprocessing	17
3.1.2 Event Candidates using Online LDA	20
3.1.3 Filtering Event-specific Topics	22
3.1.4 Credibility Assessment of Topical Events	25
3.1.5 Collection of Topical Event Streams	28
3.2 CREDBANK Overview	28
3.2.1 Agreement among Raters	29
3.2.2 Annotation Response distribution	31
3.2.3 Majority Agreement	31

3.2.4	No Majority Agreement	32
3.3	Future Research Implications	36
IV	STRATEGIES FOR OBTAINING QUALITY ANNOTATIONS FROM CROWD- WORKERS	39
4.1	Background and Related Work	40
4.1.1	Data Coding and Annotation	40
4.1.2	Crowdsourcing Qualitative Coding & Content Analysis	41
4.1.3	Crowdsourcing Data Annotations for Machine Learning	42
4.2	Strategies for Eliciting Quality Data	43
4.3	Our Tasks	46
4.4	Conduct of the Experiments	49
4.4.1	Comparative measures of correctness	50
4.4.2	Statistical Analysis	50
4.5	Experiments	51
4.5.1	Experiment 1 (Strategies 1-3, Tasks 1-3)	51
4.5.2	Experiment 2 (Strategies S3–S4 in Task 4)	54
4.5.3	Results from Experiment 1	54
4.5.4	Results from Experiment 2	56
4.6	Discussion and Conclusions	56
4.6.1	Crowd generated data annotations by non-experts can be reliable and of high-quality	58
4.6.2	Person-oriented strategies trump process-oriented strategies for encouraging high-quality data coding	59
4.6.3	Why do more to get less?	60
V	LINGUISTIC CONSTRUCTS OF CREDIBILITY	61
5.1	Method	63
5.1.1	Validating credibility classification	63
5.1.2	Response Variable: Dependent Measure	66
5.1.3	Predictive Variables: Linguistic Measures	66

5.1.4	Predictive Variables: Controls	71
5.1.5	Statistical Technique	71
5.2	Results	73
5.2.1	Model Fit Comparison	73
5.2.2	Model Accuracy	75
5.3	Accuracy Measurement: Mathematical Construct	77
5.4	Discussion	78
5.4.1	Theoretical Implications	85
5.4.2	Design Implications	86
5.5	Conclusion	87
VI	TEMPORAL DYNAMICS OF CREDIBILITY	95
6.1	Related Work	97
6.1.1	Collective Attention	97
6.1.2	Time Matters	98
6.2	Method	99
6.2.1	Pruning Corpus for Sample Independence	99
6.2.2	Credibility Classification	99
6.3	Statistical Measures	100
6.3.1	Collective Attention Metrics	100
6.3.2	Temporal Measures of Collective Attention	102
6.3.3	Statistical Analysis and Results	104
6.4	Discussion	105
VII	LIMITATION, FUTURE WORK & CONCLUSIONS	109
7.1	Limitations	111
7.2	Future Research Directions	112
7.2.1	Promoting reflective news reading	112
7.2.2	Ordering newsfeed by information quality	113
7.2.3	Bias aware social systems	113

REFERENCES 115

LIST OF TABLES

1	CREDBANK summary statistics.	30
2	Credibility classes and number of events in each class. The range of P_{ca} (proportion of annotations which are “Certainly Accurate”) for each class is also listed.	53
3	χ^2 tests of independence for Experiment 1.	55
4	Credibility classes and number of events in each class. The range of P_{ca} (proportion of annotations which are “Certainly Accurate”) for each class is also listed.	64
5	Sample of events from the CREDANK corpus grouped by their credibility classes. Events are represented with three event terms. Start and end times denote the time period during which Mitra et al. [121] collected tweets using Twitter’s search API combined with a search query containing a boolean <i>AND</i> of all three event terms. Ratings show the count of Turkers that selected an option from the 5-point, ordinal Likert scale ranging between -2 (“Certainly Inaccurate”) to +2 (“Certainly Accurate”). Each event was annotated by 30 Turkers.	89
6	List of feature categories used by our language classifier. Features are categorized as lexicon-based, non-lexicon based and control features. For the lexicon based measures I included words from each of the lexicons as features – yielding a total of 9,659 words obtained by summing the lexicon sizes. Adding the non-lexicon based features resulted in a total of 9,663 linguistic features.	90
7	Summary of different model fits sorted by % variance explained. <i>Null</i> is the intercept-only model. <i>Dev</i> denotes deviance which measures the goodness of fit. All comparisons with the Null model are statistically significant after Bonferroni correction for multiple testing. The table’s top half shows that the omnibus model containing controls and variables based on all linguistic measures for both tweets and replies is the best model. The bottom half of the table reports model performance for the omnibus model for each set of linguistic categories. It also shows deviance per linguistic category for original and replies (in gray).	91

8	Precision (P), Recall (R), F1-measure and Accuracy of two baseline classifiers: 1) Random Guess and 2) Random Weighted Guess, along with performance measures of the language classifier. I show four accuracy measurement schemes for our language classifier: 1) Unweighted is the most conservative way of measuring accuracy with no credit given for incorrect classification. It uses the unweighted credit matrix from Table 9a 2). Level-1 Weight _{0.25} gives partial credit of 0.25 if the classification is incorrect by one level only (Table 9b), 3). Level-1 Weight _{0.5} is similar but the rewarded partial credit is higher (0.5). Level-2 Weight _{0.25,0.5} gives partial credit as per the weighted matrix shown in Table 9c. Our language classifier significantly outperforms both the baselines (McNemar’s test, $p < 10^{-16}$).	92
9	Full credit is given for correct classification, denoted by 1’s along the diagonal. (a) No credit is given for incorrect classification (0’s along the non-diagonals). (b) Partial credit (0.25) is given if the classifier gets it wrong by one level and no credit is given if the predictions are off by two or more levels. (c) Partial credit (0.5) is given if the classifier gets it wrong by one level, (0.25) for two level and no credit if the predictions are wrong by three or more levels. There are four levels in the ordinal classes: Low (L), Medium (M), High (H), and Perfect (P).	92
	a Unweighted Credit Matrix	92
	b Weighted Matrix (Level 1)	92
	c Weighted Matrix (Level 2)	92
10	The top predictive words in the subjectivity category corresponding to the original tweets. Words associated with higher ($\beta > 0$) and lower ($\beta < 0$) levels of perceived credibility are shown in respective columns. All words are significant at the 0.001 level.	93
11	The top predictive phrases per linguistic category associated with higher ($\beta > 0$) and lower ($\beta < 0$) levels of credibility. Phrases corresponding to the original tweets are on the left, those corresponding to replies are on the right. All phrases are significant at the 0.001 level.	94
12	Pairwise statistical significance after Wilcoxon Rank Sum tests. P, H, M, L correspond to Perfect, High, Moderate and Low credibility classes. The top half of the diagonal corresponds to <i>message volume</i> , while bottom half shows pairwise differences in <i>people volume</i> . ns stands for non-significance. 107	
	a Minor peak attention pairwise statistical differences	107
	b Peak attention pairwise statistical differences	107

LIST OF FIGURES

1	Map of the steps taken to create the corpus (left side). The right side shows the corpus schema resulting from each step.	17
2	Turker interface for event assessment. The numbers correspond to a Turker’s workflow in annotating each item. 1) Click the search box launching a new windows of tweets corresponding to a search query, 2) Read tweets from the pop-up Twitter search window, and 3) Select one of the annotation options. If the selected option is ‘Event’, then the Turker is required to enter a short event summarization. Validation checks are put in place to ensure high-quality data.	24
3	Credibility scale, adapted from Saurí et al. [161]	26
4	Correlation between mean expert ratings and mean Turker ratings. For each number of Turkers ($n \in [1, 40]$) I compute the Pearson correlation ρ between Turker mean responses and expert mean responses. The plot shows average correlations after 10,000 resamplings.	27
5	Turker interface for credibility assessment. The numbers correspond to a Turker’s workflow. 1. Click the search box. 2. Read tweets from the pop-up Twitter search window. 3. Select one of the credibility scale options. 4. Provide a reason for the selection. Validation checks within the HIT ensure adherence to this workflow.	29
6	Frequencies of response distributions. The labels on the x-axis correspond to the splits in the annotation categories. For example, the bar at label ‘30’ correspond to the total number of events where all 30 Turkers agreed that the event is ‘Certainly True’. The bar at label ‘2,2,4,6,16’ corresponds to the count of events where the annotations are split five ways over the 5-point Likert credibility scale.	30
7	Percentage of events with majority Turker agreement. The majority agreement threshold is varied from 50% to 100% in steps of 10%.	32
8	Cluster dendograms for a sample of events and their corresponding average credibility curves.	34
9	Example pictures for three of the five possible data coding/annotation categories.	47
10	Example of the sentiment analysis annotation task.	47
11	Example of a topic list (with the intruder word highlighted with red text for illustration purposes).	48

12	Example of a tweet along with the five credibility coding/annotation categories modeled according to existing work on credibility annotation categories [23, 174].	49
13	(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 $V \in (0.15, 0.2]$, a moderate effect (Cramer’s $V \in (0.2, 0.25]$, and moderately strong Cramer’s $V \in (0.25, 0.3]$. (Bottom panel) Pearson correlation between expert and crowd annotations across all tasks.	57
14	Dendrogram from hierarchical clustering of the events from CREDBANK. The boxes show the four clusters.	65
15	The predictive power of the linguistic measures from the omnibus model. A measure’s weight is proportional to the deviance of the corresponding linguistic category (Table 7 lists the deviance numbers). The color saturation corresponds to the difference in the absolute values of positive and negative β weights of the coefficients belonging to the linguistic category. The color spans from red to green with a higher concentration of red denoting that the sum of negative β s is higher than the sum of positive β s, while the converse is true for higher concentration of green. The diagram also lists the top predictive phrases in each linguistic measure.	77
16	The time series of message volume for a sample event reported on Twitter. The event corresponds to Twitter discussions, where each tweet contained all three terms: “#chapellhillshooting”, “muslim” and “white”. The ● dot corresponds to the time window having maximum <i>message volume</i> while the ● dots correspond to the minor peaks observed in this volume. The inset diagram on the right side zooms in on one of the minor peaks, along with the rule triggering its designation.	100
17	Collective attention shown as a beanplot distribution. The shape of each half of the asymmetric bean represents the Gaussian density estimation of the distribution. The lines (in yellow) are the actual data points; the dotted long bean line is the median corresponding to the message volume, the solid line shows the median for people volume. The * denotes pairwise significant differences between cluster medians after correcting for familywise error-rate. (a). Proportion of minor peak fractions are statistically different across all credibility class pairs for both message and unique people volume. (b). Peak attention is significantly different across “Low” and “Perfect”, and “Moderate” and “Perfect” credibility classes for unique people volume, and “Perfect” and “Moderate” classes for message volume. The line charts at the bottom panel show the median trends across the credibility classes.	104

a	104
b	104
18	Time series of collective attention metrics (message volume and unique people volume) for example events in each credibility class. The examples show representative behavior of collective attention metrics in each credibility class. While events in all four classes are marked by peak attention with respect to both message and people volume, events in the low and moderate credibility classes exhibit multiple minor peaks, signifying that persistent attention is characteristic of lower credible social media events.	106

SUMMARY

Today, social media provide the means by which billions of people experience news and events happening around the world. We hear about breaking news from people we “follow” on Twitter. We engage in discussions about unfolding news stories with our “friends” on Facebook. We tend to read and respond to strangers sharing newsworthy information on Reddit. Simply put, individuals are increasingly relying on social media to share news and information quickly, without relying on established official sources [49]. While on one hand this empowers us with unparalleled information access, on the other hand it presents a new challenge — the challenge of ensuring that the unfiltered information originating from unofficial sources is credible. In fact, there is a popular narrative that social media is full of inaccurate information. But how much? Does information with dubious credibility have structure — temporal or linguistic? Are there systematic variations in such structures between highly credible and less credible information? This dissertation finds answers to such questions.

Despite many organized attempts along this line of research, social media is still opaque to the credibility of news and information. When you view your social media feed, you have no sense as to which parts are reliable and which are not. In other words, we do not understand the basic properties separating credible and non-credible content in our social feeds. This dissertation addresses this gap, by building large-scale, generalizable science around credibility in social media. Specifically, this dissertation makes the following contributions. First, it offers an iterative framework for systematically tracking the credibility of social media information. My framework combines machine and human computation in an efficient way to track both less well-known and widespread instances of newsworthy content in real-time, followed by crowd-sourcing credibility assessments. Next, by running

the framework for several months on the popular social networking site – Twitter, I present a corpus (CREDBANK) with newsworthy topics, their associated tweets and corresponding credibility scores.

Combining the massive dataset of CREDBANK with linguistic scholarship, I show that a parsimonious language model can predict the credibility of newsworthy topics with an accuracy of 68% compared to a random baseline of 25%. A deeper look at the most predictive phrases revealed that certain classes of words, like hedges, were associated with lower credibility, while affirmative booster words were indicative of higher credibility. Next, by investigating differences in temporal dynamics through the lens of collective attention, I demonstrate that recurring attentional bursts are correlated with lower credible events. These results provide the basis for addressing the online misinformation ecosystem. They also open avenues for future research in designing interventions aimed at controlling the spread of false information or cautioning social media users to be skeptical about an evolving topic’s veracity, ultimately raising an individual’s capacity to assess the credibility of content shared on social media.

CHAPTER I

INTRODUCTION

When we talk about newsworthy topics online, do we leave clues about the accuracy of the topic? If so, what are those clues? What about the way we discuss a news story or the style in which we respond to an unfolding event online? Do the words and phrases used in the commentary signal anything about the underlying credibility of the story? How many other people are paying attention to the same story? How often? When does that collective attention peak? How quickly does it die? The purpose of this thesis is to find answers to these questions and to show that such cues can be useful for inferring the credibility level of information shared in social media.

Online social networks provide a rich substrate for disseminating information through social ties. They have the unique ability to nurture looser but extensive social ties (known as *weak ties*) [10,56] and can foster shorter path lengths for information flow [190]. Together this has led to a scenario where individuals are increasingly relying on online communication technologies to consume news and information. A 2016 Pew survey shows that 62% of US adults obtain news from social media platforms [55]. While two-thirds of Facebook users (66%) get news on the site, six-in-ten or 59% of Twitter users get news on Twitter. However, modern online social networks like Facebook and Twitter are neutral towards the credibility of information. Simply put, they transmit both credible and less credible information without attending to its veracity. This is quite different from what happens in traditional mediums of information delivery. There, information is distributed only after verification by authorized gatekeepers. The lack of traditional journalistic gatekeeping in social media platforms often leads to scenarios where inaccurate information spread widely. A case in point of this phenomenon is social media's response to the Ebola outbreak in 2014.

When Ebola erupted in West Africa, a satirical website claimed that two Ebola victims had risen from the dead¹. This led to widespread panic about a potential “Ebola zombie apocalypse”, eventually flooding social media streams with inaccurate information about the disease. The detrimental effects of false tweets have also been witnessed by the financial sector. In 2013, a single false tweet claiming that Barack Obama has been injured in an explosion in the White House led to a stock market crash — \$130 billion was lost in a matter of seconds [117]. The phenomenon repeated again in 2015 when false tweets resulted in a sharp stock price plunge of two companies [2]. More recently the damaging consequences of misinformation was witnessed during the 2016 US presidential election. A sleuth of fake news websites were setup by teenagers in Macedonia to generate and distribute discredited claims through Facebook so as to earn advertisement dollars². These examples demonstrate that the ramifications of the rapid spreading of inaccurate information in social media can extend to multiple domains (politics, health, finance), ultimately leading to misinformed citizenry, and crippling our democracy. More generally, this points towards the need for studying credibility in social media.

1.1 Defining Credibility

This dissertation is about social media credibility. What is credibility *exactly*? To my surprise, I did not find one single converging answer to this simple question. Scholars from different fields have treated credibility differently. Some have defined it in terms of *believability*. Is the information worthy of being *believed*? Others consider credibility as a property of the source. Who is making the statement? Can I trust him? Yet others have treated credibility in terms of reliability. Is the information originating from a reliable source? Is the information itself reliable? Because of the lack of a single concurrent definition, it is fair to summarize that credibility is a diffuse concept and that it does not

¹<http://huzlers.com/breaking-news-ebola-victim-rises-dead-africa-fear-zombie-apocalypse/>

²https://www.buzzfeed.com/craigsilverman/how-macedonian-spammers-are-using-facebook-groups-to-feed-you?utm_term=.duOZ0VbrMK

have an exact analytical definition. This is extremely unsatisfying for most Computer scientists. As a computer scientist myself, we are used to concepts which have crisp exact definitions. But note the context of this study – online platforms where thousands of humans are receiving information from their social connections and deciding whether it is accurate. This scenario is far from being an exact analytical concept. In fact, I will argue that the absence of clear definition points to a rather diffuse viewpoint on credibility – one that refers to a general sense of accuracy constructed from multiple online social signals. Imagine a scenario where you see a breaking news story in your social feed. As you browse through the story, you find that a few of your social connections have questioned certain claims in the story, some have argued the legitimacy of the shared information, while others have straight-out denied certain statements. Based on these multiple social signals you come to the decision that the news event is not credible. This is how I have treated credibility in my work — it is a socially constructed definition of information accuracy. In other words, given a stream of social media posts representing a newsworthy event, I study the ways in which credibility manifests itself through social media traces. The dependent variable reflects responses to the following question, “How accurate do you think the event mentioned by the groups of tweets is?” The responses to this question are collected from a group of human raters and combined in a way which represents how an expert fact checker would assess the accuracy of a newsworthy topic after reading the online discussions corresponding to the news story. More on this later.

When representing credibility as a general concept of information accuracy, one obvious question which pops up is “Are you measuring whether the information is true or false?” This is a fair question, but defining what is “truth” is a philosophical debate and is related to a specific branch of philosophy called epistemology. In the history of epistemology there has been decades of controversies and arguments about the notion of truth and objectivity. Influential philosophers, such as Robert Putnam and George Lakoff have challenged the very notion of a "God's eye view of reality"—the notion that one can have the absolute

perfect knowledge or a privileged correct description of the world from outside [94]. Rather, they have argued for *internal realism*. An internalist view of ‘truth’ is based on the rational acceptability of reality by being part of that reality. Humans experience the realities of the world by being part of the world and by being within an existing social environment. Thus, the human epistemic situation is largely shaped by social relationships and social interactions. My definition of credibility is grounded in this notion of social epistemology. I operationalize the credibility of a real-world event by asking groups of human raters to assess the accuracy of the event based on social evidence.

1.2 Existing approaches to studying social media credibility

With social media’s rapid rise to prominence as a news source and its subsequent role in spreading rumor and misinformation, scholars have increasingly become interested in investigating social media information credibility. One trend in this domain is to study specific events that were subjects of misinformation or conducting indepth case studies of a handful of popular rumors. While useful in drawing out fine details underlying particular rumors, these studies fail to provide a holistic perspective of social media credibility.

On the other hand, tracking less well-known disputed information along with the widespread instances of newsworthy content is challenging. It requires sifting through massive amounts of social media posts, followed by a labor intensive task of content evaluation for credibility assessment. There has been some preliminary journalistic research on identifying, tracking and logging misinformation³. However, most of this work seems to be an arduous effort by a small number of journalists screening most of the data, or relying on externally reported instances of misinformation. My work addresses these challenges by systematically combining machine computation with multiple independent micro-labor annotations.

A parallel trend in this domain involves performing extensive quantitative analysis on

³<http://www.craigsilverman.ca/2014/09/02/researching-rumors-and-debunking-for-the-tow-center-at-columbia-university/>

social media traces—traces corresponding to historically reported cases of rumors, and building classifiers to predict the level of credibility. A common theme in this line of work is to treat credibility assessment as a two step process. The first step extracts newsworthy content, while the next step assesses the credibility of the retrieved content. My work builds on this basic two-step approach. However, I extend it by going beyond the traditional setting of post-hoc investigation of historical events, rather I investigate in real time every social media event that appears in a particular social media platform—Twitter.

1.3 Contributions of this thesis

This thesis makes the following specific contributions:

1. **A systematically constructed large-scale credibility corpus, CREDBANK.** I developed and validated a computational framework for tracking social media event streams followed by assessing the accuracy of the stream. My framework combines machine and human computation in a systematic way to track billions of streaming tweets continuously for more than three months, computationally summarizing those tweets into topics, and routing newsworthy topics to crowd workers for credibility annotations. The effort has resulted in the first large-scale systematically tracked social media corpus of credibility annotations. I have released CREDBANK’s dataset to enable future research on this rich corpus.
2. **An exhaustive comparison of strategies for obtaining high quality annotations from crowd-workers.** I performed an exhaustive set of controlled experiments to compare the performance of non-experts and experts across a variety of tasks of varying levels of difficulty. My goal here was to come up with the best strategy for obtaining high-quality credibility annotations from crowd-workers (non-experts). The exhaustive controlled experiments revealed that person-oriented annotation collection strategies (such as pre-screening workers for requisite cognitive aptitudes) results in high-quality annotations. The quality improvements when employing person-oriented strategies exceeds those

achieved by more complicated process-oriented strategies (such as iterative filtering or Bayesian Truth Serum techniques). These findings have implications for both the research-requester and the worker-coder for general Qualitative Data Analysis (QDA) tasks. Person-centric strategies, such as screening and training, have a one time cost for both parties and the cost gradually diminishes as dataset size increases.

3. **A parsimonious language model for predicting credibility.** Combining CREDBANK's dataset with linguistic scholarship, I developed a parsimonious language model which can determine credibility of newsworthy Twitter topics with 68% accuracy, where the random baseline is 25%. While not deployable as a standalone model for credibility assessment, these results show that certain linguistic categories and their associated phrases are strong predictors surrounding disparate social media events.
4. **An understanding of the temporal regularities of credibility.** Representing collective attention by the aggregate temporal signatures of an event reportage, I demonstrate that the amount of continued attention focused on an event provides information about its associated levels of perceived credibility. I show that events exhibiting sustained, intermittent bursts of attention were found to be associated with lower levels of perceived credibility. In other words, as more people showed interest during moments of transient collective attention, the associated uncertainty surrounding these events also increased.

1.4 Thesis Overview

This thesis is composed of three main parts, each focusing on progressively adding to our understanding of social media credibility.

- In Chapter 2, I review a large body of work on three related themes: multiple scholarly perspectives on credibility perceptions, credibility research done specifically in the context of online social media, and studies of event factuality and veridicality assessment.
- In Chapter 3, I describe the construction of CREDBANK. The corpus forms the basis for

understanding and modeling credibility from social media traces and guides the rest of the dissertation.

- In Chapter 4, I take a slight detour to describe a large controlled experiment that I had undertaken to perform exhaustive comparisons of the state-of-the-art crowd-sourcing techniques, so as to find the best technique to collect reliable high quality credibility annotations from crowd-workers.
- In Chapter 5, I present a parsimonious language model that maps language cues to different levels of credibility. I combine linguistic scholarship with CREDBANK's data to show that the language used by millions of people on Twitter has considerable information about an event's credibility.
- In Chapter 6, I demonstrate the temporal regularities of credibility. I also present implications based on the findings from the analysis of CREDBANK's corpus.

Finally, I conclude by outlining the future directions paved by this study and reflecting on the limitations of this work.

CHAPTER II

RELATED WORK

My work is informed by a large body of multi-disciplinary research. In this chapter, I review the literature organized around four main topics: social epistemology, perspectives on credibility, current social computing work on credibility, and scholarship on event factuality.

2.1 Perspectives on Credibility Perceptions

The study of credibility is highly interdisciplinary and scholars from different fields bring diverse perspectives to the definition of credibility [45, 148]. Credibility has been defined as believability [47], trust [69], reliability [163], accuracy [46], objectivity [40] and several other related concepts [67, 165]. It has also been defined in terms of characteristics of persuasive sources, characteristics of the message structure and content, and perceptions of the media [119]. While some studies have focused on the characteristics that make sources or information worthy of being believed, others have examined the characteristics that make sources or information likely to be believed [45]. Scholars have also argued that various dimensions of credibility overlap, and that receivers of information often do not distinguish between these dimensions, for example, between the message source and the message itself [26, 45]. Thus, despite decades of scholarly research on credibility, a single clear definition is yet to arise [67]. While communication and social psychology scholars treat credibility as a subjective perception on the part of the information receiver, information science scholars treat credibility as an objective property of the information, emphasizing on information quality as the criteria for credibility assessment [45, 47, 148]. CREDBANK's construction leans towards an information science approach and credibility assessment has been defined in terms of information quality. Moreover a significant number of studies view information quality as accuracy of information (see review by [148]). Following in their

footsteps, when I constructed CREDBANK, I focused on accuracy as a facet of information quality and instructed raters to rate the accuracy of social media events during the credibility assessment phase. Judging the accuracy of a real-world event from a social-stream of messages is an epistemic activity. In the next section, I illustrate notable concepts from epistemology which have helped operationalize this dependent measure.

One key component of credibility judgments that I did not explicitly consider is source credibility. While the classical treatment of credibility considers source of information as a key determinant of its reliability, source in online social media is a fuzzy entity because often times online information transmission involves multiple layers of source [180]. For example, a tweet from a friend shows you information about an event which the friend found from her follower, the follower saw it on a news channel and the news channel picked it up from an eye witness twitter account. Overall, this leads to a confusing multiplicity of sources of varying levels of credibility [180, 181]. As Sundar [180] rightly points out — “it is next to impossible for an average Internet user to have a well-defined sense of the credibility of various sources and message categories on the Web because of the multiplicity of sources embedded in the numerous layers of online dissemination of content”. Other studies have also shown that social media users pay much more attention to the content of the tweet than its author while assessing its credibility [123, 211]. Moreover, research by the linguistic community has demonstrated that perceptions of factuality of quoted content of tweets is not influenced by the source and the author of the content [175]. Motivated by these findings, I focus on linguistic and temporal markers. We envision that these markers can serve as meaningful cues to receivers of online content in assessing the relative accuracy of social media information.

2.2 Social Epistemology

Epistemology is the study of knowledge and justified belief. When positioned within a particular social and historical context, epistemology becomes social epistemology [54].

In the history of philosophy, traditional epistemology was heavily individualistic in focus. The emphasis was on evaluating judgments and attitudes of individuals outside of their social environment [54]. The shift from an individualistic focus to a more collective social focus started in the second half of the 20th century, when philosophers like Thomas Kuhn, Michel Foucault, and Hilary Putnam denied the notion of objectivity and absolute truth. The argument presented by traditional epistemology, that there can be “exactly one true and complete description of the way the world is”, was first challenged by Putnam in what he referred to as “internal realism” [94]. It is a perspective that articulates that humans function as part of the reality and not outside that reality. Hence it is impossible for us to ever stand outside a real world happening and observe reality with perfect knowledge and absolute awareness of the God’s eye variety. While operationalizing credibility as a social construct of accuracy, I follow Putnam’s lead and take this internalist perspective. From an internalist perspective, “truth” and “accuracy” of a real-world event corresponds to rational acceptability of reality by being part of that real human experiences and human knowledge of the world.

Sociologists and science practitioners such as Bruno Latour and Steve Woolgar have also denounced the traditional notion of absolute truth [96]. In their social studies of scientific facts, they have reasoned how the acceptance of scientific facts goes through a social construction process. Initially, when a new scientific fact is discovered, it is simply recorded in a scientific article. To reach the state where a scientific fact is broadly accepted as “true”, other people need to replicate the stated claim, cite the claim and accept the newly stated fact without contesting it. In other words, the acceptance of scientific facts needs to go through a scientific discourse and discourse inherently is a social phenomenon. Credibility of a social media event also goes through this social construction process, although at a much faster pace than scientific facts. While scientific inquiry—all the way from fact-finding to fact-acceptance—takes months or even years, social-media event reportage—from discovery to a complete interpretation of the event—takes a few hours or even minutes, and is often a

continuous journey that builds over time.

The existence of true objectivity has also been challenged in the field of journalism. While describing “journalistic truth”, journalists Bill Kovach and Tom Rosenthal emphasized the concept of “functional truth”—truth by which we can operate on a day-to-day basis. They said that a journalistic event is reported based on facts collected at the time [86]. Actions, such as police arresting suspects, judges holding trials, or juries rendering verdicts, are taken based on what we know about the event at the time. All of these “truths”, they argued, are subject to revisions, and that we operate in them in the meantime because they are necessary and they work. While studying credibility of newsworthy social media events, I ground my definition of credibility on this practical and functional form of journalistic truth and refrain from alluding to truth in the absolute or philosophical sense. Additionally, I operationalize the credibility assessment task by drawing heavily from social epistemology [54]. Social epistemology suggests that there are different ways in which an epistemic activity can be “social”. One of the ways is for the individual agent to base her decision on social evidence. The second way that an epistemic activity can be social is when a collection of individuals are making the judgment. The third branch emphasizes on assessing the epistemic quality of collective judgments. I took into consideration all these branches of social epistemology while designing the credibility assessment task. The first branch of social epistemology is reflected in how individual human raters judged the accuracy of a social media event. Each rater independently referred to a stream of tweets pertaining to that event. Their judgments were based on the social evidence found in the twitter stream. For example, if there are multiple conflicting tweets about an event, the rater might consider the event accuracy to be “uncertain”. Whereas, if a majority of tweets are disagreeing with the happenings of the event, it is quite likely that the rater would mark the event as “probably inaccurate”. The second branch of social epistemology is exhibited in the way credibility judgments were collected from a group of human raters (30 Amazon Mechanical Turk workers) and then aggregated based on majority rule. An obvious concern in a collective

activity scenario is that different choices of participants can vary the degree of epistemic success achieved. Hence, in accordance with the third branch of social epistemology, I also assess the epistemic quality of the collective judgments by systematically comparing Turk workers' collective judgment accuracy with expert-level accuracy. The comparison resulted in a configuration that achieves consistently high epistemic quality of credibility judgments at par with expert level responses.

2.3 Social Media and Credibility

Social media has quickly risen to prominence as a news source [25]. Yet lingering doubts remain about the quality of information obtained through it. Quality comprises occur due to spam content [58], stealthy advertising [126, 145, 178] and rumor and misinformation [31, 44, 79, 118]. Thus, assessing the credibility of social media information is of growing importance and has attracted the attention of various social media researchers. One trend in this domain is to study specific events that were subjects of misinformation. For example, studies have tracked the spread of rumors during the 2011 Great East Japan earthquake [107], provided descriptive analysis of rumors during the 2013 Boston bombings [111] and reported a case study of rumor dynamics in a Chinese microblogging community [105]. Together, these studies suggest the importance of anxiety, personal involvement and informational ambiguity in spreading rumors. However, their findings are based on these extreme, hand-selected cases of rumor. Scholars have also conducted survey studies, analyzing user's ability to assess credibility of microblog content [123]. Their findings suggest that users are not adept at judging credibility simply by reading the content alone. Instead their assessment is influenced by heuristics such as the author and topic of the information. While these small scale survey experiments and studies on one-off instances of rumor are useful, they fail to provide a holistic perspective of social media credibility.

A parallel trend of work within this domain is developing the capability to predict the credibility of information communicated through social media. For example, researchers

have collected documented rumor cases from popular urban legend websites (i.e., Snopes) and have analyzed their corresponding Facebook posts to find new insights on rumor mutations [49]. Similar techniques have been used to identify temporal and structural features of rumor in Twitter [91]. Predictive analysis of information credibility is also a popular trend in this area, such as building classifiers to detect whether tweets are factual or not [24], automatically assessing the credibility level of a set of tweets [142], automatically classifying rumor stances expressed in crisis events [206], or detecting controversial information from inquiry phrases [209], or identifying fake images in Twitter [62]. The common approach in this line of work treats credibility assessment as a two step process, where the first step extracts newsworthy content, while the next step determines the credibility of news topic. This basic two step process underpins the methodology of my work. However, I extend it by going beyond the traditional setting of post-hoc investigation of historical events, investigating in real time every social media event.

2.4 Event Factuality

A closely related concept to event credibility is factuality assessment of events. Social scientists and linguists have been interested in studying language dimensions of event factuality for decades. They have referred to event factuality as the factual nature of eventualities expressed in texts [159]. This factual nature can encompass facts which actually took place, possibilities that might have happened or situations which never occurred. One of the leading trends in event factuality research is generation of factuality-related corpora. For example, the TimeBank corpus was compiled from news articles annotated with temporal and factuality-relevant information of events [141]. The MPQA Opinion corpus includes annotations regarding the degree of factuality of expressions [196]. Thus, annotations categorize expressions as opinions, beliefs, thoughts or speech events, and these states convey the author's stance in terms of objective or subjective perspective. Sauri's FactBank corpus has become the leading resource for research on event factuality [159]. FactBank's

annotations are done on a rich set of newswire documents containing event descriptions. The aim of these text-based annotations is to determine ways in which lexical meanings and semantic interactions affect veridicality judgments. My work on building the CREDBANK corpus is inspired from this rich set of linguistic corpora, each capturing different aspects of event factuality. Following their example, I also took the first step in analyzing language dimensions of credibility from the CREDBANK corpus – a leading resource for research on information credibility of social media content [121].

Another noteworthy work in this area is Rubin’s theoretical framework for identifying certainty in texts [153]. Findings from her work reveal that linguistic cues present in textual information can be used to identify the text’s certainty level. Surprisingly, her results demonstrate that certainty markers vary based on content type. For example, content from editorial samples had more certainty markers per sentence than did content taken from news stories. These results prompted me to look for credibility markers in the context of social media events reported via variations in message type, for example, reporting via an original post or response to an existing post. Perhaps some of these markers share the same principles as the certainty markers of textual content. My work on building a predictive algorithm for credibility assessment hinges on using these markers as cues for assessment.

Another trend in this area of research is developing tools for automatically identifying factuality related information. Most efforts in this direction are based on existing corpus data. For example, the MPQA Opinion corpus has led to the development of OpinionFinder [199], a tool for subjectivity analysis. The TimeBank corpus has been used to identify polarity and modality using contexts and grammatical items [141]. Grammatical markers of polarity and modality have also been recognized by the Evita tool using simple linguistic and statistical techniques [160]. My work on developing the CREDBANK (corpus) followed by building a parsimonious language model for credibility assessment is inspired by this two step trend in the linguistic community – *build a corpus; develop automatic tools* to identify the key phenomenon captured by the corpus.

CHAPTER III

BUILDING A LARGE-SCALE CREDIBILITY CORPUS

With social media’s growth as a news resource [25] and its purported role in spreading false rumors and misinformation, scholars have paid significant attention to the problem of *credibility assessment*. Earlier work has looked at automatic detection of credibility [24,142], the diffusion patterns of rumors [49], building interactive tools to allow investigation of these patterns [147] and exploring the factuality of various claims [175]. However, because of the inherent difficulty associated with collecting large-scale rumor data, previous work has had to *select on the dependent variable* [187] — presuming a priori what rumors look like (i.e., constructing retrieval queries) or working from a known set of rumors or post hoc investigation of prominent events with known disputed information or credibility judgments of specific topics trending on social media. In the first quarter of 2014, I started to deliberate on the following question: how can we address the problem of selection bias, which has plagued traditional approaches that examine historically reported rumors. In other words, how can we collect large-scale social media data without making a-priori assumptions of what is a rumor or what constitutes verified news. The answer was rather simple — track all social media events and find how credible they are at that point in time. However, operationalizing this approach was a challenge in itself. It took several months of trial and experimentation to finally come up with a framework which overcomes this problem of *sampling bias*. My iterative framework combined machine learning and human computation in an efficient way to track all large-scale events happening on Twitter, followed by crowd-workers rating the credibility of Twitter coverage at the time. The effort resulted in the first large-scale corpus of credibility annotations of social media data — CREDBANK — a corpus of tweets, topics, events and associated human credibility judgments.

In this chapter, I describe the development and validation of the framework on which CREDBANK is based, as well as a brief statistical overview of the corpus. I tracked more than 160 million streaming tweets, computationally summarizing those tweets into events, and routing the events to crowd workers for credibility annotation. By guiding annotators through a framework inspired by theoretical work, I show that crowd workers can approximate the credibility judgments of University-employed reference librarians, the gold standard used in this work. In total, CREDBANK comprises more than 66M tweets grouped into 1,377 real-world events, each annotated by 30 Amazon Mechanical Turk workers for credibility (along with their rationales for choosing their annotations). The primary contribution of CREDBANK is the set of new research questions it makes possible. For example, social scientists might explore what role the mainstream media plays in online rumors; a data mining researcher might explore how the temporal patterns of rumor differ from highly credible information; a health researcher could investigate how folk theories of a new disease (the emergence of Ebola is captured in CREDBANK) diffuse through a population.

3.1 CREDBANK's Construction

CREDBANK is built on an iterative framework of the following five main phases, combining machine computation (MC) and human computation (HC) in an efficient way:

1. Streaming Data & Preprocessing (MC)
2. Online Event Candidates using Topic models (MC)
3. Filtering Event-specific Topics (HC)
4. Credibility Assessment of Topical Events (HC)
5. Collection of Topical Event Streams (MC)

Figure 1 presents the construction of CREDBANK graphically, while Algorithm 1 presents the same method, only procedurally.

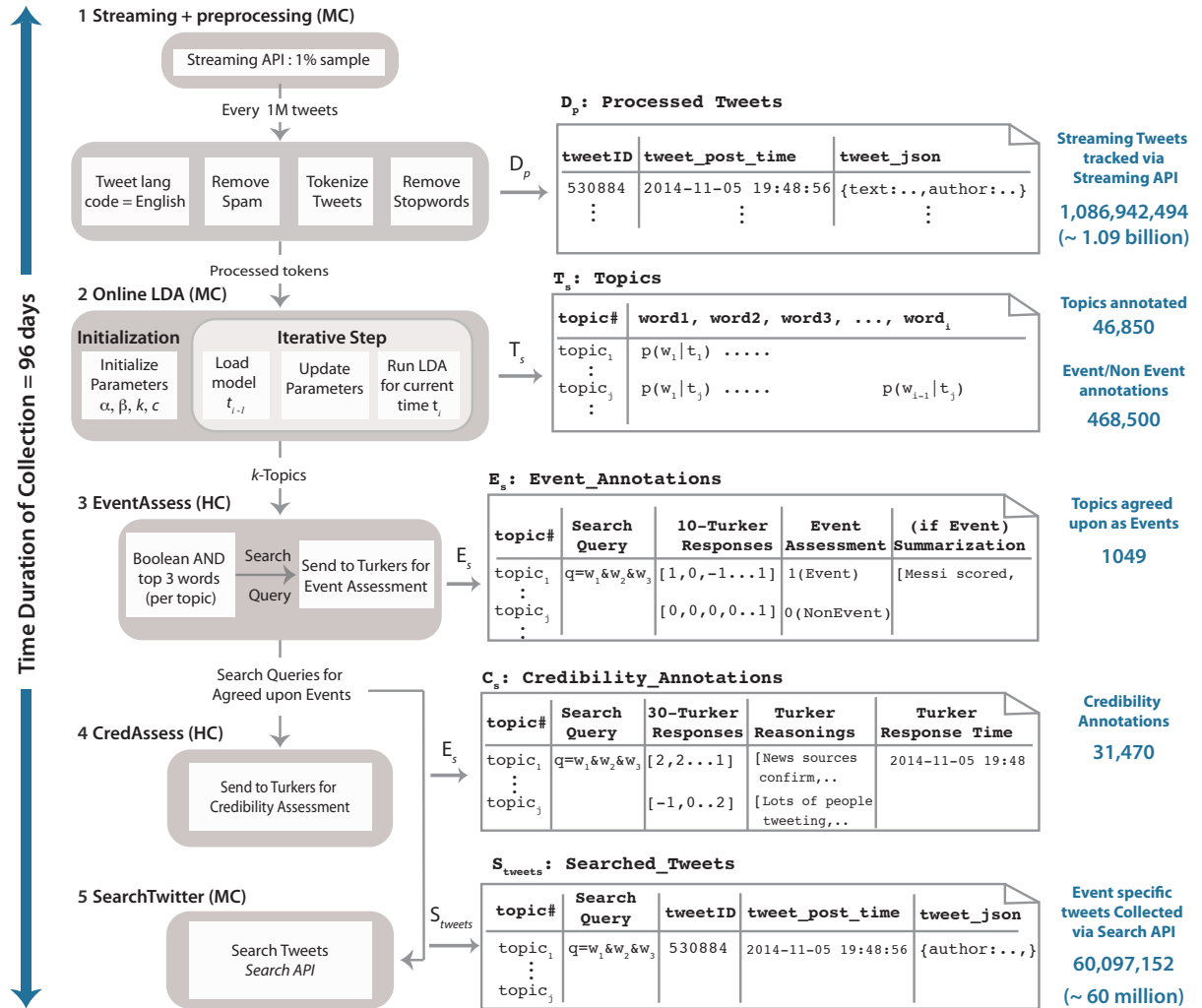


Figure 1: Map of the steps taken to create the corpus (left side). The right side shows the corpus schema resulting from each step.

3.1.1 Streaming Data and Preprocessing

I used the Twitter streaming API to collect a continuous 1% sample of global tweets¹ [140]. For every group of 1 million streaming tweets, I first filter out English only tweets identified by Twitter’s machine language detection algorithm (annotated in the tweet’s metadata returned from the API).

¹<http://github.com/reidpr/quac>

Algorithm 1: Credibility assessment workflow (streaming)

Input: D_S : Set of 1M streaming tweets
1 $|W|$: Time window of collected tweets
2 K : Number of topics
3 i : Iteration count
Output: T_S : Set of topical events
4 C_S : Set of credibility annotations of topical events
5 S_{tweets} : Set of searched tweets
6 **begin**
7 $TO_p \leftarrow PreProcess(D_S)$
8 $T_S \leftarrow OnlineLDA(TO_p, |W|, k, i)$
9 $E_S \leftarrow EventAssess(T_S)$
10 $C_S \leftarrow CredAssess(E_S)$
11 $S_{tweets} \leftarrow SearchTwitter(E_S)$

3.1.1.1 Spam Removal

Next, I apply standard spam-removal techniques to filter out tweets which might have escaped Twitter’s spam detection algorithms. Adopting the practices of Firefox’s Clean Tweets [1] add on, I filter out tweets from accounts less than a day old and eliminate tweets containing three or more hashtags [89]. Additionally, I also remove tweets where the entire text is in upper case. Next, following [29]’s approach, I check whether text patterns within a tweet correspond to spam content. This step detects tweets which look like prototypical spam, such as *buy Viagra online without a prescription* or *get car loan with bad credit*. I first compiled a list of textual spam patterns based on existing lists of spam trigger words [3, 152, 170]. A tweet is marked as spam if it contains a URL and a spam phrase—a standard technique used by spammers to direct users to spam websites [29]. To illustrate, an example tweet filtered by this step is “100% **all natural** way to get rid of pigmentations and uneven skin tone [URL].”

3.1.1.2 Tokenization

I tokenized tweets using a regex-based Twitter tokenizer [130]. While Traditional Penn Treebank style tokenizers work well on more-structured texts (like news articles), they perform poorly on social media text, often breaking punctuation, emoticons and unicode characters into a single token per character [133].

1 PreProcess**Input:** D : Set of tweets as documents**Output:** TO_p : Set of tokens**2 begin**

```
3    $D_p \leftarrow D$ ;  
4   foreach  $tweet \in D$  do  
5     if  $isLangEnglish(tweet.text)$  then  
6       // Remove spam  
7       if ( $oneDayAge(tweet.account)$  OR  
8          $isAllCaps(tweet.text)$  OR  
9          $countHashTags(tweet.text) \geq 3$  OR  
10         $hasSpamPattern(tweet.text)$  AND  $hasURL(tweet.text)$ ) then  
11          $D_p \leftarrow D_p.remove(tweet)$   
12    $TO_p \leftarrow regexTokenizer(D_p)$   
return  $TO_p$ ;
```

13 OnlineLDA**Input:** $TO_p, |W|, K, i$ **Output:** T_s : Set of k topics as word probability distributions**14 begin**

```
15   if  $i == 1$  then  
16      $\alpha_0 \leftarrow 0.001, \beta_0 \leftarrow 0.01, c \leftarrow 0.5$ ;  
17     Run LDA on tokens  $TK_p$  corresponding to tweets in window  $|W|$  ;  
18   else  
19     // Iterative step  
20      $m_{i-1} \leftarrow$  Model from iteration  $i - 1$ ;  
21      $TK'_p \leftarrow$  Updated tokens for tweets in iteration  $i$ ;  
22      $\alpha', \beta' \leftarrow$  UpdatePriors( $m_{i-1}$ );  
return  $T_s$ ;
```

23 EventAssess**Input:** T_s : Set of k topics as word probability distributions**Output:** E_s : Set of Twitter Search Queries agreed upon as Events**Output:** B_s : Set of event summarizations entered by Turkers**24 begin**

```
25    $W_{top} \leftarrow$  Set of top 3 most probable word tuple sequences =  $\{w_1, w_2, w_3\}$   $q_e \leftarrow$  constructQuery( $w_1$  AND  $w_2$  AND  $w_3$ );  
26    $EventAssess.PostHits(Q_s)$ ;  
27   if  $hitFinished$  then  
28     foreach  $topic \in T_s$  do  
29       if  $|Turkers\ mark\ as\ events| \geq 6$  then  
30          $E_s.add(topic, q_e)$ ;  
31          $B_s.add(summarizations)$ ;  
32   return  $E_s, B_s$ ;
```

33 CredAssess**Input:** E_s : Set of Twitter Search Queries corresponding to real-world Events**Output:** C_s, R_s : Set of 30 credibility ratings and reasons for each event in E_s **34 begin**

```
35    $C_s \leftarrow ()$ ;  
36   if  $notBlank(E_s)$  then  
37      $CredAssess.PostHits(E_s)$ ;  
38     if ( $hitFinished$ ) then  
39       foreach  $event \in E_s$  do  
40          $C_s \leftarrow C_s.append(rating)$ ;  
41          $R_s \leftarrow R_s.append(reason)$ ;  
42   return  $C_s, R_s$ ;
```

3.1.1.3 Stop-word Removal

Next, I employ multi-stage stop-word removal. The purpose of doing this is to select a vocabulary for topic modeling (discussed shortly) that eliminates overly generic terms while retaining terms which are frequent in the corpus, but infrequent among general Twitter messages [130]. Specifically, I had the following stages of stop word removal:

1. First, I remove a list of function words based on the standard SMART stop list dictionary [103].
2. Next, I filter out Twitter specific stop words, like *RT*, *follow*, *follows*, *@username*.
3. The earlier two approaches of removing stop words based on a static list is often plagued by being too generic. To overcome this limitation, several methods to automatically generate stop words have emerged. I adapt one such method for our purposes—a method based on Zipf’s law [113, 154, 186]. Zipf [210] observed that in a data collection the frequency of a term (TF) is inversely proportional to its rank. Inspired by this observation, several popular stop-word removal methods assume that stopwords correspond to top ranking terms (i.e., the most frequent words or TF-High) [186, 204], while others label both top (TF-High) and low ranked words (TF-Low) as stopwords [113]. Thus, in addition to the static stop-word list, I remove both: 1) the most frequent words (TF-High), corresponding to the top 50 words at every iteration, and 2) words that occur fewer than 10 times (TF-Low).

Overall, this results in a dynamic vocabulary generated at each iteration that I send to the event detection phase of the workflow.

3.1.2 Event Candidates using Online LDA

The basic premise of this phase is rooted in our main objective: gathering credibility assessments of real-world events as reported by social media streams. In this regard, the concept of events defined by the information retrieval community fits my study objective

[5, 14]. As per their definition, an event is a real-world occurrence happening at a specific time and associated with a time-ordered stream of messages. Hence, the primary goal of this phase is identifying such real-world events.

Event detection work in the information retrieval community has mostly proceeded in two branches: *Retrospective event detection*, where events are detected post-hoc on static sets of historical data [5] and *Online event detection*, where events are detected in near real-time by processing dynamic sets of streaming data [97, 104]. The online approach has been argued to be more useful over traditional retrospective methods because of its ability to provide real-time responsiveness and the option to allow trend analysis of streaming data. Hence, I chose to apply online event detection in the present work. Next, I had to decide on the specific technique for online event detection. A range of online approaches exist, ranging from simple keyword-based methods [35, 194], bursty term analysis techniques [132, 138] to more sophisticated topic-modeling based methods [6, 68, 97]. The disadvantage of keyword-based approaches using a pre-defined set of keywords is that it fails to capture rapidly evolving real-world events. Bursty term analysis is based on bursts in term frequencies to overcome the limitations of pre-defined keyword-based techniques. However, it still cannot capture multiple co-occurring terms associated with an event [205]. On the other hand, topic models can learn term co-occurrences associated with an event, making them a much better choice to capture quickly evolving real-world events. Among several variants of online topic models I opted for Lau et al's [97] online Latent Dirichlet allocation (LDA) [17] with Gibbs sampling [57] primarily for two reasons. Firstly, their online LDA model allows iterative processing of streaming text while controlling the model from growing progressively over time. Secondly, unlike other models with assumptions of fixed vocabularies, this model provides the flexibility to regenerate the vocabulary at each iteration, thus facilitating the data preprocessing steps discussed in the previous section.

I ran the online LDA model iteratively for every set of 1 million streaming tweets. The input to the LDA is a bag-of-words representation of individual tweets that passed the

preprocessing steps. The model output is a set of latent topics represented as a set of related words which tend to co-occur in similar tweets. In other words, the topic modeling step segregates a collection of tweets into coherent topics, where each topic can be interpreted by the top N terms with the highest marginal probability $\rho(w_j | \phi_k)$ — the probability associated with each term w_j in a given topic ϕ_k .

One important step in our process is setting the model parameters. The model is initialized with Dirichlet prior values $\alpha = 0.001$ and $\beta = 0.01$, where α controls the sparsity of document-topic distribution. A low value of α is preferred, because it produces a sparse distribution, leading to very few topic assignments per tweet. This intuitively makes sense, because it is almost impossible to mention large number of topics in a 140 character long tweet. The model also takes the number of topics k as an input parameter. I empirically evaluated the sensitivity of our model against a range of k -value settings, while keeping the other parameters constant. I tested this on a dataset of 1 million tweets. Recall, that this is the size of our streaming data at every iteration of our workflow. The value k converges to $k = 50$. Such a high value of k also allows us to capture more granular topics rather than some high level general topics. Another input parameter to the model is the contribution factor c , which determines the contribution of the model learned in the previous iteration ($i-1^{th}$). Following a related approach [97], I set $c = 0.5$, so as to give equal weighting to tweets collected in successive iterations. The set of topics from each iteration forms our set of candidate events.

3.1.3 Filtering Event-specific Topics

One potential problem of a purely computational approach for event detection is occasional false positives in our set of candidate events. In other words, non-event activities such as conversations or opinions may be marked as events. To prevent these from being distractors in our main credibility assessment task, I turn to human judgments. I recruited independent human annotators from Amazon Mechanical Turk (AMT) to decide whether a tweet is

truly about an event or not. AMT is a widely used micro-labor market for crowd-sourcing annotation tasks.

3.1.3.1 Task Design (Event Annotation)

What is the best strategy to obtain high quality annotations from crowd workers? In order to answer this question, I ran an exhaustive set of experiments testing I designed a large controlled experiment to perform exhaustive comparisons of various crowd-sourcing techniques, including targeted screening, recruiting Turkers iteratively, offering them financial incentives, incentivizing on Bayesian Truth Serum, etc.—a total of 34 intervention strategies and 68,000 Turker annotations for a range of tasks of varying levels of difficulty were explored. Chapter 4 outlines the details of the controlled experiment. The outcome of the study found that selectively screening and training workers, followed by offering financial incentives, is the best strategy for obtaining quality data annotations [122]. Thus, I designed our crowd event annotation framework to first selectively screen and train workers via a qualification test. The test first provides rubrics to differentiate between events and non-events, followed by accompanying examples describing how to do the task. To come up with a definition of events, I traced back to research done for topic detection and tracking tasks [5] and presented the following definition to our human judges:

Event Tweets: Tweets which are about a real-world occurrence, which is considered to happen at a specific time and place, and which is of interest to others and not just to one’s Twitter friends.

Non-Event Tweets: Tweets which are personal opinions, personal updates, conversations among friends, random thoughts and musings, promotions and advertisements.

Overall, this served as a training phase for the workers. Next the workers were screened on the basis of their score on the qualification test. The qualification test questions were specifically designed for performing the event annotation task. The purpose was to provide task-specific orientation to the workers. Only those who scored a minimum of 75% were

Identify whether a group of tweets is about an event or not

Instructions

Tweets are short messages posted by users in a social media site called Twitter. In this task you will be presented with a group of tweets searched using a few keywords in Twitter. Your task will be to identify if **MOST** of the tweets in the group is about an event or not. The qualification test you passed earlier provided you with instructions of how to do the task. Just to recap, below is the definition of **Event** and **Non-Event** tweets:

- **Event Tweets** - Tweets which are about a real-world occurrence, which is considered to happen at a specific time and place, and which is of interest to others and not just to ones Twitter friends.
- **Non-Event Tweets** - Tweets which are personal opinions, personal updates, conversations among friends, random thoughts and musings, promotions and advertisements.

NOTE: Grading for bonus only occurs after all the assignments for the the majority at least 80% of the time, then you will be given a \$0.15 bo

Please wait while we load the tweets ...

Task:

Tweet Group 1

Click to Open Twitter Search box

1

The group of tweets can be categorized as **Required**

Event Tweet Non-Event Tweet Not Sure

Summarize the event in a few keywords

Tweet Group 2

Click to Open Twitter Search box

3

The group of tweets can be categorized as **Required**

Event Tweet Non-Event Tweet Not Sure

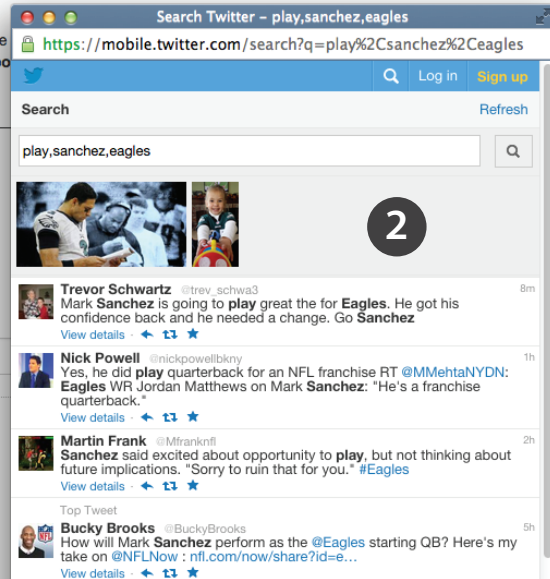


Figure 2: Turker interface for event assessment. The numbers correspond to a Turker’s workflow in annotating each item. 1) Click the search box launching a new windows of tweets corresponding to a search query, 2) Read tweets from the pop-up Twitter search window, and 3) Select one of the annotation options. If the selected option is ‘Event’, then the Turker is required to enter a short event summarization. Validation checks are put in place to ensure high-quality data.

allowed to work on the annotation task (also called Human Intelligence Tasks, or HITs).

I designed the annotation task such that it closely mimics the way a person would search Twitter to see information related to an event. The advantage of this approach is that our data collection closely reflects everyday practices. For each of the $k = 50$ topics, I first take the top 3 terms and create a Twitter search query by using a boolean AND over all three terms. Next, for each of these k queries corresponding to the k topics, I create an annotation item by embedding a Twitter search query box inside a HIT (see Figure 2). While annotating an item in a HIT, the worker has to first click the search box, see a set of real-time tweets and choose to label the tweet stream as representing: 1) Event Tweets, 2) Non-Event Tweets,

or 3) Not Sure. If she selects the **Event** option, she is further prompted to summarize the event in a free-form text field. There were two factors which guided my design decision to elicit such subjective responses. First, subjective responses tend to improve rating quality and minimize random clicking [84]. Second, it allows me to filter out any ill-formed search queries—queries which are too broad to surface any specific Twitter event. The intuition is that such a query will likely return non-coherent tweets, making it difficult for humans to find a coherent theme and come up with a meaningful summary.

3.1.3.2 Determining Events

An initial pilot study confirmed this intuition. During the pilot phase, I also determined the number of Turkers needed to correctly annotate an item as an event. I manually checked the Turker annotations and found that if 6 out of 10 independent Turkers agreed that the topic is an event, then the majority label matched expert annotation. The pilot study also helped me determine the number of items to post per HIT so that a single HIT is completed within a short duration (under 2 mins), ensuring lower cognitive load per HIT. I posted 10 items per HIT and paid \$0.15, adhering to minimum wage requirements. I also offered an additional \$0.15 bonus to workers whose responses matched the modal response of the crowd. For each item I asked 10 Turkers for their independent ratings, and if a majority agreed that the topic is an event, then I added the topic to the queue for credibility assessment. I purposely choose a conservative inter-rater agreement (6 out of 10 agreements) to ensure maximum precision in the event detection step.

3.1.4 Credibility Assessment of Topical Events

In this step, I gather all the topics which were agreed upon as **Events** in the previous phase and recruit Turkers to rate their accuracy. Here I need to address two important factors while designing the credibility annotation task: 1) the scale of annotation, and 2), the correct number of workers to recruit per item so as to get reliable and quality annotations.

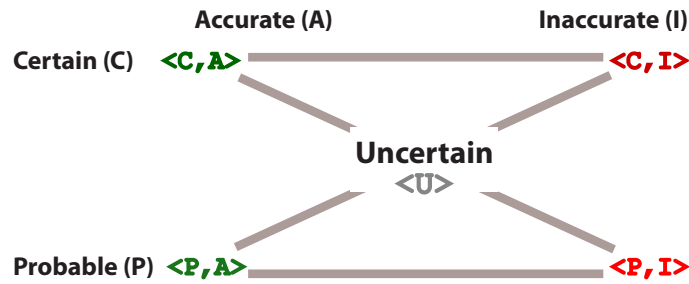


Figure 3: Credibility scale, adapted from Saurí et al. [161]

3.1.4.1 Determining the Credibility Scale

Credibility assessment bears close resemblance to research done on ‘Event Factuality’ by the linguistic community [38, 153, 161], where factuality of an utterance is expressed in terms of polarity and modality markers. Building on their work, I represent event credibility as an interaction between two dimensions: *Polarity* which distinguishes among ‘Accurate’, ‘Inaccurate’, and ‘Uncertain’, and *Degree of certainty* which differentiates among ‘Certainly’, ‘Probably’ and ‘Uncertain’. Figure 3 shows the interaction between the two dimensions by mapping this into a Square of Opposition (SO) diagram—a diagrammatic representation of a set of logical relationships [136]. Thus, event credibility is represented as a tuple $[degree, polarity]$, forming a set of tuple sequences as: { “*Certainly Accurate*”, “*Probably Accurate*”, “*Uncertain*”, “*Probably Inaccurate*”, “*Certainly Inaccurate*”}. In others words, credibility assessment is based on a 5-point Likert scale ranging from ‘[-2] Certainly Inaccurate’ to ‘[+2] Certainly Accurate.’

3.1.4.2 Determining Number of AMT Workers Necessary for Quality Annotations

In order for the credibility annotations to be useful and trustworthy, an important criteria is collecting high quality annotations—annotations which are at par with expert level responses. While fact-checking services have successfully recruited librarians as expert information providers [87], their limited time and availability makes it impossible to scale real-time expert annotation tasks. Moreover, with a small pool of in-house experts and a constant update of Twitter streams, near real-time annotation is infeasible. Crowd-sourced

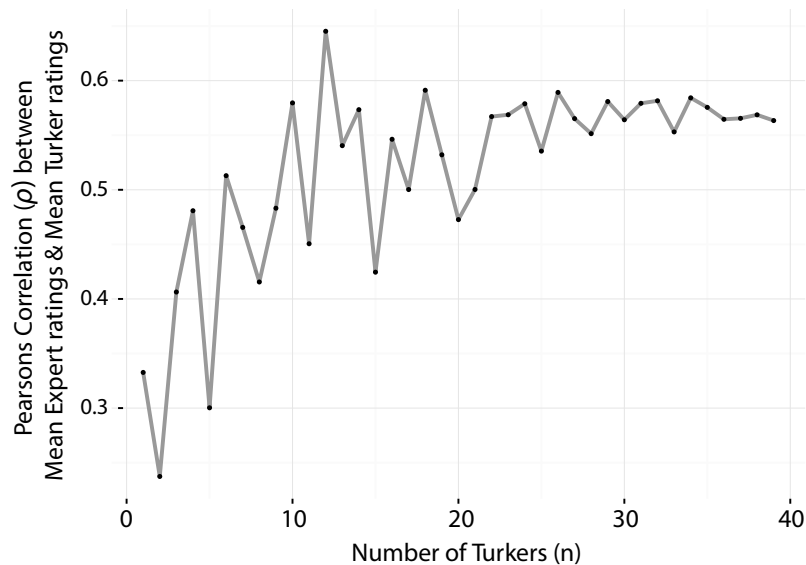


Figure 4: Correlation between mean expert ratings and mean Turker ratings. For each number of Turkers ($n \in [1, 40]$) I compute the Pearson correlation ρ between Turker mean responses and expert mean responses. The plot shows average correlations after 10,000 resamplings.

micro-labor markets like AMT are a promising option for addressing these challenges. But can Turkers provide credibility annotations roughly equivalent to those provided by experts? One standard way of addressing the issue of collecting quality responses is by redundancy—taking the majority or averaging over multiple independent responses [182].

Naturally a question arises as to how many Turker responses will closely approximate an expert’s judgment. To determine this number, I piloted this system over a span of 5 days until I finished collecting and annotating 50 events both by Turkers and expert annotators. I recruited reference librarians from our University library as a our expert raters. For each event, I collected 40 Turker ratings and 3 expert ratings. The web interface shown to librarians were similar to the one shown to Turkers. To estimate the effect of using n Turkers, I randomly sampled n ratings for each annotation item ($n \in \{1, 40\}$). Then I took the mean of these n ratings and computed Pearson correlation between Turker mean responses and expert mean responses. I controlled for sampling error via bootstrapping—recalculating the correlations 10,000 times for each n and then averaging over the 10,000 re-computations (Figure 4). The correlations keeps increasing and finally levels off at 30. Hence I fixed the

number of workers to 30 to obtain reliable and quality credibility annotations.

3.1.4.3 Task Design (Credibility Annotation)

For designing the annotation tasks (or HITs), I follow the same principles as in the earlier phase. Each HIT had items corresponding to only those topics which were determined as events in the previous step. A Twitter search query box corresponding to the topic is embedded in the HIT (see Figure 5). A worker performing the annotation task has to first click the search box to see the real-time tweets and then choose one of the options from the 5-point Likert scale. Next, the Turker is prompted to enter a reason behind their choice. Asking workers for such free-form subjective responses, while on one hand improves annotation quality, on the other hand adds an extra dimension to our annotation framework.

Similar to the previous phase, I selectively screen and train workers through a task specific qualification test, requiring workers to score at least 75%. An initial pilot study helped us determine the number of items to post per HIT so as to ensure shorter time durations per HIT (under 3 mins) and lower per-HIT cognitive load. I allowed a maximum of 5 items to be posted in a single credibility assessment HIT.

3.1.5 Collection of Topical Event Streams

In the final phase, I collected tweets specific to each of the topical events returned by `EventAssess`. Using the Twitter Search API with search queries corresponding to each of these events (q_e), I collect the most recent tweets, going as far back as the last 7 days—the limit imposed by the search API. Intuitively, this doesn't seem to be a severe limitation because my method tracks recent events as they appear in the stream, followed by their annotation.

3.2 CREDBANK Overview

I provide a brief overview of the CREDBANK corpus, considering aspects such as agreement among raters and events that group into similar credibility distributions. Table 1 presents

Rate the credibility of tweets

Instructions

Tweets are short messages posted by users in a social media site called Twitter. An **Event** in Twitter is often depicted by tweets, mentioning real-world occurrences, which is considered to happen at a specific time and place, and which is of interest to others and not just to ones Twitter friends.

In this task, you will be presented with a group of tweets where **MOST** of the tweets are mentioning an **Event**. You need to rate the credibility level of the **Event**. In others how accurate do you think the event mentioned by the group of tweets are? You either need to be knowledgeable about the information stated in the tweets, or need to search for it online before you can make a reasonable credibility judgement. The qualification test you passed earlier provided you with instructions of how to do the task.

NOTE: Grading for bonus only occurs after all the assignments for the have a better view of the "majority" from multiple workers).

If your answers match the majority then you will be given a bonus

Please wait while the tweets get loaded ...

Task:

Event 1

Click to Open Twitter Search box

1

The Event can be categorized as : **Required**

You need a reason to support your choice.

- [-2]. Certainly Inaccurate
- [-1]. Probably Inaccurate
- [0]. Uncertain (Doubtful)
- [1]. Probably Accurate
- [2]. Certainly Accurate

3

Required

Please provide a reason behind your choice.

4

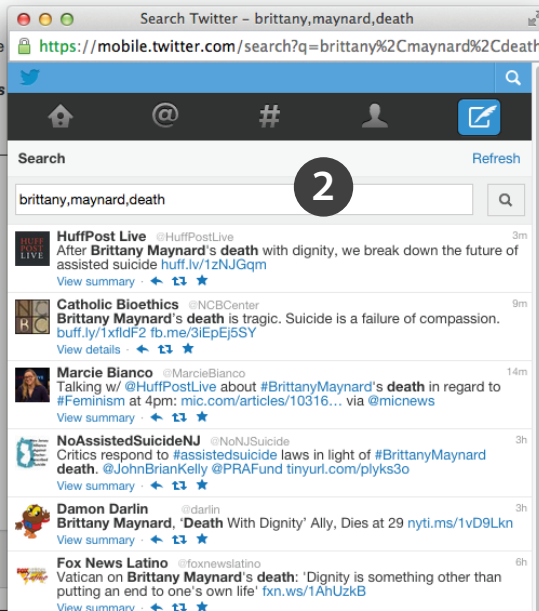


Figure 5: Turker interface for credibility assessment. The numbers correspond to a Turker's workflow. 1. Click the search box. 2. Read tweets from the pop-up Twitter search window. 3. Select one of the credibility scale options. 4. Provide a reason for the selection. Validation checks within the HIT ensure adherence to this workflow.

descriptive statistics of the corpus. Each set of items resulted from executing the steps of our workflow (see Figure 1).

3.2.1 Agreement among Raters

I used intraclass correlation (ICC) to quantify the extent of agreement among raters [171]. ICC is argued to be a better measure compared to chance corrected measures (e.g., Cohen and Fleiss Kappa) because unlike chance-corrected measures, ICC does not rely on the notion of perfect agreement. Instead, it measures the reliability of ratings by comparing the portion of variation in the data that is due to the item being rated and the variation that is due to raters. If the rater-induced variation exceeds the item-induced variation then the raters are said to have low-inter rater reliability. Moreover, ICC is flexible enough to adapt

Table 1: CREDBANK summary statistics.

	Counts
Time Duration of Collection	26 days
Streaming tweets	161,716,260
Topics annotated as Event/NonEvent	12,300
Event/NonEvent Annotations	123,000
Topics Agreed upon as Events	265
Credibility Annotations	7,950
Event specific tweets searched and collected	11,658,768

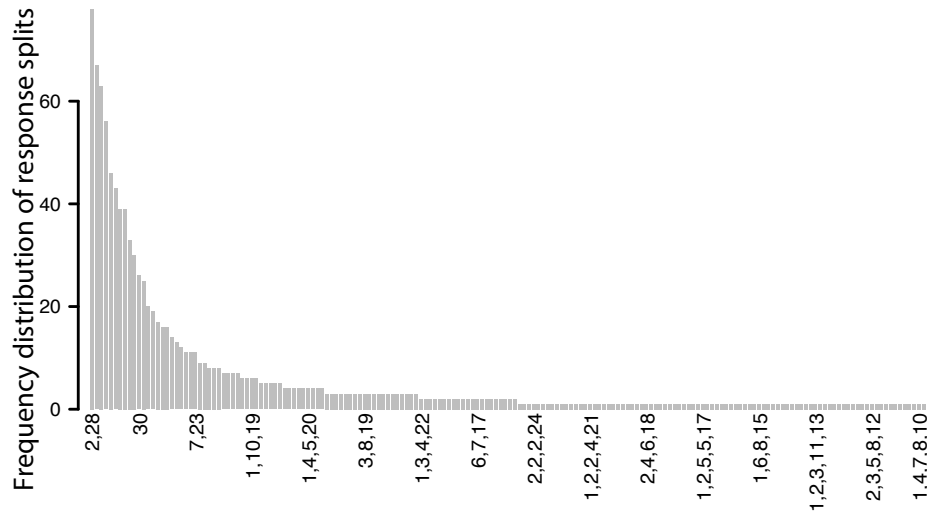


Figure 6: Frequencies of response distributions. The labels on the x-axis correspond to the splits in the annotation categories. For example, the bar at label ‘30’ correspond to the total number of events where all 30 Turkers agreed that the event is ‘Certainly True’. The bar at label ‘2,2,4,6,16’ corresponds to the count of events where the annotations are split five ways over the 5-point Likert credibility scale.

to different rater agreement study designs. In my study design I have a random sample of r raters rating each event. Hence, I use the Class 2 ICC measure [171] and obtain a fairly high ICC coefficient (Average Random Raters ICC = 0.77, 95% C.I. = [0.77, 0.81]) indicating high reliability of the collected annotations. A total of 1736 unique Turkers participated in the credibility annotation task.

3.2.2 Annotation Response distribution

How often did Turkers agree with one another on a credibility label? To explore this question I study the entire distribution of credibility annotations. I plot the frequency distribution of the splits in the annotation categories (see Figure 6). Hereafter I refer to these as *response splits*; the labels on the x-axis correspond to the different *response splits*. The label names are formed in the same order as the credibility scale presented to Turkers. For example, the label ‘2,7,21’ groups the items for which the ratings are split three ways—2 Turkers agreeing on ‘Uncertain’, 7 agreeing on ‘Probably True’ and 27 agreeing on ‘Certainly True’.

The long tail of the distribution suggests the difficulty associated in converging on a single credibility rating. In fact, there are a significant number of events where the response splits span across the 5-point credibility scale. For further examination of these cases, I first divide annotated events into two sets—those where more than 50% of Turkers agreed on a credibility label (*Majority set*) and those where there is no clear majority (*No Majority set*). Next, I provide closer examination of these groups.

3.2.3 Majority Agreement

I explore the *Majority set* by varying the majority agreement threshold and plotting the percentage of event annotations falling within that threshold. More than 95% of events had 50% Turkers agreeing on a single credibility label, ‘Certainly Accurate.’ Increasing the majority threshold results in rapid drop in the agreement percentages, with only 76.54% of events having 70% Turker agreement, while only 55% of events had 80% Turker agreement. All 30 Turkers agreed on only 2% of events being ‘Certainly Accurate’. In other words, considering moderate threshold of inter-rater agreement (70% majority threshold), an alarming 23.46% of events were not perceived to be credible. An important implication of this finding is the presence of Twitter events where credibility assessment did not converge on ‘Certainly Accurate,’ hinting at the presence of non-trivial percentages of inaccurate information in Twitter. Figure 7 summarizes these results.

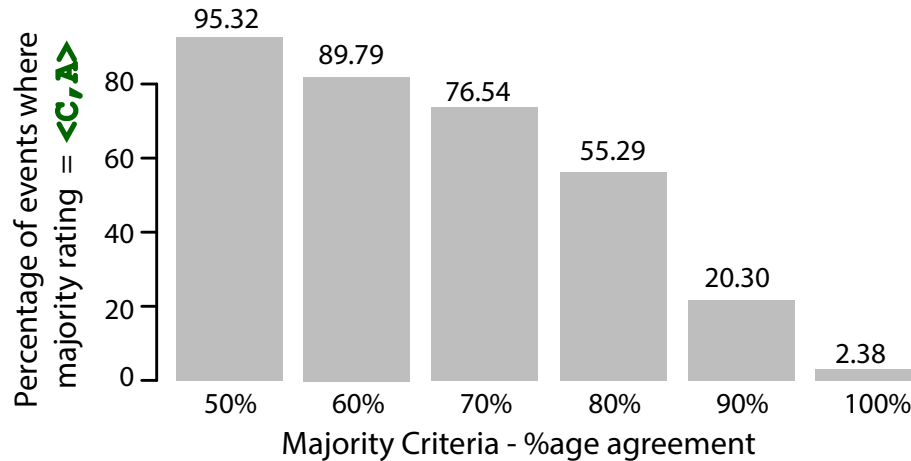


Figure 7: Percentage of events with majority Turker agreement. The majority agreement threshold is varied from 50% to 100% in steps of 10%.


3.2.4 No Majority Agreement

Next I examine cases where 50% of the Turkers did not converge on a single credibility score. In my dataset there are 49 such events, each with 30 annotations, resulting in a total of 1470 ratings. To compare the variations in these ratings I introduce the concept of a *credibility curve*—a histogram of rating frequencies at each credibility score. Are there different types of curves? Is it possible to group these based on their shape and magnitude? I turn to clustering techniques to look for meaningful groups based on shape and magnitude.

3.2.4.1 Credibility Clusters

In the absence of any prior hypothesis concerning the number of clusters, I apply hierarchical agglomerative clustering to group these credibility curves. The clustering process starts with each annotation item as a cluster and then merges pairs of clusters based on a similarity metric. I used the Euclidean distance similarity metric and Ward’s fusion strategy for merging. Ward’s method is a preferred strategy because it uses an analysis of variance approach and is very efficient [193]. A common concern with clustering is determining the validity of the cluster. I check cluster validity by using an *R* implementation of multi-scale bootstrap resampling: *pvclust*. *pvclust* performs hierarchical clustering, computes *p*-values

via multi-scale bootstrap analysis for each cluster and in the process returns clusters which are strongly supported by the data [184]. Applying *pvclust* on the *No Majority set* results in four clusters. I qualitatively compare them by plotting their corresponding *credibility curves*. The credibility curve of each cluster is a normalized plot of the rating counts at each credibility label for all events in that cluster. Figure 8 illustrates these curves and their associated event clusters. Here I focus mainly on the trends among these four groups and also highlight a few events in these clusters.

Step Curve  This group is the largest of all. The credibility ratings of the events in this group are spread all over the 5-point scale. The shape of the credibility curve further suggests the even split between the two categories: ‘Uncertain’ and ‘Probably Accurate’. A closer examination of these events reveals instances with high degree of speculation and uncertainty. For example, the topics *kobe, dwight, nigga* and *kobe, dwight, something* refer to the verbal altercation between the basketball players Kobe Bryant and Dwight Howard in a recent NBA game, followed by attempts to lip read the words exchanged². I find that the reasons provided by the Turkers also reflect this speculation.

“Although I see him mouthing what looks liek [*sic*] soft there is no audio and no confirmation from either person.”(*Tuker rating: [0] Uncertain*)

More seriously, other examples of events in this group refer to the Ebola pandemic. During the time when I collected data, Twitter witnessed a massive increase in conversations related to the Ebola virus along with several rumors about how the virus spreads, as well as false reports of potential cases of the disease³. Our corpus annotations capture these posts along with human judgments of credibility as this event was unfolding on Twitter. Although a third of the Turkers rated the event as accurate, the bump around the ‘Uncertain’ and ‘Probably True’ categories suggests the uncertainty associated with this event.

Low Shoulder Curve  Events in this group had very few ratings in the ‘Certainly

²http://espn.go.com/los-angeles/nba/story/_/id/11783332

³<http://time.com/3478452/ebola-Twitter/>

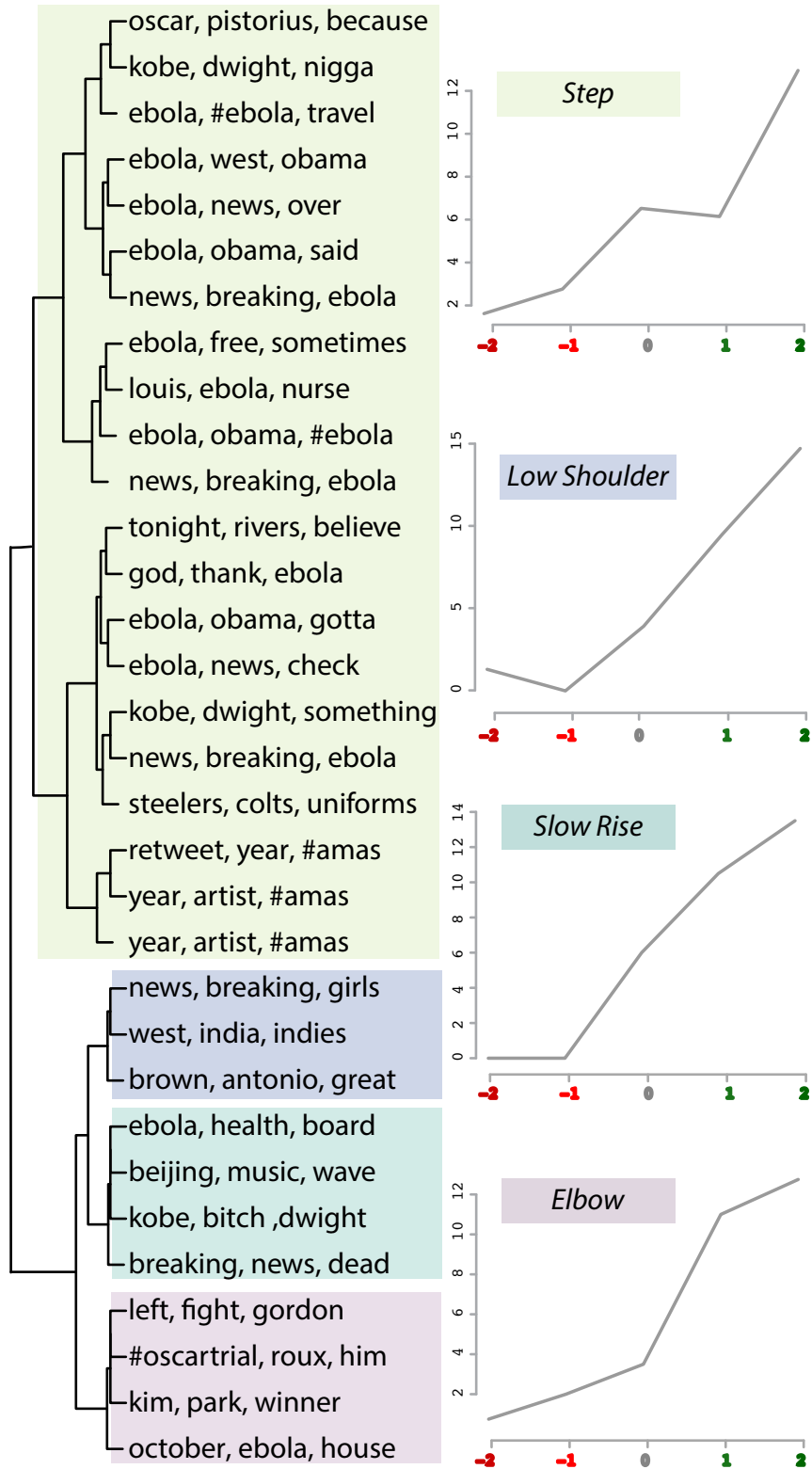


Figure 8: Cluster dendrograms for a sample of events and their corresponding average credibility curves.


Inaccurate' category, none in the 'Probably Inaccurate' category, while the other three categories had progressively increasing membership. Closer examination of the reasons entered by Turkers for rating as 'Certainly Inaccurate' revealed that sometimes when they were unable to find a coherent theme in the tweets they would pick this category.

“All more or less about the same thing with credible sources, but not on any specific event. Scattered reports with different related topics.” (*Turker rating: [-2] Certainly Inaccurate*)

However, there were only a few these instances. An interesting observation here are the different strategies used to assess event accuracy. Consider the event *news, breaking, girls*. It is about the kidnapping of 200 Nigerian school girls by an Islamist militant group. The event sparked a global outrage, with speculations of their release after a truce between the Nigerian government and the militant group⁴. Looking at the reasons provided by Turkers while rating this event, I see assessment strategies ranging from simply trusting the source of the information [102] to using suspicion as a cue for judgment [82].

“Credible sources in the tweets and confirmed by internet research.” (*Turker rating: [1] Probably Accurate*)

“219 Kidnapped Nigerian girls to be released, BBC reports "cautiously optimistic.” (*Turker rating: [1] Probably Accurate*)

Slow Rise Curve  This group had events with ratings spanning the positive polarity dimension including the 'Uncertain' category. Most reasons for the choice of 'Uncertain' were circulating conflicting reports about an event.

Conflicting stories on Ebola and what is happening. (*Turker rating: [0] Uncertain*)

I find that similar to the previous group, there were some instances where Turkers were unable to detect the coherent theme due to ambiguity of search terms. I found one such case

⁴<http://www.bbc.com/news/world-africa-29665165>

in our data set. This relates to the topic *breaking,news,dead*. The time during which these terms surfaced, it referred to the killing of sixteen people at a concert accident in Korea⁵. While the terms are fairly general, it was encouraging to see that most of the Turkers were able to relate it to the Korean concert accident. A few of the annotations, however, reflected the confusion arising from the generality of terms.

Elbow Curve ↗ Events in this group had credibility ratings spanning the entire 5-point scale. In this sense, the group has similar characteristics to the *Step Curve* group. Closer investigation reveals events marked with considerable speculations and dynamically evolving over time. For example, the event *#oscartrial, roux, him* in this group refers to the trial of Oscar Pistorius for the murder of his girlfriend⁶—a seven-month long, emotional trial which was accompanied by speculations regarding the prosecution.

3.3 Future Research Implications

My central goal in this thesis is to systematically study social media credibility. In moving towards that goal I had to first compile a dataset (CREDBANK) linking social media event streams with human credibility judgments in a systematic and comprehensive way. This is the first attempt I am aware of to do so at such a scale and in a naturalistic setting, that is in an environment which closely resembles the way a person would search Twitter for event information. I envision that this will be a useful data resource for the community to not just further the study of social media credibility, but to also enable a set of new research questions. For example, social scientists might explore what role does the mainstream media plays in online rumors; a data mining researcher might explore how the temporal patterns of rumors differ from highly credible information or study the interplay between highly disputed event and other less disputed ones occurring in the same time span; a health

⁵<http://www.aljazeera.com/news/asia-pacific/2014/10/deaths-reported-s-korea-concert-accident-20141017112748873969.html>

⁶<http://www.washingtontimes.com/news/2014/oct/17/pistorius-sentencing-final-arguments-begin>

researcher could investigate how folk theories of a new disease (the emergence of Ebola is captured in CREDBANK) diffuse through a population.

In Chapters 4 and 5, I unravel linguistic and temporal regularities of credibility by analyzing and modeling CREDBANK's data corpus. Here, I sketch additional possible future CREDBANK research directions.

Social and structural dynamics of events across credibility. What role do users of a social network play in spreading information varying in credibility level? How does audience size and the level of information credibility affect information propagation? Our corpus enables delving into these questions. Investigating the following and follower graphs of sets of user posts and their corresponding credibility ratings might be a first step in this direction. I had started down this road, but quickly released that unless we have access to Twitter's firehose API, it will take months, if not years to traverse the social graphs of millions of users. I tried to address this limitation by reaching out to Twitter's data partner GNIP, but soon realized that the data cost is beyond the scope of our current academic budgets. I hope that researchers with larger funding pools or institutions with firehose access can continue working in this direction.

What role the mainstream media play in online rumors? Studies have demonstrated social media's importance in news production, highlighting several instances where news surfaced in social media before mainstream media reports [129]. With this in mind, it seems very worthwhile to investigate the role played by mainstream media in the propagation of online misinformation. With the available user profile information in Twitter posts, CREDBANK allows unpacking of these questions.

Could credibility be modeled as a distribution? The long tail of our credibility rating distributions (Figure 6) suggests the nuances associated with finding a single unique credibility label for an item. Perhaps I need to rethink the widely held assumption of the existence of distinct, single-valued credibility labels. CREDBANK's large set of per-item credibility

ratings allows future work on probabilistic modeling of credibility.

What are the strategies used to evaluate credibility? Individuals use a wide variety of strategies when assessing information accuracy. Research focused on mental models have found that people often look for coherence in the story, as it is difficult to interpret a single piece of information in isolation [78]. I see echoes of this strategy in the reasons provided by CREDBANK’s Turkers. I think that this corpus allows a systematic study of strategies used for credibility assessment via CREDBANK’s rationales.

Studies in cognitive psychology have demonstrated the tendency of individuals to estimate the likelihood of an event “by the ease with which instances or associations come to mind” [188] (also known as *availability heuristics*). This might result in judgment biases, with people attaching more value to information they can easily recall, such as information that is more emotionally resonant or more recent [164]. Are there any repeated biases associated with human annotators’ credibility assessments?

Supplementary data. A byproduct of the corpus-building process are the event annotations—groups of tweets annotated as events or non-events, along with short keyword-based summarizations upon being judged as events. I envision one use of this data may be in event-extraction and summarization systems. Imagine an automatic system which needs to reason about whether a set of posts is coherent enough to be considered as an event, or a system which generates short summaries from a set of event-based posts. CREDBANK could provide valuable ground truth.

CHAPTER IV

STRATEGIES FOR OBTAINING QUALITY ANNOTATIONS FROM CROWD-WORKERS

In this chapter, I present a small but necessary detour from the overall effort of understanding social media credibility. In the previous chapter, I presented CREDBANK which is the backbone of this dissertation. Constructing CREDBANK hinged on answering the following question — *How can we obtain high quality annotations using crowd workers?* This chapter is an exhaustive attempt to answer this question.

The emergence of crowd-sourced 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 acquiring well-measured high quality judgments using an online workforce is often seen as a challenge [66, 72, 144, 166, 176]. This has led to scholarly work suggesting quality control measures to address the problem of noisy data [41, 85, 115, 166]. 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 following study addresses these questions. Building on some of the most promising strategies identified by prior work (e.g., [166]), I designed and conducted a large empirical study to compare the relative impacts and interactions of 34 intervention strategies. Specifically, I collected and analyzed 68,000 human annotations across more than 280 pairwise statistical comparisons for strategies related to worker screening and selection, interpretive convergence modeling, social motivations, financial incentives, and hybrid combinations. Further, I compare these

interactions against a range of representative subjective judgment-oriented coding activities of varying difficulty. The exhaustive experiments led to the following three principal contributions:

- I show 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, transient, untrained, anonymous workforce, I found that prescreening workers for requisite cognitive aptitudes, and providing basic training in collaborative qualitative coding methods results in better agreement and improved interpretive convergence of non-expert workers.
- I further demonstrate 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., [139, 166]).
- Finally, of particular importance for contemporary AMT researchers, I demonstrate 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 [35], AMT is not nearly the “Wild West” that it used to be.

4.1 Background and Related Work

4.1.1 Data Coding and Annotation

Qualitative Data Analysis (QDA)—that is, systematically analyzing non-numeric data such as interview transcripts, open-ended survey responses, field notes or observations, and a wide range of text documents, images, video, or audio data is generally a specialized skill most often acquired through formal training. Such skills are costly, both in terms of the financial demand required to obtain the skill-set (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, and a reduction in potential coding errors. What if we could quickly, inexpensively, and yet reliably obtain high-quality annotations from a massive, distributed, untrained, anonymous, transient labor force?

4.1.2 Crowdsourcing Qualitative Coding & Content Analysis

Crowd-sourced 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, prior work by Wang, Kraut, & Levine [192] asked workers to code discussion forum messages for whether they offered information or provided emotional support. Sorokin & Forsyth [176] had coders annotate images to locate people. Hutto & Gilbert [72] asked coders to annotate the intensity of sentiments in social media texts. Soni and colleagues [174] had workers mark the degree of factuality for statements reported by journalists and bloggers. Andre, Kittur, and Dow [7] asked workers to extract thematic categories for messages shared amid Wikipedians.

Clearly, crowd sourcing does enable quick, inexpensive content analysis and data coding at large scales (c.f., [7, 72, 174, 176, 192]). However, these types of QDA activities are often quite subjective in nature. As such, they are susceptible to conflicting interpretations, dissimilar rubrics used for 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 transience, and fast cash disbursements can entice the online labor workforce to trade speed for quality [41]. 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 [166]. Scholars using AMT must therefore carefully consider strategies for ensuring that the codes and annotations produced by non-experts are of high quality, that is, ensuring that the coding produced by anonymous workers is accurate and reliable [66, 72, 144, 166, 176]. Previous research suggests several quality control measures to tackle the problem of noisy data [7, 41, 71, 85, 95, 115, 166, 168]. 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, I designed and conducted 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. I discuss these strategies in greater detail later). Further, I compare these interactions against a range of coding activities that have varying degrees of subjective interpretation required.

4.1.3 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 inexperienced raters [76, 167, 173]. For example, Snow and colleagues [173] 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 [167] 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 data sets. Ipeirotis, Provost, & Wang [77] 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 inexperienced coders subsequent to data collection, I follow in the same vein as Shaw et al. [166] 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. [166], I include numerous other person-centered and process-centered strategies for facilitating high quality data coding from non-experts across a range of annotation tasks. I describe these strategies in the next section.

4.2 Strategies for Eliciting Quality Data

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

Challenge 1 – Undisclosed cognitive 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 [81]. 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.

Strategy 1 – Screen workers: 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%) [16, 114, 137]. 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 [72, 174].

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.

Strategy 2 – 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 [198]. 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 [155]. 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 and calibrates them to the nature of desired responses. This strategy helps improve inter-coder/inter-rater agreement, or interpretive convergence - i.e., the degree to which coders agree and remain consistent with their assignment of particular codes to particular data [155].

Challenge 3 – Existing financial incentives are oriented around minimizing 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 workers may attempt to maximize their earnings while minimizing their effort.

Strategy 3 – Financially incentivize workers to produce high-quality results: In an effort to incentivize carefully considered responses, rewarding high quality responses has shown to improve annotation accuracy [72, 166]. For every intervention strategy examined, I include both a non-incentivized and an incentivized group, and I confirm whether financial incentives continue to have significant impacts above and beyond those of a particular intervention strategy.

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.

Strategy 4 – Aggregating, iteratively filtering, 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 and accepting only the majority agreement from multiple independent workers [72, 183]. 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 [179]. Prior research has addressed this challenge by adding iterative steps to the basic parallel process of collecting multiple judgments [66, 110]. 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 [106, 114].

4.3 Our Tasks

In order to establish a framework of strategies for obtaining high quality labeled data, I 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 (I describe these in more depth in a moment).

Each of annotation task varied in the degree of subjective interpretation required. I deployed four HITs on AMT using a modified version of the NASA-TLX workload inventory scale to assess subjective judgment difficulty [65]. Response options ranged from “Very Low” to “Very High” on a seven-point scale. Each HIT asked 20 workers to perform the four qualitative coding tasks described below, and paid \$0.75 per HIT. To account for item effects, I used different content for each annotation task in each of the four HITs. Also, to account for ordering effects, I randomized the order in which the tasks were presented. Thus, I collected a total of 80 responses regarding the difficulty of each type of task, providing us with a range of tasks that vary in their underlying subjective judgment difficulty.



Figure 9: Example pictures for three of the five possible data coding/annotation categories.

TASK 1: People in Pictures (PP), median difficulty = 1: In this task, I 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 [36, 135]. I selected 50 images containing different numbers of people from the Creative Commons on Flickr. 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 – I determined ground truth at the time the image was selected from Flickr. I purposefully selected images based on a stratified sampling technique such that exactly ten pictures were chosen for each coding/annotation category.

TASK 2: Sentiment Analysis (SA), median difficulty = 2: In this task, I mimic a sentiment intensity rating annotation task similar to the one presented in [72] whereby I presented workers with short social media texts (tweets) and asked them to annotate the degree of positive or negative sentiment intensity of the text. I selected 50 random tweets from the public dataset provided by [72]; however, I 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 I can maintain consistent levels of chance for coding the correct annotations across all our subjective judgment tasks.



Figure 10: Example of the sentiment analysis annotation task.

Expert Annotation / Ground Truth – I derived ground truth from the validated “gold

standard” public dataset provided by [72], and adjusted by simple binning into a five point annotation scale (rather than the original nine point scale). One of the authors then manually verified each transformed sentiment ratings categorization into one of the five coding/annotation category options.

TASK 3: Word Intrusion (WI), median difficulty = 2: In this task, I mimic a human data annotation task that is devised to measure the semantic cohesiveness of computational topic models [27]. I presented workers with 50 “topics” (lists of words produced by a computational Latent Dirichlet Allocation (LDA) process [18]) 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 [27], I inserted an “intruder word” into each of the 50 LDA topics, and asked workers to identify the word that did not belong.



Figure 11: 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.

TASK 4: Credibility Assessment (CA), median difficulty = 3: In this task, I 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 [23, 124, 143]. To build a dataset of annotation items that closely resembles real-world information credibility needs, I first ensure that the dataset contains information sharing tweets, specifically



Figure 12: Example of a tweet along with the five credibility coding/annotation categories modeled according to existing work on credibility annotation categories [23,174].

those mentioning real world event occurrences [127]. To this end, I borrowed existing computational approaches to filter event specific tweets from the continuous 1% sample of tweets provided by the Twitter Streaming API [23, 98, 207].

Next, I recruited independent human annotators to decide whether a tweet was 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, I add the tweet as a potential candidate for credibility assessment. Next, the first author manually 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 [88]. I 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.

4.4 Conduct of the Experiments

A full factorial design to evaluate all strategies across all coding/annotation tasks results in combinatorial explosion, making a full factorial experiment intractable. As a work-around, I evaluate the strategies across tasks in stages. A total of 34 combinations were explored (see Table 1). I recruited non-expert content analysis data coders from Amazon Mechanical Turk, and employed a between-group experimental design to ensure that I 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, I asked workers to make coding/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 annotations were collected (50 items * 40 annotations * 34 intervention strategy combinations).

In the design of our HITs, I 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.

4.4.1 Comparative measures of correctness

I 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 of the crowd (i.e., the mode), the second counts the number of workers who match the mode of experts. I purposely choose mode over other measures of central tendency to establish a strictly conservative comparison metric which can be applied consistently across all comparisons.

4.4.2 Statistical Analysis

For all our experimental conditions I 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, I investigate the cell-wise residuals by performing all possible pairwise comparisons. Because simultaneous comparisons are prone to increased probability of Type 1 error, I 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 [110, 114].

4.5 Experiments

Next, I present two experiments. Briefly, 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, I compare the “winner” from Experiment 1 against more complex, process-oriented a priori strategies such as BTS, competition, and iteration.

4.5.1 Experiment 1 (Strategies 1-3, Tasks 1-3)

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

1. **Control condition, no bonus (Control NB):** Workers were presented with simple instructions for completing the data coding/annotation task. No workers were screened, trained, or offered a financial incentive for high-quality annotations. “NB” stands for No Bonus.
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 code/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, 34, 66, 134]. I 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. [95] 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 coding rubrics and heuristics. I 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

Table 2: Credibility classes and number of events in each class. The range of P_{ca} (proportion of annotations which are “Certainly Accurate”) for each class is also listed.

		Subjective Judgment 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	
		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		
Intervention	Control	✓	✓		✓	✓			✓	✓			
	Baseline	✓	✓		✓	✓			✓	✓			
	Screen	✓	✓		✓	✓			✓	✓			
	Train	✓	✓		✓	✓			✓	✓			
	Both (Screen+Train)	✓	✓		✓	✓			✓	✓			✓ ✓ ✓
	Iterative Filtering												

were qualified using the same task-specific demonstration and training techniques, 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).
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 2 summarizes the control and treatment conditions used for Experiment 1 (described above), and previews the test conditions for Experiment 2 (described later). Likewise, the χ^2 statistic is also highly significant when comparing worker annotations to an **expert**: χ^2 (df=9, N= 59,375) = 149.12, $p < 10^{-15}$. Furthermore, Table 3 shows that these significant differences are robust across three diverse types of qualitative data coding/annotation tasks. After seeing a statistically significant omnibus test, I perform post-hoc analyses of all pairwise comparisons using Bonferroni corrections for a more rigorous alpha criterion. Specifically, there are 45 multiple hypothesis tests, so I test statistical significance with

respect to for all paired comparisons. In other words, our between-group experimental study design supports 6 sets of 45 comparisons (i.e., 45 pairs) across 3 tasks and across 2 accuracy metrics, for a total of $45 \times 3 \times 2 = 270$ pairwise comparisons; and for all pairs, p-values must be less than 0.001 in order to be deemed statistically significant. Figure 13 depicts the percentage of correct annotations in each intervention strategy for each type of coding/annotation task, with indicators of the associated effect sizes for pairs with statistically significant differences.

4.5.2 Experiment 2 (Strategies S3–S4 in 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, I found that screening workers for task-specific aptitude and training them to use a standardized, consistent rubric for subjective judgments improves the quality of annotations. Thus I keep screening and training constant across the conditions of Experiment 2. Our Experiment 2 subjective judgment difficulty is even higher than that of the word intrusion task. Based on these observations, I repeat the **Screen + Train + (Bonus-M)** as a benchmark condition for Experiment 2.

4.5.3 Results from Experiment 1

Table 3 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, Table 3 shows that the χ^2 statistic related to the number of correct annotations when compared to the crowd is highly significant: χ^2 (df=9, N= 59,375) = 388.86, $p < 10^{-15}$. As test conditions, I then 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

Table 3: 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 (People in Pictures)	9	20,000	345.73	$<10^{-15}$
Worker-to-Expert	PP (People in Pictures)	9	20,000	46.66	$<10^{-15}$
Worker-to-Crowd	SA (Sentiment Analysis)	9	19,675	185.49	$<10^{-15}$
Worker-to-Expert	SA (Sentiment Analysis)	9	19,675	160.95	$<10^{-15}$
Worker-to-Crowd	WI (Word Intrusion)	9	19,700	90.74	$<10^{-15}$
Worker-to-Expert	WI (Word Intrusion)	9	19,700	59.82	$<10^{-15}$

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. [166]. BTS asks people to prospectively consider other’s responses to improve quality. Thus, in this intervention condition, I ask workers for their own individual responses, but I 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 [139, 166]).
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.

4.5.4 Results from Experiment 2

I compare the proportion of correct response using our two measures of correctness. In Experiment 2, I did not find any 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 I again conduct pairwise comparisons with Bonferroni correction. For our four experimental conditions, I conduct a total of comparisons, thus increasing the rigor of our alpha significance criterion to .

U find that across all conditions the winning strategy is the one in which workers are screened for cognitive aptitude, trained on task-specific qualitative annotation methods, 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 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 statistical impact of each strategy:

$$Screen + Train + Bonus_{Majority} > [Competition <> Iteration] > BTS$$

4.6 Discussion and Conclusions

I systematically compared the relative impacts of numerous a priori strategies for improving the quality of annotations from non-experts, and I checked their robustness across a variety of different content analysis coding and data annotation tasks. Here, I offer several reasons for focusing on a priori techniques, as opposed to complex statistical data cleaning techniques

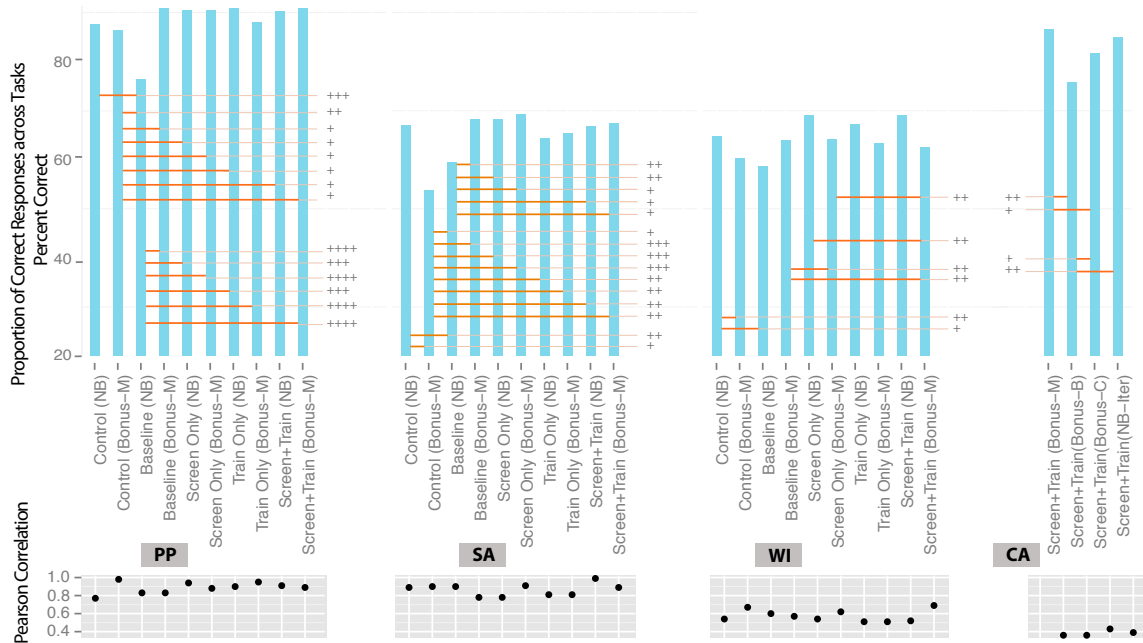


Figure 13: (Top panel) Proportion of correct responses across all tasks with respect to crowd. Pairwise comparisons which are statistically significant are shown with connecting lines (all p-values significant at 0.001 after Bonferroni correction). Effect sizes, as measured by Cramer’s V coefficient, are indicated using “+” symbols at four levels: +, ++, +++, and ++++ indicate a very weak effect Cramer’s V < 0.15, a weak effect Cramer’s V ∈ (0.15, 0.2], a moderate effect (Cramer’s V ∈ (0.2, 0.25]), and moderately strong Cramer’s V ∈ (0.25, 0.3]. (Bottom panel) Pearson correlation between expert and crowd annotations across all tasks.

performed post-collection. First, the value of a priori strategies are not as well explored, lending novelty to our contributions. Second, a priori person-oriented strategies emulate the procedures of sharing a common QDA codebook. Our results demonstrate the value of applying a well-established method for qualitative data coding to crowd-sourced data annotations by non-experts. Third, screening and training techniques have a onetime up-front cost which soon amortizes with increases in the size of datasets, so the techniques scale exceptionally well. Fourth, 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, post hoc procedures are of limited use for subjective oriented judgment tasks. Fifth, for time sensitive judgments (e.g., credibility decisions for rapidly unfolding events), simple a priori methods trump complex post hoc methods.

4.6.1 Crowd generated data annotations by non-experts can be reliable and of high quality

I find that our crowd-generated data annotations have relatively high data quality (in comparison to prior research, e.g., [166]), even though I use aggressive criteria for measuring accuracy; that is, I purposely choose exact matching with the mode over other potential measures (e.g., mean or median) as a strict metric for all comparisons. Further, the effects of interventions are generally robust across a range of representative QDA data annotation tasks of varying difficulty and with varying degrees of subjective interpretation required. For example, the top panel of Figure 13 shows the agreement between individual coders and the crowd provided ground truth. In every task, the agreement is well above chance (20% for all tasks).

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 for both the top and the bottom panels in Figure 13. When compared to an accepted

expert (Figure 13, bottom), I find generally high agreement between experts and the crowd-produced accuracy measure for the two easier subjective judgment tasks (i.e., judging the number of People in Pictures [PP] and Sentiment Analysis [SA] for tweets). The subjective judgment difficulty of the Credibility Assessment (CA) task is quite high, and so correlation of the crowd to the expert librarians is understandably decreased, though still moderately strong in the 0.4 to 0.45 range. (Note: the lower correlation for the Word Intrusion [WI] task is related more to the poor performance of the computational topic model algorithm as a non-human “expert” than the ability of the crowd to match that expert).

4.6.2 Person-oriented strategies trump process-oriented strategies for encouraging high-quality data coding

In general, I find that screening and training workers are successful strategies for improving data annotation quality. Financial incentives do not appear to help improve quality, except in the simplest baseline condition (impact becomes negligible when stronger person-centric strategies are used).

The insightful work from Shaw and colleagues [166] noted that process-centric strategies like BTS were effective at promoting better quality annotations. An 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 annotation quality. By verifying that a person has the requisite cognitive aptitude (knowledge, skill, or ability) necessary to perform a particular qualitative data annotation task, together with training workers on qualitative data coding expectations, we can significantly improve effectiveness above and beyond the effects of BTS (see Figure 13, top). Person-centric strategies, such as a prior screening for requisite aptitudes and prior training on task specific coding rules and heuristics, emulate the processes of personnel selection and sharing a common “codebook”. Qualitative scholars have been using such strategies for years to facilitate accurate and reliable data annotations among collaborative data coders [155]. By

applying these techniques to crowd-sourced non-expert workers, researchers are more likely to achieve greater degrees of interpretive convergence – and do so more quickly, with less variation (c.f., [52]) – because workers are thinking 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 more complex strategies actually do not perform as well as screening and training, why do more to get less quality?

The results of this study are not intended necessarily to be prescriptive. Even in a study this size, my focus was still on just a subset of potential intervention strategies, subjective judgment tasks, and various social and financial based incentives. Future work should directly compare the efficiency and effectiveness of a priori person-centric techniques to peer-centric methods (c.f., [71]) and more complex post hoc statistical consensus finding techniques (c.f., [168]). Nonetheless, the person-centric results reported in this paper help illustrate the value of applying established qualitative data analysis methods to crowd-sourced QDA coding by non-experts.

CHAPTER V

LINGUISTIC CONSTRUCTS OF CREDIBILITY

When people discuss a newsworthy topic via a social medium, like Twitter or Facebook, do they leave recognizable linguistic cues hinting at the story's underlying credibility? When other users interact or simply glance through the ongoing discussions, can they make reasonable judgments about the story's credibility? Can we design an algorithm, based only on linguistic signatures, such that it can infer the story's credibility and perform at par with human judgments of credibility? In this chapter, I describe a parsimonious language model which does exactly that. But first, consider the two Twitter events below:

EVENT: TRANSASIA PLANE CRASH

TWEET 1: Dashcams capture apparent footage of Taiwanese plane crash. Crash video may hold crucial clues.

TWEET 2: Hard to believe photos purporting to show #TransAsia plane crash in Taiwan are real. But maybe. Working to verify.

TWEET 3: If you haven't seen this plane crash video yet, it's chilling.

EVENT: GIANTS VS. ROYALS WORLD SERIES GAME

TWEET 1: Wow #Royals shut out the Giants 10-0. Bring on game 7, the atmosphere at The K will be insane. #WorldSeries

TWEET 2: The #Royals evened up the #WorldSeries in convincing fashion.

TWEET 3: @marisolchavez switching between the Spurs game and the Royals-Giants game. I agree! SO GOOD!!! #WorldSeries.

Of the two Twitter events, which would you consider to be highly credible and which less credible? The first event, about the TransAsia plane crash, contains expressions of skepticism such as *hard to believe*, *may hold*, hedging like *apparent footage*, *purporting to show*, *but*

maybe and anxiety in the word *chilling*. The second report, about a baseball game between the Kansas City Royals and the San Francisco Giants, exhibits high positive sentiment through *Wow*, *winning*, and *SO GOOD!!*, and general agreement, with the expressions *I agree* and *convincing*. As you may have guessed, the first event would most likely be perceived as less credible while the second one would be viewed as highly credible. This chapter is about linguistic constructs such as these and the credibility perceptions of social media event reportage that they signal.

While research in social media credibility has gained significant traction in recent years [24, 92, 145, 209], we still know very little, for example, about what types of words and phrases surround the credibility perceptions of rapidly unfolding social media events. Existing approaches to identifying credibility correlates of social media event reportage are based on retrospective investigation of popular events with known credibility levels, and thus suffer from dependent variable selection effects [187]. My analysis overcomes this sampling bias by adopting the CREDBANK corpus [121]. Merging the data from CREDBANK with linguistic scholarship, I built a statistical model to predict perceived credibility from language. My model takes 15 theoretically driven linguistic categories spread over more than 9,000 phrases as input, controls for 9 twitter specific variables, and applies penalized ordinal regression to show that several linguistic categories have significant predictive power. The most conservative accuracy measurement is 42.59%, while relaxing the measurement scheme brings the accuracy to 67.78%—significantly higher than a random baseline of 25%. This suggests that the language of social media event reportage has considerable predictive power in determining the perceived credibility level of Twitter events. In essence, our results show that the language used by millions of people on Twitter has considerable information about an event’s perceived credibility.

5.1 Method

To search for language cues indicating credibility, I employed data from the CREDBANK corpus [121]. My unit of analysis is an individual event and the perceived credibility level of its reportage on Twitter. My measurement of perceived credibility level is based on the number of annotators that rated the event’s reportage as “Certainly Accurate”. More formally, for each event, I found the proportion of annotations (P_{ca}) rating the reportage as “Certainly Accurate”.

$$P_{ca} = \frac{\text{“Certainly Accurate” ratings for an event}}{\text{Total ratings for that event}}$$

To have a reasonable comparison it is impractical to treat P_{ca} as a continuous variable and have a category corresponding to every value of P_{ca} . Hence, I placed P_{ca} into four classes that cover a range of values. I named the classes based on the perceived degree of accuracy of the event in that class. For example, events which were rated as “Certainly Accurate” by almost all annotators belonged to the “Perfect Credibility” class, with $0.9 \leq P_{ca} \leq 1$. Table 4 shows the credibility classes and the number of events in each class. Table 5 lists a representative sample of collected events in each class, their duration of collection, the credibility rating distribution of their corresponding reportages on a 5-point Likert scale, and the proportion (P_{ca}) of ratings marked as “Certainly Accurate”. To ensure that our P_{ca} based credibility classification was reasonable, I compared classes generated by the P_{ca} method to those obtained via a data-driven classification technique (refer to the next section for details). I found a high degree of agreement between the P_{ca} -based and data-driven classification approaches. I favor our proportion-based (P_{ca}) technique over data-driven approaches because the former is much more interpretable, readily generalizable and adaptable to domains other than Twitter, on which CREDBANK was constructed.

5.1.1 Validating credibility classification

This subsection details the steps taken to validate our four class credibility classification scheme based on the proportion of “Certainly Accurate” annotations for an event (P_{ca}). To

Table 4: Credibility classes and number of events in each class. The range of P_{ca} (proportion of annotations which are “Certainly Accurate”) for each class is also listed.

Credibility Class	P_{ca} range	Number of Events
Perfect Credibility	$0.9 \leq P_{ca} \leq 1.0$	421
High Credibility	$0.8 \leq P_{ca} < 0.9$	433
Moderate Credibility	$0.6 \leq P_{ca} < 0.8$	414
Low Credibility	$0.0 \leq P_{ca} < 0.6$	109

ensure that our P_{ca} based credibility classification is a reasonable classification, I compare classes generated by the P_{ca} method against those obtained via data-driven classification.

5.1.1.1 Generating data-driven credibility classes

I used hierarchical agglomerative clustering (HAC) [112] to generate data-driven classes of the credibility rating distributions. HAC is a bottom-up clustering approach which starts with each observation in its own cluster followed by merging pairs of clusters based on a similarity metric. In the absence of a prior hypothesis regarding the number of clusters, HAC is the preferred clustering method. HAC-based clustering approach groups the events based on the shape of their credibility curves on the 5-point Likert scale. Such shape based clustering approach has been used in prior work to cluster based on the shape of popularity peaks [33, 203]. I used the Euclidean distance similarity metric and Ward’s fusion strategy for merging [193]. The choice of this strategy minimizes the within-cluster variance thus maximizing within-group similarity [193]. Figure 14 shows the resulting dendrogram from hierarchical clustering. The boxes correspond to the credibility groups when the dendrogram is cut into four clusters.

5.1.1.2 Comparing P_{ca} based classes to HAC-based classes

Is the P_{ca} based credibility classification a close approximation of the HAC based classification? Essentially, I need a metric to compare two clusterings of the same dataset. In other words, I need to measure how often both clustering methods classify the same set of

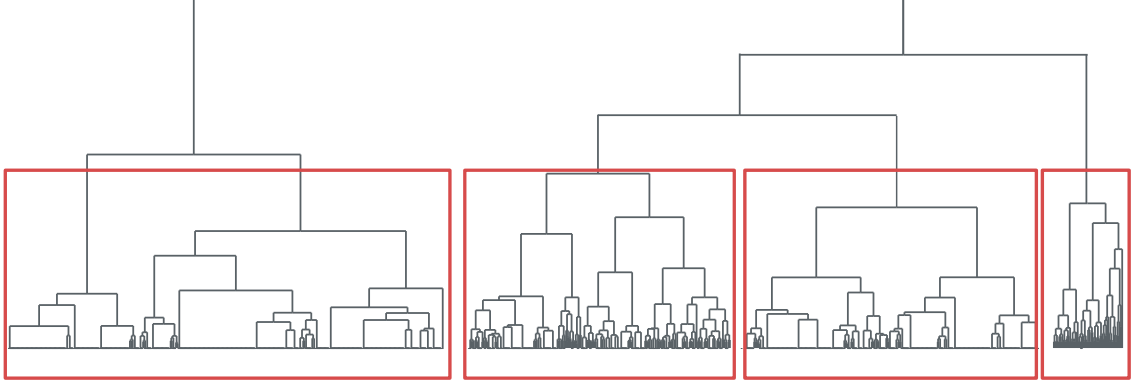


Figure 14: Dendrogram from hierarchical clustering of the events from CREDBANK. The boxes show the four clusters.

observations as members of the same cluster. I borrow a technique proposed by Tibshirani et al. [48]. Let $P_{clust} = \{x_{1c_1}, x_{2c_1}, x_{3c_2}, \dots, x_{nc_4}\}$ denote the cluster labels from P_{ca} based classification and $H_{clust} = \{x_{1h_1}, x_{2h_3}, x_{3h_4}, \dots, x_{nh_3}\}$ the labels from HAC-based classification of the same dataset D of n observations. Here, x_{ic_j} denotes that the i^{th} observation belongs to cluster c_j as per the P_{ca} classification and x_{ih_j} denotes that the i^{th} observation belongs to cluster h_j as per the HAC classification. I see that x_{1c_1} and x_{2c_1} belong to the same cluster. Such pairs are called “co-members”. While (x_{1c_1}, x_{2c_1}) are co-members as per P_{ca} classification, (x_{2h_3}, x_{nh_3}) are co-members from HAC classification. For each clustering method, I first compute all pairwise co-membership of all pairs of observations belonging to the same cluster. Next, I measure agreement between the clustering methods by computing the Rand similarity coefficient from the co-memberships as follows:

$$R = \frac{N_{11} + N_{00}}{N_{11} + N_{10} + N_{01} + N_{00}}$$

N_{11} : the number of observation pairs where both are co-members in both clustering methods.

N_{10} : the number of observation pairs where the observations are co-members in the first clustering method, but not in the second.

N_{01} : the number of observation pairs where the observations are co-members in the second clustering method, but not in the first.

N_{00} : the number of observation pairs where neither pair is co-member in either clustering

method.

Rand similarity coefficients range between 0 and 1, with 1 corresponding to perfect agreement between the two clustering methods. I obtain a fairly high R of 0.774 denoting high agreement between our P_{ca} based and HAC-based clustering approaches. I favor our proportion-based (P_{ca}) clustering technique over data-driven approaches because the former is much more interpretable and readily generalizable and adaptable to domains other than Twitter on which CREDBANK was constructed.

5.1.2 Response Variable: Dependent Measure

With each event from CREDBANK as our unit of analysis, our dependent variable is an ordinal response variable representing the credibility level of the event from “Low” to “Perfect” (a ranked category with “Low” < “Medium” < “High” < “Perfect”). I chose an ordinal representation of perceived credibility level for two main reasons. Firstly, representing credibility perceptions with the continuous variable P_{ca} instead of a few representative categories would add overhead to the interpretability of results. Secondly, the literature has not yet resolved the issue of representation of event credibility perceptions. The closest reports are from linguists studying veridicality, with some favoring categorical representation and assigning a single majority annotator agreement to each item [161] and others advocating for probabilistic modeling of differing annotator judgments so as to capture their inherent uncertainty [38]. By selecting a proportion-based (P_{ca}) ordinal scale, I achieved a compromise between the two extremes. Rather than a single majority agreement category, P_{ca} captures the extent of disagreement with the “Certainly Accurate” rating.

5.1.3 Predictive Variables: Linguistic Measures

To detect linguistic strategies corresponding to credibility assessment, I compiled several language-based measures after reviewing the principles underlying factuality judgments and veracity assessments [38, 80, 83, 162]. Building on lexical and computational insights, I identified 15 linguistic measures as potential predictors of perceived credibility level. Using

standard methods from computational linguistics, I incorporated these measures as features in our statistical model (discussed shortly). Specifically, I used specialized lexicons designed to operationalize language-based measures. Below I justify our choice of each measure as a potential credibility marker.

Modality: Modality is an expression of an individual’s “subjective attitude” [22] and “psychological stance” [120] towards a proposition or claim. It signals an individual’s level of commitment to the claim. While, words like *should* and *sure* denote assertion of a claim, *possibly* and *may* express speculations. Past research on certainty assessment has demonstrated the importance of such modal words [38,153]. Investigation of the distribution of *weak* and *strong* modality in veridicality assessments showed that *weak* modals *can*, *could* and *may* strongly correlate with the “possible” veridicality judgment category, while *strong* modals like *must*, *will* and *would* were evenly distributed across categories. Inspired by past research, I measured the modality expressed in an event’s reportage by using Sauri et al’s list of modal words [159], which have been successfully used in prior research on veridicality assessment [37, 162, 175]. By counting the occurrences of each modal word in an event reportage, I incorporated them as input features of our statistical model. I followed the same technique for other lexicon-based measures.

Subjectivity: Subjectivity is used to express opinions and evaluations [11, 197]. Hence, detecting the presence of subjectivity can differentiate opinions from factual information (often called objective reporting) [11, 195, 197]. Prior research has shown that knowledge of subjective language can be useful in analyzing objectivity in news reporting [195] and in recognizing certainty in textual information [153]. Drawing on these prior works, I hypothesized that subjectivity can provide meaningful signals for credibility assessment and used OpinionFinder’s subjectivity lexicon comprising 8,222 words [200].

Hedges: Hedges refer to terms “whose job is to make things more or less fuzzy” [93]. They are often used to express lack of commitment to the truth value of a claim or to display

skepticism and caution [73]. People who are uncertain about a topic tend to use such tentative language [185]. Work on certainty categorization in newspaper articles found that hedges were used to classify statements into low or moderate levels of certainty [153], thus demonstrating the intrinsic connection between hedges and expressions of certainty – a concept closely related to credibility assessment. Hence, I included hedges as potential credibility markers of an event’s reportage. To measure hedges, I used two sets of lexicons signaling tentative language: 1) list of hedge words from Hyland [75] and 2) tentative words from the LIWC dictionary [185].

Evidentiality: Evidentials are recognized as a means of expressing the degree of reliability of reported information [15, 153]. These are verbs like *claim, suggest, think*, nouns like *promise, hope, love*, adverbs such as *supposedly, allegedly* and adjectives like *ready, eager, able*. They qualify the factuality information of an event [153, 159]. Thus the choice of these attributive predicates can express the level of commitment in the reported information [15], or indicate a speaker’s evidential stance or even express the level of factuality in events [159]. Evidentials can be used to report (e.g. *say, tell*), express knowledge (e.g., *know, discover, learn*), convey belief & opinion (e.g., *suggest, guess, believe*) or show psychological reaction (e.g., *regret*). Such predicates can be used to emphasize a claim made in an information snippet or evade from making any strong claims, thus implicitly lowering the credibility signaling of the expressed information [153]. Recent studies have shown that evidentiality predicates can affect credibility perceptions of quoted content in journalistic tweets [175]. These manifestations of evidentiality prompted us to add them to the list of potential credibility markers.

Negation: Negation is used to express negative contexts. Social psychologists have shown that individuals who have truly witnessed an event can discuss exactly what did and did not happen, thereby resulting in higher usage of distinction markers such as negations like *no, neither, non* [64]. Thus negations might be associated with higher levels of perceived

credibility. Other studies on event veridicality have also used negations as features for veridicality assessment of news events [37]. Hence I include negation as a potential credibility marker. I measure it by using a lexicon of negation particles from the De Facto lexicon—a factuality profiler for event mentions in texts [159].

Exclusions and Conjunctions: Both exclusions and conjunctions are components for reasoning [185]. While exclusion words like *but, either, except* are useful in determining if something belongs to a category [185], conjunctions are used to join thoughts together. Prior research has demonstrated that an increased usage of exclusion words is associated with individuals telling the truth [64, 128]. Thus exclusions might be associated with positive polarity of credibility. On the other hand, conjunctions are useful for creating a coherent narrative. Hypothesizing that a coherent narrative can be associated with higher levels of credibility, I employed LIWC’s list of “exclusion” and “conjunction” words to incorporate features corresponding to these language markers.

Anxiety: Small scale laboratory and field research studies have shown anxiety to be a key variable in rumor generation and transmission [20, 151]. Since apprehensive statements typically manifest anxiety in the context of information transmission [19], I measured anxiety using LIWC’s list of anxiety words.

Positive and Negative Emotion: Moments of uncertainty are often marked with statements containing negative valence expressions. This aligns with work on rumor discourse where negative emotion statements were found to accompany undesirable events [19, 20]. To measure the extent of emotions expressed in event specific tweets, I used LIWC’s comprehensive list of positive and negative emotion words [185].

Boosters and Capitalization: Boosters are expressions of assertiveness. Words like *establish, clearly, certainly* are used to express the strength of a claim and the certainty of expected outcomes [74]. Hypothesizing that booster words can be useful credibility markers, I used the list of “booster” words compiled by Hyland [75] and the list of “certainty” words listed

in the LIWC dictionary [185] to incorporate features corresponding to booster markers in our model.

Like boosters, individuals often use capitalization as a way of emphasizing expressions. To measure capitalization, I computed the number of capitalized terms in an event's tweets.

Quotation: Quotations serve as a reliable indicator for veridicality assessment in newswire documents, with quoted content mostly correlating with the "Uncertain" category [38]. More recent research has shown that both linguistic and extra-linguistic factors influence certainty perceptions of quoted content in social media platforms such as Twitter [175]. Based on these studies, I hypothesized that quotation can be a potential indicator of the credibility levels associated with a social media event's reportage. By counting the occurrence of quoted content, I mapped this predictor onto its corresponding feature in the statistical model.

Questions: Posing questions to social media connections is a common practice and serves the purpose of satisfying information needs, advertising current interests and activities, or creating social awareness [125]. Linguists have found that question asking is a key strategy for dialogue involvement, increasing engagement and encouraging the reader to share the curiosity of the writer and his reported point of view [74, 75]. In a parallel line of work, social psychologists studying people's communicative styles during rumor transmission observed that some people might act as "investigators" asking lots of questions and seeking information [19, 20, 169]. These studies suggest the importance of asking questions in the face of uncertainty. Hence, I propose inclusion of questions as a potential indicator of perceived credibility level. I computed this measure by counting the number of question marks present in the tweets corresponding to an event stream.

Hashtags: Hashtags are twitter specific features which have been shown to serve as useful signals for identifying rumors [24]. Hence I include the count of hashtag terms in tweets associated with an event as a potential credibility marker.

5.1.4 Predictive Variables: Controls

I include the following nine variables as controls:

1. Number of original tweets, retweets, replies
2. Average length of original tweets, retweets, replies
3. Number of words in original tweets, retweets, replies

Including these variables in our model allows us to control for the effect of content popularity – trending events generating large number of tweets, replies and retweets.

5.1.4.1 Model Limitations

Like any other statistical model building technique, a limitation with our approach is that I might be missing potential confounding variables. For example, the author of the message or the message source may have an effect on credibility perceptions. While it is impractical to have a complete coverage of all potential confounds, many of these variables is perhaps implicitly manifested as language. Consider the scenario where you judge the message author's credibility on how she reports the event or when you assess different sources based on how contradictory their views are on a particular event.

5.1.5 Statistical Technique

My goal is to understand the linguistic strategies that affect the perceived credibility level of an event's reportage as it unfolds on social media. With perceived credibility level being a rank ordered dependent variable, I treat this problem as an ordered logistic regression problem. Our regression model takes linguistic features computed from all tweets posted for an event as input variables and outputs the four-level ordered outcome variable – *credibility class*. Table 6 outlines the control features, non-lexicon based and lexicon-based features along with the size of each lexicon. Certain phrases were found to be present in multiple lexicons. For example, the word *possibly* is present in both *subjectivity* and *hedge* dictionaries. To prevent double counting of features I included phrases spanning multiple dictionaries once under the *Mixed* lexicon category. There were 111 such overlapping phrases and the

mixed category comprised only 1.14% of the total feature set.

While regression performs best when input features to the model are independent from one another, phrase collinearity is a common property in natural language expressions. For example, phrases like *no doubt* and *undoubtedly* (both of which are present in our lexicon-based feature set) might frequently co-occur in tweets related to a definitive event. Moreover, phrase datasets can be highly sparse. Hence I used a penalized version of ordered logistic regression which handles the multi-collinearity and sparsity problem. It is also well-suited for scenarios where the number of input features is large relative to the size of the data sample. For example, our feature set comprises over 9,000 linguistic phrases while our data sample covers 1,377 events. This regression technique has also been widely used in identifying the power hierarchy in an email corpus [51], family relationships from Facebook messages [21] and in mapping sociocultural identity in tweets [43].

This regression technique has a parameter α (with $0 \leq \alpha \leq 1$) which determines the distribution of weights among the predictive variables. When $\alpha = 0$ (as in ridge regression), all correlated terms are included with coefficient weights shrunk towards each other, while $\alpha = 1$ includes only one representative term per correlated cluster with other coefficients set to zero. After testing our model's performance with varying levels of $\alpha \in [0, 0.1, 0.5, 1]$, I selected a parsimonious model with $\alpha = 1$. I used the `glmnetcr`¹ implementation from the R package, which predicts an ordinal response variable while addressing issues of sparsity, collinearity and large feature size relative to data sample size.

As our first step in building our statistical model, I included only control variables so as to measure their explanatory power. Next, I included all 15 linguistic categories (a total of 9,663 linguistic features). This means that any predictive power assigned to the linguistic features comes after taking into account the explanatory power of the controls. The top half of Table 7 outlines our iterative model building process. I first added features corresponding to all the original tweets in our dataset. For example, for a feature phrase such as *wow* from

¹<https://cran.r-project.org/web/packages/glmnetcr>

our positive emotion lexicon, I counted its cumulative number of occurrences in all original tweets associated with an event. Imagine a feature matrix with rows corresponding to an event and columns corresponding to linguistic features or controls. The values in each cell then represents the raw count of occurrences of the feature in an event’s original tweets. I also tested our feature space with normalized counts, logs of normalized counts, tf-idf (term frequency-inverse document frequency) and logs of tf-idf based counts but did not detect any significant improvements in model performance. Therefore, I adopted the simplest representation – raw counts of linguistic features – as our model’s independent variables.

Our next phase involved repeating feature expansion with all the reply tweets. Thus, the model’s independent variables consisted of cumulative occurrences of features within replies to original tweets associated with an event. For our reply tweet model I included all linguistic measures except *subjectivity*. I made this choice so as to retain only those language features which captured reactions present in the user’s replies. As subjectivity is primarily used to express opinions, it is more meaningful in the context of an original post. Furthermore, prior work has shown that reactions and inquiring tweets carry useful signals in assessing the certainty of information [209]. Our decision to explore feature phrases in original posts and replies differently was based on the intuition that people use different mechanisms while posting original content than when reacting to already-posted content through replies. By treating these separately, our goal was to capture these differences in linguistic tactics. I did not repeat the process for retweets because retweets essentially re-iterate what the original poster said. Instead, I simply add retweet count, number of words and average length of retweets to act as control variables to our model.

5.2 Results

5.2.1 Model Fit Comparison

I calculated the goodness of fit of our language model by comparing the model’s deviance against that of the Controls-Only model. Comparing with the Controls-Only model instead

of the Null model allowed us to capture the relative predictive power of the linguistic measures in contrast to the control variables. Deviance is analogous to the R^2 statistic of linear regression models and is related to a model's log-likelihood. It measures the model's fit to the data with lower values denoting a better fit, and difference in deviances approximately following a χ^2 distribution. While the Null model deviance was 3,937.37, addition of controls reduced the deviance to 3,499.54. The Controls-Only model explained 11% of the variability in the data and had significant explanatory power: $\chi^2(13, N=1,377) = 3937.37 - 3499.54 = 437.84, p < 10^{-15}$.

Adding linguistic measures corresponding to only the original tweets resulted in further drop in deviance, and our "Original Tweets + Controls" model explained 64.52% of the variability observed in the data. I also observed a significant reduction in deviance when I added features corresponding to only the replies. The deviance of our "Reply Tweets + Controls" model was 1,603.30 and the model accounts for 59.28% of the variance observed in the data. The model with the highest explanatory power incorporated a combination of linguistic measures corresponding to both original and reply tweets. The resulting omnibus model has a deviance of 1,181.65 with significant explanatory power: $\chi^2(1234, N=1,377) = 3,937.37 - 1,181.65 = 2,755.22, p < 10^{-15}$. It explains 69.99% of the variability in the data. From this point on, I term this omnibus model as our language classifier and report its accuracy in the next section.

The bottom half of Table 7 also lists model fits per linguistic class measure, i.e., how the model performed when I added features corresponding to a single linguistic category while keeping the control variables constant. Examining each model separately allowed us to compare the explanatory power of the different feature categories. I found that subjectivity has the highest explanatory power, followed by positive and negative emotion categories. The mixed category came next, followed by anxiety, booster and hedges. Figure 15 maps out the predictive power found for the linguistic categories and lists top representative positive and negative β weights per category. Phrases with positive β predicted an event to have

high perceived credibility. Conversely negative β s were indicative of an event with lower perceived credibility. The thickness of the arcs in Figure 15 is proportional to the deviance explained by each of the linguistic categories in their respective standalone models. Each arcs' degree of color saturation is based on the difference in the absolute values of the positive and negative β coefficients. I observe that color saturation inverts for original and reply tweets along a few linguistic categories, such as booster, hedges, anxiety and the emotion categories. I interpret these results in our Discussion section.

How did our control variables perform? I found that the only control variables with non-zero positive β weights were: *average retweet length* ($\beta = 0.25$), *average reply length* ($\beta = 0.18$). Controls with non-zero negative β s were: *number of retweets* ($\beta = -0.27$), *average original tweet length* ($\beta = -0.14$), *number of words in retweets* ($\beta = -0.02$).

How did our non-lexicon based features perform? Non-lexicon based features include variables such as: fraction of *capitalized* terms, *questions*, *quotations* and proportion of *hashtags*. I found that, with the exception of proportion of quotations in original tweets ($\beta = -0.097$), the non-lexicon based features lacked predictive power ($\beta = 0$).

5.2.2 Model Accuracy

I computed the performance of our language classifier according to four accuracy measurement schemes; the next section contains the mathematical implementation of the metrics. Prediction accuracy of each scheme is computed via stratified 10-fold cross validation on a 75/25 train/test split. Stratification was done to ensure that the proportion of the four credibility classes in each data fold is representative of the proportion in the entire dataset. Table 8 displays performance comparisons.

Unweighted Accuracy: This scheme represents the most conservative approach for measuring our model performance since it ignores the partial ordering present among the credibility classes. Model performance was assessed based on whether the predicted credibility class label for an event exactly matches the true label.

Level-1 Weight_{0.25} Accuracy: The unweighted accuracy measurement treated all errors equally by penalizing every misclassification. However, since credibility classes are ordered with "Low" < "Medium" < "High" < "Perfect", not all misclassifications are equally serious. Hence, our weighted accuracy schemes relaxed our penalizing criteria and rewarded partial credit for certain misclassifications. In the Level-1 Weight_{0.25} accuracy, a partial credit of 0.25 was rewarded if the classifier mispredicted the credibility class by one level (for example: classifier predicted "High" when the true credibility class is "Perfect"). Table 9(b) displays the corresponding credit matrix.

Level-1 Weight_{0.5} Accuracy: Here a partial credit of 0.5 was rewarded if the classifier prediction was incorrect by one level. The credit matrix corresponds to the one shown in Table 9(b), but 0.25 replaced with 0.5.

Level-2 Weight_{0.25,0.5} Accuracy: This is our most lenient classifier which rewarded a partial credit of 0.5 for mis-classification by one level and a partial credit of 0.25 for mis-classifications by two levels (Table 9(c)).

I compared the performance of our language classifier against two baseline classifiers: 1) Random-Guess baseline and 2) Random-Weighted-Guess baseline. In the random guess classifier, every credibility class had an equal probability of being selected. Hence the classifier randomly guessed and predicted any of the four possible credibility categories. On the other hand, the predictions of random-weighted guess classifier were based on the proportion of instances that belonged to each credibility class in our dataset. I opted for the random guess baseline classifiers over a choose-most-frequent-class baseline so as to illustrate a sensible baseline performance for each credibility category. I performed McNemar's test of significance to compare the accuracy of our language classifier with that of the baseline. McNemar's test, which assessed whether the proportion of correct and incorrect classifications in the two systems are significantly different, indicated that even with the most conservative approach employing an unweighted credit matrix, our language

a confusion matrix with actual and predicted class instances mapped along the rows and columns of the matrix respectively. Accuracy is then measured as the number of agreements between the predicted and true classes. Agreements are captured along the diagonal of the matrix, while off-diagonals represent mis-classifications.

$$Accuracy = \sum_{a=1}^K \sum_{r=1}^K \frac{x_{a,r}}{n} * w_{a,r} \quad (1)$$

where $x_{a,r}$ =number of instances from the a^{th} actual class predicted as being from r^{th} class, n = total number of instances classified, $w_{a,r}$ =credit for correct/incorrect classification. For naive accuracy, the credits are drawn from an unweighted confusion matrix corresponding to Table 9(a). The diagonals represent agreement between actual and predicted classes, while off-diagonals correspond to different mis-classifications. All off-diagonal elements for the naive accuracy are 0 indicating that there is no credit for any mis-classification. Hence naive accuracy measures the proportion of instances along the diagonal of a confusion matrix.

However, an ordinal classification task, such as ours, is a form of multi-class classification where there is an inherent order between the classes, but there is no meaningful numeric difference between them. Naive accuracy measure for evaluating ordinal classification models suffer from an important shortcoming – it ignores the order and penalizes for every misclassification. Hence, following an established approach by Cohen et al. [30], I employ an alternative measure defined directly in the confusion matrix. Table 9(b) and (c) displays our additional weighted confusion matrices. The off-diagonals of these matrices can be in the range of $(0, \dots, 1]$. As the values increase towards 1, the corresponding mis-classification is considered decreasingly serious. A value of 1 means that the two classes are considered identical for accuracy assessment. These additional weighted matrices allow us to capture how much the ordinal model diverges from the ideal prediction.

5.4 Discussion

According to our findings, the top predictive linguistic features associated with higher perceived credibility mostly comprise linguistic measures. The only control variable in the

top 100 predictors of high credibility scores was *average retweet length* ($\beta = 0.25$), while *average reply length* ($\beta = 0.18$) fell within the top 200 positive predictors. Similarly, top predictive features of lower perceived credibility scores were phrases from our language categories. *Number of retweets* ($\beta = -0.27$) was the only control in the topmost 50 predictors of low perceived credibility scores. This indicates that while higher number of retweets were correlated with lower credibility scores, retweets and replies with longer message lengths were associated with higher credibility scores. An explanation of this could be that longer message length of retweets and replies denote more information and greater reasoning, leading to higher perceived credibility. On the other hand, higher number of retweets (marker of lower perceived credibility score) might represent an attempt to elicit collective reasoning or ascertain situational awareness during times of crisis and uncertainty [157].

Among non-lexicon based features, fraction of quotations in original tweets was negatively correlated with credibility ($\beta = -0.097$). Indeed, a key pragmatic goal of messages with quoted content is to convey uncertainty and provenance of information while refraining from taking complete accountability of the message's claim [175].

Taking a closer look at our most predictive words in each lexicon, I found a striking view of words and phrases signaling perceived credibility level of social media event reportages (see Figure 15 for an overview). Table 10 and 11 list the top predictive words in each linguistic category. Below I present the relation of different linguistic measures to perceived levels of credibility.

Subjectivity: Our results show that subjectivity in the original tweets had substantial predictive power. As is perhaps to be expected, subjective words indicating perfection (*immaculate*_[$\beta=0.61$], *precise*_[$\beta=0.43$], *close*_[$\beta=0.45$]) and agreement (*unanimous*_[$\beta=0.12$], *reliability*_[$\beta=0.11$]) were correlated with high levels of credibility. Subjective phrases suggesting newness (*unique*_[$\beta=0.75$], *distinctive*_[$\beta=0.082$]) or signaling a state of aI and wonder (*vibrant*_[$\beta=0.85$], *amazement*_[$\beta=0.67$], *charismatic*_[$\beta=0.08$], *brilliant*_[$\beta=0.08$], *awed*_[$\beta=0.07$], *bright*_[$\beta=0.07$], *miraculously*_[$\beta=0.06$], *radiant*_[$\beta=0.06$]) were also associated with higher perceived credibility

levels. This suggests that when a new piece of information unfolds in social media or when the information is surprising and sufficiently awe-inspiring, people tend to perceive it as credible. Perhaps the newness of the information contributes to a paucity of detail to assess. While linguistic markers are efficient in determining the perceived credibility level of *newness*, temporal or structural signals can only be utilized after the information has circulated for a while. Additional subjective words associated with higher levels of perceived credibility hinted at the existence of complex, convoluted phenomena (*inexplicable*_[$\beta=0.61$], *intricate*_[$\beta=0.69$], *strangely*_[$\beta=0.07$]). Social psychologists argue that when faced with complex, difficult to explain phenomenon, individuals often take the “cognitive shortcut” of believing the phenomenon instead of assessing and analyzing it [189].

I also found that subjective words associated with narratives of trauma, fear, and anxiety were associated with higher perceived levels of credibility. Such words are, for example, *darn*_[$\beta=0.54$], *mortified*_[$\beta=0.35$], *mishap*_[$\beta=0.32$], *calamity*_[$\beta=0.30$], *catastrophic*_[$\beta=0.30$], *anxiously*_[$\beta=0.15$], *distressed*_[$\beta=0.11$], *unforeseen*_[$\beta=0.08$]. This finding aligns with results from prior psychology research showing that the more threatening and distressing the situation, the more critical is the need to reduce one’s feelings of anxiety; individuals under such scenarios often tend to be more credulous [151].

In contrast, subjective words denoting exasperation (*damn*_[$\beta=-0.38$], *goddam*_[$\beta=-0.69$]), expressions denoting feelings of shock and disappointment (*awfulness*_[$\beta=-0.28$], *appalled*_[$\beta=-0.19$], *shockingly*_[$\beta=-0.11$]) were associated with lower levels of credibility. This finding echoes findings from prior work, in which the presence of swear words in tweets denotes reactions to an event and are less likely to contain information about the event [60]. Thus event reportages with lower informational content would be perceived as less credible. Other correlates of negative β s include subjective words signaling enquiry and assessment (*contradict*_[$\beta=-0.44$], *pry*_[$\beta=-0.37$], *perspective*_[$\beta=-0.09$], *unspecified*_[$\beta=-0.07$], *ponder*_[$\beta=-0.05$], *scrutinize*_[$\beta=-0.03$]) and words expressing ambiguity (*peculiar*_[$\beta=-0.13$], *confusing*_[$\beta=-0.05$], *obscurity*_[$\beta=-0.05$], *disbelief*_[$\beta=-0.04$]). Research on identifying rumors in social media have demonstrated that, when exposed to a

rumor, people act as information seekers and thus make enquiries and express doubt before deciding to believe or debunk a rumor [150, 209].

Moreover, subjective words pointing out impracticality and unreasonableness (*unexpected*_[$\beta=-0.05$], *delusional*_[$\beta=-0.05$], *fanatical*_[$\beta=-0.06$], *paranoid*_[$\beta=-0.01$], *lunatic*_[$\beta=-0.02$]) and words conveying doubt (*lacking*_[$\beta=-0.26$], *nevertheless*_[$\beta=-0.36$], *likelihood*_[$\beta=-0.26$], *tentative*_[$\beta=-0.02$], *suspicion*_[$\beta=-0.07$], *dispute*_[$\beta=-0.10$], *moot*_[$\beta=-0.07$], *best known*_[$\beta=-0.19$]) were also associated with lower perceptions of credibility. These findings demonstrate the underlying sense-making activity undertaken as an attempt to assess dubious information before deciding on its accuracy. Furthermore, I find that subjective words denoting fast and frantic reaction were weak predictors of lower credibility levels: (*fleeting*_[$\beta=-0.05$], *speedy*_[$\beta=-0.04$], *frenetic*_[$\beta=-0.01$]). This suggests that quick and speedy information is often viewed as having lower levels of credibility.

Positive & Negative Emotion: The phrases in both the emotion categories were found to have substantial predictive power when included as features in original and reply tweets. While the color saturation of the emotion category (Figure 15) with respect to the original tweets tends to green, color saturation for replies tends to red. This suggests a fundamental difference in the way emotion-laden words were perceived in originals and replies while assessing credibility level of information. While replies associate negative sentiment with lower perceptions of credibility, originals relate positive sentiment with higher perceived credibility. Moreover, the prominent green color saturation for the positive emotion in original tweets and strong red color saturation for the negative emotion in reply tweets further emphasized this difference. These observations also indicate that replies play a key role in the collective sense-making process when faced with less credible information.

Looking at the emotion words with non-zero β weights, I found an intriguing view of how sentiment words provide cues of high and low credibility perceptions. Similar to subjectivity category, negative emotion words denoting extreme distress and loss in original tweets were associated with higher levels of perceived credibility (*sucky*_[$\beta=0.57$], *piti*^{*}_[$\beta=0.34$], *aggravat*^{*}_[$\beta=0.21$], *loser*^{*}_[$\beta=0.2$], *troub*^{*}_[$\beta=0.20$], *misses*_[$\beta=0.17$], *heartbroke*_[$\beta=0.12$], *sobbed*_[$\beta=0.04$],

*weep*_[\beta=0.02]^{*}, *fail*_[\beta=0.75]^{*} 0.02, *defeat*_[\beta=0.02]) . I found a similar trend in replies. Negative emotion category in replies correlated with higher perceived credibility and was expressed with words such as *stink*_[\beta=0.51]^{*}, *griev*_[\beta=0.29]^{*}, *sucky*_[\beta=0.24], *devastating*_[\beta=0.24], *victim*_[\beta=0.07]^{*}.

On the other hand, positive emotion words indicating agreement were predictors of higher levels of perceived credibility, both in original and reply tweets. Example predictive phrases from originals include: *eager*_[\beta=0.28], *dynam*_[\beta=0.25]^{*}, *wins*_[\beta=0.24], *terrific*_[\beta=0.07], *okays*_[\beta=0.04], while reply tweets had predictive phrases like *yay*_[\beta=0.47], *convinc*_[\beta=0.43]^{*}, *agreed*_[\beta=0.28], *impress*_[\beta=0.25]^{*}, *loved*_[\beta=0.20], *brilliant*_[\beta=0.19]^{*}, *fantastic*_[\beta=0.18], *wonderf*_[\beta=0.06]^{*}. Note that adjectives like *eager*, *dynamic*, *terrific*, *brilliant*, *fantastic*, *wonderful* are commonly used to qualify the factuality of information in an event [153, 159].

One of the most compelling findings were the list of emotion phrases correlating with lower levels of credibility. For the positive emotion category with respect to original tweets, such predictive words included *ha*_[\beta=-0.11], *please*_[\beta=-0.13], *joking*_[\beta=-0.03]. For the replies I found predictive words such as, *grins*_[\beta=-0.19], *ha*_[\beta=-0.07], *heh*_[\beta=-0.06], *silli*_[\beta=-0.02]^{*}, *joking*_[\beta=-0.01]. These phrases ridicule the absurdity of information – a characteristic commonly seen in fake news and rumors. The negative emotion phrases associated with lower levels of credibility painted a similar picture with predictive words like, *lame*_[\beta=-0.18], *cheat*_[\beta=-0.13]^{*}, *careless*_[\beta=-0.31] from the original tweets and *grave*_[\beta=-0.27], *liar*_[\beta=-0.16], *mocks*_[\beta=-0.16], *distrust*_[\beta=-0.12] from the replies.

Hedges & Boosters: While hedges and booster words have significant predictive power, the color saturation shows reverse trends in original and reply tweets. This suggests a vital difference in the way expressions of certainty and tentativeness are perceived in originals and replies during credibility assessments. While boosters in original tweets were more strongly related to lower perceived credibility, boosters in replies contributed to higher levels of credibility. A similar inversion was observed for hedges, indicating that emphasizing

¹A word ending in * denotes a word stem. For example, the stem *troubl** would match with any target word starting with the first five letters, such as *troublesome*, *troubles*, *troubled*.

claims made in an original tweet through the use of booster words provides a good signal of lower credibility levels (*without doubt*_[β=-0.25], *invariabl*^{*}_[β=-0.13], *musn't*_[β=-0.06]). In contrast, booster words in replies, cues the presence of credible information by emphasizing assertions (*undeniable*_[β=0.36], *shows*_[β=0.23], *guarant*^{*}_[β=0.05]) or signaling past knowledge acquisition (*defined*_[β=0.34], *shown*_[β=0.17], *completed*_[β=0.003]).

Hedges paint a different picture. Hedge words in original tweets conveying information uncertainty (*appeared*_[β=0.26], *halfass*^{*}_[β=0.13], *to my knowledge*_[β=0.12], *tends to*_[β=0.02]) or qualifying claims with conditions (*depending*_[β=0.23] *contingen*^{*}_[β=0.14]) were viewed as having higher credibility. In contrast, hedging in replies was used to express suspicion and raise questions regarding a dubious original tweet. Hence hedge words like *certain level*_[β=-0.16], *dubious*^{*}_[β=-0.12], *suspects*_[β=-0.08] were correlated with lower levels of credibility. As before, when hedges corroborate information with conditions in the reply tweets, they signaled higher levels of credibility (*guessed*_[β=0.28], *borderline*_[β=0.27], *in general*_[β=0.09], *fuzz*^{*}_[β=0.02]).

Evidentials: Evidentials contribute different shades of factuality information to an event's reportage. Phrases from the evidential category alone were able to explain more than 16% of the variance observed in the data. The top predictive evidentials associated with higher credibility illustrate event reportage (*tell*_[β=0.14], *express*_[β=0.06], *describe*_[β=0.05] in originals, *declare*_[β=0.22], *post*_[β=0.02], *according*_[β=0.05] in replies), fact checking (*verify*_[β=0.05], *assert*_[β=0.18] in replies) and knowledge acquisition (*discover* in both replies and original tweets). In contrast, evidentials correlating with lower levels of credibility indicated losing knowledge (*forget*_[β=-0.50] in replies), expressing uncertainty (*reckon*_[β=-0.03], *predict*_[β=-0.01] in replies) and fact checking in originals (*check*_[β=-0.09], *verify*_[β=-0.02]). *told*_[β=-0.13], one of the top predicates correlating with lower credibility was used in positioning a tweet's claim as uncommitted with respect to the factuality:

Roux's snide remark when arbitrary lawyer **told** Roux to get Ubuntu book- as if legal world support him?*smh*. #OscarPistorius #OscarTrial

Anxiety: Words expressing anxiety had significant predictive power as well. As before, I observed reverse color saturation trends in original and reply tweets, suggesting anxiety utterances are perceived differently in originals and replies during credibility assessments. In original tweets, anxiety words questioning the practicality of a claim (*crazz*_[$\beta=0.11$]^{*}, *irrational*_[$\beta=-0.03$], *embarrass*_[$\beta=-0.02$]) were associated with lower levels of credibility. On the other hand, anxiety words with $\beta > 0$ exuded disappointment with the situation: *distress*_[$\beta=0.24$]^{*}, *miser*_[$\beta=0.15$]^{*}, *startl*_[$\beta=0.08$]^{*}. Essentially this set of anxiety words were used to express opinion on an already existing event.

Only 1 **miserable** goal??

Watching the Eric Garner video was so **distressing**, sick bastards going unpunished for killing an innocent man in broad daylight

16 disgusting and **distressing** abuses detailed in the CIA torture report.

Additionally, apprehensive expressions in both originals and replies (*vulnerabl*_[$\beta=-0.34$]^{*}, *uncontrol*_[$\beta=-0.08$]^{*}, *turmoil*_[$\beta=-0.04$]), and words indicating fear in replies (*fear*_[$\beta=-0.15$], *petrif*_[$\beta=-0.12$]^{*}) were associated with lower perceived credibility. This finding aligns with findings from social psychology, which emphasizes the role of anxiety in rumormongering. All these negative β words stressed on the severity of the threat, and prior studies have shown that during threatening situations rumors are aimed at relieving tensions of anxiety [20].

Conjunctions, Exclusions, Negation & Modality: Words from the conjunction category associated with lower levels of credibility ($\beta < 0$) were used for reasoning and drawing inferences: *because*_[$\beta=-0.07$], *then*_[$\beta=-0.37$], *when*_[$\beta=-0.18$], whereas words correlated with higher credibility levels ($\beta > 0$) were used for creating coherent narratives: *while*_[$\beta=0.47$], *as*_[$\beta=0.14$], *til*_[$\beta=0.04$]^{*}. These findings suggest that presence of conjunctions to facilitate coherent narrative is a signal for high credibility.

Additionally, I found that predictive words in the exclusion category exhibited characteristics similar to that of hedges outlined earlier. While words associated with lower levels

of credibility ($\beta < 0$) signaled the presence of ambiguity (*something*_[\beta=-0.09]), words with positive β qualified claims with conditions (*exclu**_[\beta=-0.03]). Words from the modality and negation categories did not emerge as predictive features in the context of original tweets. For reply tweets, the only modal word associated with lower levels of credibility indicated use of the evidential strategy (*reportedly*_[\beta=-0.53]). The negation words corresponding to reply tweets surfaced as predictors of lower perceived credibility. Example words included *neither*, *nowhere*, both of which were used to signal disagreements.

Mixed Category: Recall that our mixed category contained phrases belonging to multiple lexicons and was added to tackle the double counting of features. Phrases in the mixed category had substantial predictive power. A deeper look into the phrases revealed that words denoting agreement were associated with higher perceived credibility (*clear*_[\beta=0.18], *established*_[\beta=0.23], *agree*_[\beta=0.08] in the context of originals and *glad*_[\beta=0.60], *definite*_[\beta=0.06], *established*_[\beta=0.04] as features in reply tweets). Conversely, words with a ring of hedging (*apparently*_[\beta=-0.11], *fairly*_[\beta=-0.19], *messy*_[\beta=-0.12], *if*_[\beta=-0.10]), phrases expressing disagreement (*impossible*_[\beta=-0.17]) and words mocking at the irrationality of statements (*hilarious*_[\beta=-0.06], *fun*_[\beta=-0.08]) were correlated with lower levels of credibility.

5.4.1 Theoretical Implications

Despite the popularity of multi-media based interactions, social conversations on most CMC systems are largely done through texts. Methods, such as ours which can automatically analyze CMC generated textual content and draw meaningful inferences about human behavior can be of immense value to researchers from different domains. For instance, a linguist might investigate the relationship between language and speaker commitment or study textual factors shaping reader's perspective. A social scientist might explore types of language which drive collective sense making in times of uncertainty. A behavioral psychologist can use our findings to understand the types of behaviors exhibited in information assessment. For example, studies have shown that question asking is a common behavior in

social media and is often used for seeking information about real-world events including rumors [208, 209]. Our results indicate the importance of questioning the rationality of claims through the use of anxiety and positive emotion words, expressing suspicion through the use of hedges and emphasizing a less credible claim with language boosters. These findings can be the starting point for understanding the common information assessment behaviors exhibited on online social media and how these behaviors manifest at scale.

While I know a great deal about the relationship between language and sentiments or language and opinion, I know very little about how people perceive credibility of events in textual conversations. By studying social media credibility through a linguistically well-grounded model, I believe that in addition to providing theoretical insights on the relationship between language and credibility perceptions, our work can also complement current predictive modeling techniques. Moreover, unlike previous explorations of language signals of credibility, our work is based on a comprehensive collection of a large set of social media events. Hence the subsequent inferences drawn by this study circumvents the problem of sampling bias otherwise present in studies based on a handful of pre-selected social media event reportages.

5.4.2 Design Implications

I believe that our work can inform the design of a wide-array of systems. For example, imagine a news reporting tool which surfaces eye witness reports from social media and highlights those which are associated with high versus low perceptions of credibility, or consider a fact checking system which highlights high versus low credible slices of event reportage. While I do not claim that our classifier can be deployed as a standalone system to verify facts or debunk rumors, but at the least it can be used to extract reliable credibility signals from text alone. I believe that when used in combination with other extra-linguistic variables, it can complement and add value to existing fact checking systems. For example, extra-linguistic features such as author of the content, the involvement of the author in the

topic of the content (such as, proportion of prior tweets posted by the individual), the type of source (an established news source or an eyewitness account) and content novelty (whether it is a first time report of an event or emerging information about an already reported event) can be useful additions to a language-based fact checker.

Further, most existing approaches that attempt to classify the credibility of online content utilize information beyond the content of the posts, usually by analyzing the collective behavior of users involved in content circulation. For example, temporal patterns of content [61, 92], popularity of the post (measured by the number retweets or replies) [61] or the network structure of content diffusion [24, 92, 145]. While useful, these features can only be collected after the content (whether accurate or not) have disseminated for a while [209]. Utilizing language markers is a key towards early detection of low credible content, thereby limiting their damage.

Additionally, our results can enable a new class of systems to underscore degrees of uncertainty in news reporting, in medical records or even in scientific discourses. Our findings can also equip systems to highlight apprehensions in event reporting or surface the irrationality of claims. Moreover, event reportage is not limited to one CMC system, such as Twitter. In addition to a plethora of existing systems enabling reporting of events, often new CMC systems emerge. Hence a designer would want to build a tool which is domain-independent or one which can be easily adapted to a new domain. Given that most linguistic expressions are not domain specific, it might be possible to build such a tool without the overhead of domain adaptation. At most, it will involve refining the current set of language markers. For example, refining the set of hedge markers or booster words for the new domain.

5.5 Conclusion

In this study I uncovered words and phrases which indicate whether an event will be perceived as highly credible or less credible. By developing a theory driven, parsimonious

model working on millions of tweets corresponding to thousands of events and their corresponding credibility annotations, I unfold ways in which social media text carry signals of information credibility. This is an empirical result, not a deployable system; however, when combined with other signals (e.g., temporality, structural information, event type, event topic etc.) the linguistic result reported here could be an important building block of an automated system. In brief, these results show that the language used by millions of people on Twitter has considerable information about an event's perceived credibility. I hope these findings motivate future researchers to explore dynamics of event credibility through linguistically-oriented computational models or extend this line of work to include higher level interaction terms, such as including discourse relations and syntactic constructions.

Table 5: Sample of events from the CREDANK corpus grouped by their credibility classes. Events are represented with three event terms. Start and end times denote the time period during which Mitra et al. [121] collected tweets using Twitter’s search API combined with a search query containing a boolean *AND* of all three event terms. Ratings show the count of Turkers that selected an option from the 5-point, ordinal Likert scale ranging between -2 (“Certainly Inaccurate”) to +2 (“Certainly Accurate”). Each event was annotated by 30 Turkers.

Event Terms	# Tweets	Start time	End Time	Ratings	P_{ca}
Perfect Credibility: $0.9 \leq P_{ca} \leq 1$					
george clooney #goldenglobes	10350	2015-01-12 08:50	2015-01-12 18:10	[0 0 1 1 28]	0.93
king mlk martin	88045	2015-01-15 22:00	2015-01-15 22:00	[0 0 0 2 28]	0.93
win pakistan test	5478	2014-10-26 18:10	2014-11-03 21:00	[0 0 0 3 27]	0.90
george arrested zimmerman	45645	2015-01-07 19:40	2015-01-11 00:50	[0 0 0 3 27]	0.90
scott rip sad	26006	2014-12-29 07:50	2015-01-05 18:10	[0 0 0 3 27]	0.90
hughes rip phil	157258	2014-11-25 11:40	2014-11-28 09:00	[0 0 1 2 27]	0.90
breaking jones positive	19973	2015-01-07 03:30	2015-01-07 16:00	[0 0 0 3 27]	0.90
apple ipad air	169182	2014-10-09 13:10	2014-10-17 09:40	[0 1 1 1 27]	0.90
george arrested zimmerman	45645	2015-01-07 19:40	2015-01-11 00:50	[0 0 0 3 27]	0.90
missing flight singapore	88144	2014-12-27 18:50	2014-12-28 21:00	[0 0 1 2 27]	0.90
High Credibility: $0.8 \leq P_{ca} < 0.9$					
beckham odell catches	21848	2014-11-04 04:10	2014-11-04 22:20	[0 0 0 4 26]	0.87
eric garner death	180582	2014-11-26 08:30	2014-12-04 07:10	[1 1 0 2 26]	0.87
windows microsoft holographic	18306	2015-01-21 23:40	2015-01-25 10:00	[0 0 0 4 26]	0.87
kayla mueller isis	65819	2015-02-06 21:10	2015-02-12 00:10	[0 0 0 8 52]	0.87
liverpool arsenal goal	16713	2014-12-14 05:20	2014-12-14 05:20	[0 1 0 4 25]	0.83
korea north sanctions	57529	2014-12-27 19:30	2014-12-27 19:30	[0 0 0 5 25]	0.83
copenhagen police shooting	26986	2015-02-14 20:40	2015-02-16 04:00	[1 0 0 4 25]	0.83
paris charlie attack	224673	2015-01-07 15:50	2015-01-10 15:30	[0 0 1 5 24]	0.80
nigeria free ebola	32412	2014-10-20 17:00	2014-10-21 07:30	[1 0 1 4 24]	0.80
japanese video hostages	23759	2015-01-20 11:00	2015-01-24 12:30	[0 0 2 4 24]	0.80
Moderate Credibility: $0.6 \leq P_{ca} < 0.8$					
children pakistan #peshawarattack	24239	2014-12-16 12:30	2014-12-17 20:10	[0 1 1 5 23]	0.77
obama president #immigrationaction	57385	2014-11-19 23:00	2014-11-21 12:50	[1 0 0 6 23]	0.77
#ericgarner protesters police	12510	2014-12-04 00:50	2014-12-05 10:20	[0 0 2 6 22]	0.73
sydney hostage #sydnaysiege	21835	2014-12-15 04:20	2014-12-15 17:20	[0 0 2 6 22]	0.73
bobby shmurda bail	22362	2014-12-17 21:40	2014-12-19 17:30	[0 0 1 7 22]	0.73
news isis breaking	17408	2015-02-11 02:30	2015-02-18 19:30	[1 1 1 7 20]	0.67
chris #oscar evans	3096	2015-02-16 18:50	2015-02-23 19:50	[1 0 4 5 20]	0.67
torture report cia	61045	2014-12-10 01:00	2014-12-12 19:50	[1 1 2 5 21]	0.60
chelsea game goal	544	2014-11-15 01:50	2014-11-23 04:40	[1 0 5 6 18]	0.60
#antoniomartin ambulance shot	6330	2014-12-24 11:30	2014-12-24 23:10	[0 0 3 9 18]	0.60
Low Credibility: $0 \leq P_{ca} < 0.6$					
syria isis state	6547	2015-02-17 11:00	2015-02-24 19:50	[0 0 2 11 17]	0.57
gerrard liverpool steven	204026	2014-12-26 03:40	2015-01-02 20:20	[0 1 3 9 17]	0.57
police #antoniomartin officer	13141	2014-12-24 11:20	2014-12-25 01:50	[0 1 3 9 17]	0.57
#charliehebd #jesuischarlie religion	4939	2015-01-07 17:30	2015-01-08 08:50	[0 2 7 4 17]	0.57
#chapelhillshooting muslim white	35282	2015-02-11 11:20	2015-02-13 06:20	[2 2 8 16 32]	0.53
paris boko killed	3917	2015-01-07 22:50	2015-01-11 01:50	[0 3 1 11 15]	0.50
next coach michigan	7811	2015-02-04 05:00	2015-02-09 16:20	[0 4 6 20 30]	0.50
news breaking ebola	45633	2014-10-11 06:40	2014-10-19 06:20	[1 3 6 8 12]	0.40
ebola #ebola travel	27796	2014-10-09 06:10	2014-10-17 09:10	[2 2 6 10 10]	0.33
baylor kicker dead	31341	2015-01-02 02:30	2015-01-02 23:20	[15 3 6 1 5]	0.17

Table 6: List of feature categories used by our language classifier. Features are categorized as lexicon-based, non-lexicon based and control features. For the lexicon based measures I included words from each of the lexicons as features – yielding a total of 9,659 words obtained by summing the lexicon sizes. Adding the non-lexicon based features resulted in a total of 9,663 linguistic features.

Lexicon-based measures		Lexicon Size
Modality		30
Subjectivity		8222
Hedges		125
Evidentiality		82
Negation		12
Exclusions		17
Conjunction		28
Boosters		145
Anxiety		91
Positive Emotion		499
Negative Emotion		408
Non-lexicon based		
Hashtags		Quotation
Questions		Capitalization
Controls		
Tweet count	Reply count	Retweet count
Avg. tweet length	Avg. reply length	Avg. retweet length
Tweet word count	Reply word count	Retweet words count

Table 7: Summary of different model fits sorted by % variance explained. *Null* is the intercept-only model. *Dev* denotes deviance which measures the goodness of fit. All comparisons with the Null model are statistically significant after Bonferroni correction for multiple testing. The table's top half shows that the omnibus model containing controls and variables based on all linguistic measures for both tweets and replies is the best model. The bottom half of the table reports model performance for the omnibus model for each set of linguistic categories. It also shows deviance per linguistic category for original and replies (in gray).

Model	Dev	% Var	df	χ^2
Null	3,937.37		0	
Controls Only	3,499.54	11.12	13	437.84
Reply Tweets + Controls	1,603.30	59.28	1227	2,326.99
Original Tweets + Controls	1,396.98	64.52	1143	2,540.39
Original + Replies + Controls (Omni)	1,181.65	69.99	1234	2,755.72

Omnibus Model by Linguistic Feature				
Linguistic Feature	Dev	% Var	df	χ^2
All Subjectivity (omni)	1,539.91	60.89	1063	2,397.47
Positive Emotion (omni)	2,331.71	40.78	621	1,605.66
original + control	2,860.11	27.36	325	1,077.27
reply + control	2,922.32	25.78	309	1,015.06
Negative Emotion (omni)	2,360.85	40.04	665	1,576.52
original + control	2,792.39	29.08	349	1,144.99
reply + control	2,882.16	26.8	330	1,055.22
Mixed (omni)	2387.62	39.36	633	3,936.98
original + control	2,829.00	28.15	332	3,937.09
reply + control	2,962.87	24.75	308	974.5
Anxiety (omni)	3111.71	20.97	191	3,937.16
original + control	3,245.97	17.56	103	3,937.20
reply + control	3,350.71	14.9	86	586.67
Boosters (omni)	3158.56	19.78	160	778.81
original + control	3,325.51	15.54	89	611.87
reply + control	3,316.45	15.77	83	620.92
Hedges (omni)	3221.56	18.18	143	715.81
original + control	3,338.89	15.2	85	598.48
reply + control	3,373.54	14.32	70	563.83
Evidentiality (omni)	3284.95	16.57	96	652.42
original + control	3,371.57	14.37	54	565.8
reply + control	3,372.75	14.34	54	564.62
Conjunction (omni)	3384.17	14.05	48	553.2
original + control	3,434.97	12.76	30	502.41
reply + control	3,446.38	12.47	30	490.99
Exclusions (omni)	3444.81	12.51	28	492.57
original + control	3,465.28	11.99	20	472.09
reply + control	3,460.16	12.12	20	477.21
Negation (omni)	3452.68	12.31	47	484.69
original + control	3,452.68	12.31	24	484.69
reply + control	3,451.90	12.33	19	485.48
Modality (omni)	3455.83	12.23	24	481.54
original + control	3,461.35	12.09	18	476.03
reply + control	3,475.52	11.73	18	461.85

Table 8: Precision (P), Recall (R), F1-measure and Accuracy of two baseline classifiers: 1) Random Guess and 2) Random Weighted Guess, along with performance measures of the language classifier. I show four accuracy measurement schemes for our language classifier: 1) Unweighted is the most conservative way of measuring accuracy with no credit given for incorrect classification. It uses the unweighted credit matrix from Table 9a 2). Level-1 Weight_{0.25} gives partial credit of 0.25 if the classification is incorrect by one level only (Table 9b), 3). Level-1 Weight_{0.5} is similar but the rewarded partial credit is higher (0.5). Level-2 Weight_{0.25,0.5} gives partial credit as per the weighted matrix shown in Table 9c. Our language classifier significantly outperforms both the baselines (McNemar’s test, $p < 10^{-16}$).

	Baseline Classifiers						Language Classifier											
	Random Guess			Random Weight			Unweighted			Level-1 Wt. _{0.25}			Level-1 Wt. _{0.5}			Level-2 Wt. _{0.25,0.5}		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Low	5.30	25.0	8.80	5.30	5.30	5.30	17.9	11.8	14.2	29.1	29.1	29.1	40.2	46.4	43.1	46.2	50.9	48.4
Moderate	40.7	25.0	31.0	40.7	40.7	40.7	41.2	56.0	47.5	51.6	64.8	57.4	62.0	73.5	67.3	66.3	75.8	70.7
High	31.7	25.0	27.9	31.7	31.7	31.7	38.2	38.5	38.4	52.5	52.9	52.7	66.8	67.2	67.0	68.0	68.2	68.1
Perfect	22.3	25.0	23.6	22.3	22.3	22.3	57.2	41.8	48.3	64.8	50.0	56.5	72.4	58.2	64.5	75.4	64.0	69.2
Accuracy	25.00			31.84			42.59			53.63			64.92			67.78		

Table 9: Full credit is given for correct classification, denoted by 1’s along the diagonal. (a) No credit is given for incorrect classification (0’s along the non-diagonals). (b) Partial credit (0.25) is given if the classifier gets it wrong by one level and no credit is given if the predictions are off by two or more levels. (c) Partial credit (0.5) is given if the classifier gets it wrong by one level, (0.25) for two level and no credit if the predictions are wrong by three or more levels. There are four levels in the ordinal classes: Low (L), Medium (M), High (H), and Perfect (P).

	L	M	H	P
L	1	0	0	0
M	0	1	0	0
H	0	0	1	0
P	0	0	0	1

(a) Unweighted Credit Matrix

	L	M	H	P
L	1	0.25	0	0
M	0.25	1	0.25	0
H	0	0.25	1	0.25
P	0	0	0.25	1

(b) Weighted Matrix (Level 1)

	L	M	H	P
L	1	0.50	0.25	0
M	0.50	1	0.50	0.25
H	0.25	0.50	1	0.50
P	0	0.25	0.50	1

(c) Weighted Matrix (Level 2)

Table 10: The top predictive words in the subjectivity category corresponding to the original tweets. Words associated with higher ($\beta > 0$) and lower ($\beta < 0$) levels of perceived credibility are shown in respective columns. All words are significant at the 0.001 level.

Subjectivity	$\beta > 0$	Subjectivity	$\beta > 0$	Subjectivity	$\beta < 0$	Subjectivity	$\beta < 0$
vibrant	0.85	unique	0.75	goddam	-0.69	contradict	-0.44
intricate	0.69	amazement	0.67	damn	-0.38	pry	-0.37
inexplicable	0.61	immaculate	0.61	nevertheless	-0.36	awfulness	-0.28
darn	0.54	close	0.45	likelihood	-0.26	lacking	-0.26
precise	0.43	mortified	0.35	best known	-0.19	appalled	-0.19
validity	0.34	promising	0.33	shockingly	-0.11	confuse	-0.10
mishap	0.32	calamity	0.30	dispute	-0.10	perspective	-0.09
catastrophic	0.30	ecstatic	0.23	moot	-0.07	suspicion	-0.07
exceptionally	0.19	anxiously	0.15	unspecified	-0.07	fanatical	-0.06
unanimous	0.12	distressed	0.11	delusional	-0.05	ponder	-0.05
reliability	0.11	distinctive	0.08	unexpected	-0.05	fleeting	-0.05
charismatic	0.08	unforeseen	0.08	obscurity	-0.05	speedy	-0.04
brilliant	0.08	strangely	0.07	disbelief	-0.04	scrutinize	-0.03
awed	0.07	bright	0.07	tentative	-0.02	lunatic	-0.02
miraculously	0.06	radiant	0.06	paranoid	-0.01	frenetic	-0.01

Table 11: The top predictive phrases per linguistic category associated with higher ($\beta > 0$) and lower ($\beta < 0$) levels of credibility. Phrases corresponding to the original tweets are on the left, those corresponding to replies are on the right. All phrases are significant at the 0.001 level.

Original Tweet				Reply Tweet			
Positive Emotion	$\beta > 0$	Positive Emotion	$\beta < 0$	Positive Emotion	$\beta > 0$	Positive Emotion	$\beta < 0$
eager*	0.28	yays	-0.20	yay	0.47	grins	-0.19
dynam*	0.25	reassur*	-0.20	convinc*	0.43	ha	-0.07
wins	0.24	please*	-0.13	agreed	0.28	heh	-0.06
terrific*	0.07	ha	-0.11	impress*	0.26	silli*	-0.02
okays	0.04	joking	-0.03	loved	0.20	joking	-0.01
splend*	0.04			brilliant*	0.19		
wonderf*	0.03			fantastic*	0.18		
Negative Emotion	$\beta > 0$	Negative Emotion	$\beta < 0$	Negative Emotion	$\beta > 0$	Negative Emotion	$\beta < 0$
sucky	0.57	careless*	-0.31	stink*	0.51	woe*	-0.63
piti*	0.34	lame*	-0.19	griev*	0.29	smother*	-0.57
aggravat*	0.21	fuck	-0.14	devastat*	0.24	grave*	-0.27
loser*	0.20	cheat*	-0.13	sucky	0.20	mocks	-0.16
troubl*	0.20	egotis*	-0.09	obnoxious*	0.09	liar*	-0.16
misses	0.17	unsuccessful*	-0.03	troubl*	0.08	distrust*	-0.12
missed	0.12	distrust*	-0.01	victim*	0.07	fuck	-0.05
heartbroke*	0.12	contradic*	-0.01	ugl*	0.04	paranoi*	-0.05
sobbed	0.04			heartbroke*	0.02	weird*	-0.04
weep*	0.02					Negation	$\beta < 0$
fail*	0.02					neither	-0.02
defeat*	0.02					nowhere	-0.12
Hedges	$\beta > 0$	Hedges	$\beta < 0$	Hedges	$\beta > 0$	Hedges	$\beta < 0$
appeared	0.26	indicates	-0.18	guessed	0.28	certain level	-0.16
depending	0.23	from my perspective	-0.15	borderline*	0.27	dubious*	-0.12
contingen*	0.14	suggested	-0.07	in general	0.09	suspects	-0.08
halfass*	0.13	dunno	-0.04	fuzz*	0.02	approximately	-0.04
to my knowledge	0.12	borderline*	0.00	almost	0.01	dunno	-0.04
tends to	0.02					Exclusion	$\beta < 0$
						exclu*	0.12 something*
						Booster	$\beta < 0$
						undeniable	0.36 implicit*
						shows	0.23 total
						guarant*	0.05 factual*
						Anxiety	$\beta < 0$
						distress*	0.24 fear
						miser*	0.15 petrif*
						startl*	0.08 inadequa.
						nervous*	0.04 desperat*
						impatien*	0.03 shaki*
						Modality	$\beta < 0$
							reportedly
						Evidentials	$\beta < 0$
						verify	0.05 predict
						post	0.02 reckon
						discover	0.01 forgot
						Conjunction	$\beta < 0$
						as	0.14 then
						how	0.06 when
						til	0.04 because
						Mixed	$\beta < 0$
						glad	0.59 fairly
						definite	0.06 messy
						established	0.04 if

CHAPTER VI

TEMPORAL DYNAMICS OF CREDIBILITY

Information spreading through social media exhibit rich temporal dynamics. While information diffusing through micro-blogging platforms like Twitter have a short life span [201], with content rising and falling in popularity within hours; short quoted phrases (known as *memes*) take several days to rise and fade away [101]. On the other hand, general themes (like ‘politics’, ‘economy’, ‘terrorism’) have an even larger temporal life span [4, 59, 191]. Social psychologists studying the spread of news and rumor have also noted the importance of temporal patterns in rumor transmission – different types of rumor mongering statements persist over varying temporal spans [20, 169]. Despite the importance of temporal patterns in information diffusion and rumor transmission, there has been little work in understanding temporal trends in events and its associated credibility assessments. In this chapter, In this chapter, I investigate this less treaded research area. One reason behind the shortage of temporal studies is the lack of time stamped data with in-situ credibility annotations. CREDBANK’s corpus containing timing information at every step of its building process overcomes this shortcoming. I operationalize the study of temporal dynamics through the lens of collective attention, that is, how many people are collectively paying attention to a social media event at a certain point in time, and how that attention changes over time. Nearly universally, all social systems are designed around time as the central axes. Their interfaces are designed and driven around when people post messages and how often they post. Hence, studying temporal dynamics of credibility through the lens of collective human attention seems natural and intuitive to me. The central question which this chapter addresses is the following:

How do the dynamics of collective attention directed toward an event reported on social

media vary with its credibility?

Collective human attention is essential for information propagation in social networks [33,201]. It drives various social, economic and technological phenomenon, such as herding behavior in financial markets [172], formation of trends [9], popularity of news [201], web pages [146], and music [156], propagation of memes [100], ideas, opinions and topics [149], person-to-person word-of-mouth advertising and viral marketing [100], and diffusion of product and innovation [13]. Moreover, it is the key phenomenon underlying social media reporting of emerging topics and breaking news [116].

A fundamental attribute underlying any collective human behavior is how that behavior unfolds over time [12,33]. Is there a relationship between allocation of collective attention and perceived credibility of events reported through social media? Do occasional bursts in collective attention—as more eyes and voices are drawn to the event’s reportage—correspond to less certain information concerning the event? Uncovering the relationship between collective human behavior and information credibility is important for assessing the veracity of event reportage as it unfolds on social media. This relationship, if it exists, can provide insights into ways to disambiguate misinformation from accurate news stories in social networks—a medium central to the way we consume information [25] and one where digital misinformation is pervasive [39].

Empirical attempts at answering these questions in naturalistic settings have been constrained by difficulties in tracking social media posts in conjunction with judgments concerning the accuracy of the underlying information. Previous studies have instead focused on individual case studies involving specific news events [8,107,111], or have retrospectively studied a set of multiple prominent events [39,49] which were known to contain misinformation. While useful, these approaches raise sampling concerns. In particular, they are based on the post-hoc investigation of events with known credibility levels, and thus select on the dependent variable [187]. Although these studies suggest the possibility of spikes in collective attention when false rumors propagate through social networks, the relation

between collective attention and information credibility has not been systematically tested. This chapter attempts to fill this gap. Although the nature of this data limits causal inference, I tested the correspondence between collective attention and the level of information credibility. After filtering out unique event instances, I was left with a pruned corpus of 1,138 real-world events spread over 47M tweets. Analyzing this massive dataset, I find that the amount of recurring collective attention bursts could be used to determine the level of perceived credibility of an event. Specifically, I demonstrate that multiple occasional bursts of collective attention toward an event is associated with lower levels of perceived credibility. This finding opens a new perspective in the understanding of human collective attention and its relation to the certainty of information. In doing so, the subsequent results can have widespread implications in fields where predictive inference based on online collective interests dictates economic decisions, emergency responses, resource allocation or product recommendations [53,202]; hence trusting the credibility of the collective reports is essential for an accurate anticipation by the predictive process.

6.1 Related Work

6.1.1 Collective Attention

A phenomenon which is vital towards the spread of social media information is “collective attention” [146]. Hence, researchers have been attracted toward understanding how attention to new information propagates among large groups of people. While some studies have shown that dynamics of collective attention of online content is characterized by bursts signifying popularity changes [99, 146], others have demonstrated a natural time scale over which attention fades [201]. A study investigating the emergence of collective attention on Twitter, found that although people’s attention is dispersed over a wide variety of concerns, it can concentrate on particular events and shift elsewhere either very rapidly or gradually [158]. Another parallel study focusing on spikes of collective attention in Twitter, analyzed the popularity peaks of Twitter hashtags [99]. They found that the evolution of

hashtag popularity over time defined discrete classes of hashtags. Drawing on the progress of these studies, I ask: does the process of evolving collective attention reflect the underlying credibility of a social media story? Unraveling the relation between collective attention rhythms and corresponding credibility level is a complex empirical problem. It requires longitudinal tracking of collective mentions of newsworthy stories in social media along with their in-situ credibility judgments. To that end, CREDBANK provides the most consistently tracked social media information and its associated credibility scores.

6.1.2 Time Matters

One useful way to understand the interplay between collective attention and information credibility is to examine user activity and information patterns through the lens of time. For decades social scientists have investigated the timing of individual activity to understand the complexity of collective human action. They have reported that timing can range from random [63] to well correlated bursty activity patterns [12]. The bursts in human collective action have not only led to social media reporting of emerging topics, but have also exhibited rich temporal dynamics of social media information spread [116]. For example, information diffusing through micro-blogging platforms like Twitter have demonstrated a short life span [201], with content rising and falling in popularity within hours; whereas, short quoted phrases (known as *memes*) have displayed several days to rise and fade away [101]. On the other hand, general themes (like ‘politics’, ‘economy’, ‘terrorism’) have shown an even larger temporal life span [59, 191]. Social psychologists studying the spread of news and rumor have also noted the importance of temporal patterns in rumor transmission – different types of rumor mongering statements persist over varying temporal spans [20, 169]. However, despite the importance of temporal patterns in information diffusion and rumor transmission, there has been little work in understanding temporal trends in events and its associated credibility assessments. This paper is a step towards unraveling that relation.

6.2 Method

I tested the relation between collective attention and the level of information credibility by analyzing data from the longitudinal credibility corpus, CREDBANK [121]. Recall that the massive corpus was constructed by iteratively tracking millions of public Twitter posts between October 2014 and February 2015.

6.2.1 Pruning Corpus for Sample Independence

During the iterative building of CREDBANK, if an event trended on Twitter for a sufficiently long time period, it is possible that the event is curated multiple times. For example, the event "arsenal", "win", "city" corresponds to the Arsenal's winning the football match against Stoke city. People on Twitter had active conversations about the event for several hours, resulting in the event being captured more than once in CREDBANK. However, our statistical analysis (discussed shortly) required sample independence. Occurrence of multiple instances of the same event will likely violate the independence assumption. Hence, I pruned our event sample to keep single distinct instances of each event. By matching the three terms in each event topic, I looked for duplicate event occurrences. Thereafter, if multiple instances of the same event existed, I picked the event which had the earliest curation time. Restricting events by earliest curation times ensured that I only retained crowd worker annotations corresponding to the very first time that they performed the annotation task; hence preventing any potential prior knowledge bias. The pruning step resulted in a dataset of 1,138 events spanning 47,000,127 tweets.

6.2.2 Credibility Classification

I measured an event's perceived credibility level by considering how many human raters agreed that the event was "Certainly Accurate" (I treat the dependent variable same as I did in Chapter 4). More formally, for each event I find the proportion P_{ca} of ratings marked as "Certainly Accurate". Instead of considering P_{ca} as a continuous variable and have a category corresponding to every value of P_{ca} , I compute four credibility classes that cover

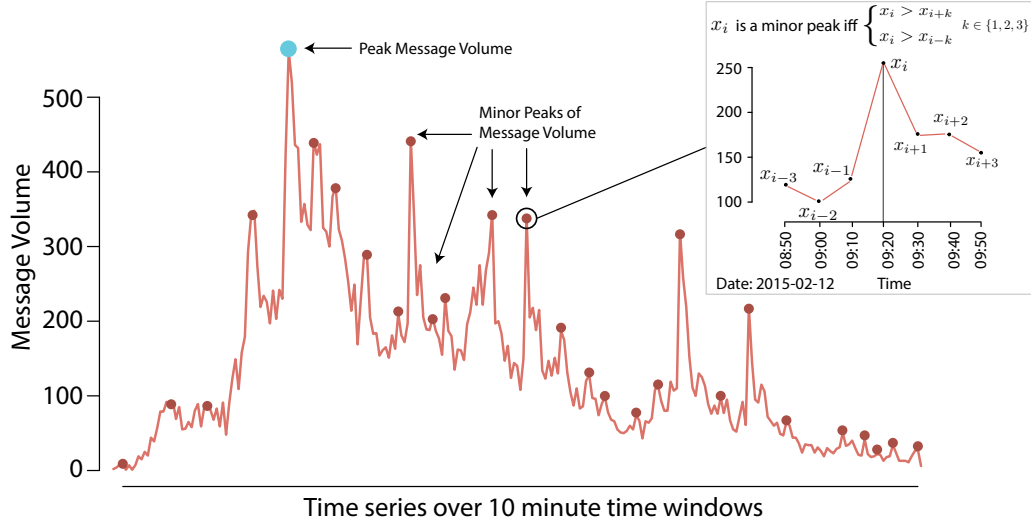


Figure 16: The time series of message volume for a sample event reported on Twitter. The event corresponds to Twitter discussions, where each tweet contained all three terms: “#chapellhillshooting”, “muslim” and “white”. The ● dot corresponds to the time window having maximum *message volume* while the ● dots correspond to the minor peaks observed in this volume. The inset diagram on the right side zooms in on one of the minor peaks, along with the rule triggering its designation.

a range of P_{ca} values (see Table 4). The class names are based on the perceived degree of accuracy of the event in that class. For example, events belonging to the “Perfect Credibility” class were rated as “Certainly Accurate” by almost all raters ($0.9 \leq P_{ca} < 1$).

6.3 Statistical Measures

To understand the relation between collective attention and information credibility, I computed statistical measures using time-stamped tweets from the CREDBANK corpus, where groups of tweets corresponded to discussions of an event by multiple Twitter users over a certain time span.

6.3.1 Collective Attention Metrics

Collective attention of each event reportage was measured using two metrics:

1. *Message Volume*: message volume tracks the aggregate number of messages over time
2. *People Volume*: people volume records the aggregate count of unique users paying attention to the story over time.

Each measure is represented as a time series with message (or unique user) counts aggregated over 10-minute time intervals. My choice of a 10-minute window is supported by studies showing that Twitter acts as a medium for reporting breaking news and hence is characterized by fast diffusion of information [70,90]. Thus, tracking collective attention on the order of minutes is a reasonable representation of a rapidly evolving phenomenon. Each event may differ in the temporal dynamics of its collective attention; thus inferences drawn on a small set of events tracked for a few days may be confounded by temporal traits peculiar to certain news stories. However, by tracking news stories over several months and averaging over hundreds of such collective attention rhythms, the results represent the most consistent relations between the dynamics of collective attention and perceptions of information credibility. My rationale for using *people volume*, in addition to *message volume*, as a collective attention metric is to ensure that the collective attention measured is not confounded by superfluous posting activity from potential Twitter bots, automated programs posing as human beings [28]. Since *people volume* corresponds to the unique number of individuals paying attention to the event over time, it aims to counteract any extreme posting activity by such bots.

As an example illustrating my data and methods, Figure 16 shows aggregate message volume for an event reported on Twitter where every message contained the terms “#chapell-hillshooting”, “muslim” and “white”. On February 10, 2015, three Muslim students in Chapel Hill, North Carolina were shot to death by a white neighbor and speculations concerning the motives of the shooter surrounded the event [50]. While authorities suggested the motive to be an ongoing dispute between neighbors over a parking space, many social media users suggested a hate crime as the motive. Twitter messages concerning this specific topic on February 12 blamed the media for ignoring the coverage of an event involving Muslim killings and suggested the shooting was an act of terrorism and so a hate crime. It was not until February 13 that authorities opened an investigation to determine if the shooting was in fact a hate crime. Credibility rating distributions showed that less than

50% of raters agreed that the social media reportage of the event was “Certainly Accurate”, thereby questioning the alleged terror claims underlying the act.

6.3.2 Temporal Measures of Collective Attention

To quantify the importance of the time when collective attention maximized, I first computed the strict global maximum in the time series [177]. I call this the *peak attention*. This is the ratio of messages (or unique people) within the peak time window to the total cumulative volume of messages (or unique people) over the entire event time series:

$$Peak\ Attention = \frac{\max(x_1, \dots, x_n)}{\sum_{i=1}^n x_i} \quad (2)$$

where x_i is the count of messages (or unique people) in time window i in an event time series x_1, x_2, \dots, x_n . Our choice of the above measure is based on the success of prior studies using peak fraction based metrics to successfully characterize herding behavior over time [33, 203]. To illustrate how peak attention measure can characterize variations in time series, consider the example of an event reportage marked by a sudden spike in collective attention followed by a subsequent drop. The lack of precursory growth suggests that most of the attention was concentrated on the peak, thereby resulting in high *peak attention* (Figure 18c and 18d). Whereas, an event with steady growth in collective attention, followed by a gradual decay would imply a relatively smaller fraction of attention in the peak, thus leading to lower value of *peak attention* (Figure 18b).

While *peak attention* captures the importance of the maximal burst in collective attention, it does not take into account the presence or absence of spikes in the precursory growth and in the subsequent decay following the burst. Hence, I define a measure to quantify the spikiness in collective attention. I detect all strict local maxima [177] in each of the event time series. A strict local maxima corresponds to an instance in the time series when the volume of messages (or unique people) is larger than the volume in the neighboring time windows. I define this neighborhood as three time windows on either side of the local

maxima and call these local maxima *minor peaks*. Thus, the attention in a minor peak is higher than the attention 30 minutes (i.e., three time windows times 10-minute window size) before and after the occurrence of a peak.

$$x_i \text{ is a minor peak iff } \begin{cases} x_i > x_{i+k} \\ x_i > x_{i-k} \end{cases}, k \in \{1, 2, 3\} \quad (3)$$

We then define minor peak attention as the ratio of messages (or unique people) in the local maxima relative to the total cumulative volume of messages (or unique people) over the entire event time series. Formally, if \mathcal{M} is the set of all minor peak indices in an event's message (or unique people) time series, then minor peak attention is defined as follows:

$$\text{Minor Peak Attention} = \frac{\sum_{j \in \mathcal{M}} x_j}{\sum_{i=1}^n x_i} \quad (4)$$

The inset diagram in Figure 16 shows a local maximum. While the peak attention captures the maximum momentary interest that an event acquires during its lifetime on Twitter, the points representing minor peak attention reflect renewed and ongoing recurrences of momentary interest. Additionally, both these measures have two important properties: both are invariant with respect to scaling and shifting [203]. First, since both measures are proportions based on cumulative collective attention, they are invariant to the overall volume of attention. Hence, two event time series having similar peaky shapes but different total attention volumes would be treated similarly. Secondly, both measures are computed independent of the maxima position on the time axis. Thus, if two event time series peaks occur at different times but possess a similar peaky structure, the measures will be invariant to the translations on the time axis. Hence, both these measures—despite being simple representations of temporal dynamics—are useful in interpreting the relationship between collective attention rhythms and event credibility across a range of different events exhibiting high variability in overall popularity and time of popularity.

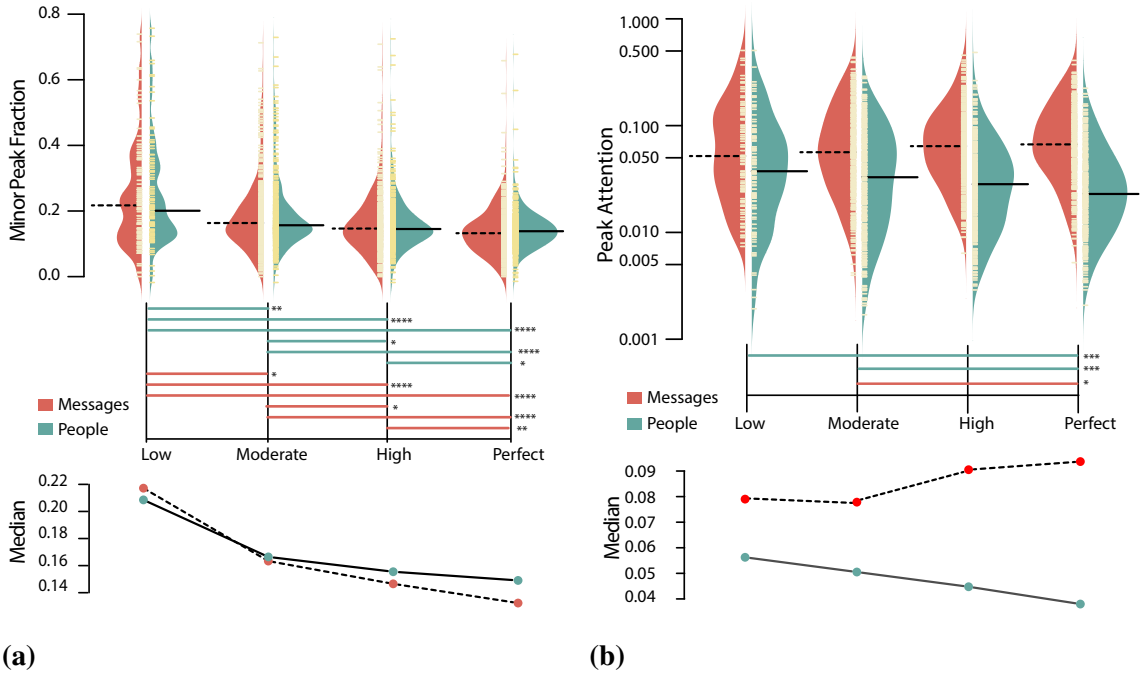


Figure 17: Collective attention shown as a beanplot distribution. The shape of each half of the asymmetric bean represents the Gaussian density estimation of the distribution. The lines (in yellow) are the actual data points; the dotted long bean line is the median corresponding to the message volume, the solid line shows the median for people volume. The * denotes pairwise significant differences between cluster medians after correcting for familywise error-rate. (a). Proportion of minor peak fractions are statistically different across all credibility class pairs for both message and unique people volume. (b). Peak attention is significantly different across “Low” and “Perfect”, and “Moderate” and “Perfect” credibility classes for unique people volume, and “Perfect” and “Moderate” classes for message volume. The line charts at the bottom panel show the median trends across the credibility classes.

6.3.3 Statistical Analysis and Results

I tested the differences in collective attention measures across the credibility classes using the Wilcoxon Rank Sum or Mann-Whitney U test. For each temporal measure (peak and minor peak attention) and for each collective attention metric (message and people volume), I performed pairwise Wilcoxon Rank Sum tests, followed by Bonferroni corrections [42] to control for potential inflation of the family-wise error rate by multiple test comparisons. I found that, for both *message volume* and *people volume*, differences in the minor peak fraction are statistically significant ($p < 0.00833$ after Bonferroni corrections and using Wilcoxon Rank Sum tests). As shown in Figure 17, median *minor peak attention* decreases as credibility level increases from “Low” to “Perfect”. I also found a significant moderate

degree of negative correlation between P_{ca} and minor peak fraction for both *message volume* ($r = -0.33$) and *people volume* ($r = -0.33$). These results suggest that an event attracting renewed interest is associated with lower perceived credibility. On the other hand, *peak attention* of messages was only statistically different between “Perfect” and “Moderate” credibility classes. Peak attention for *people volume* could only provide coarse-grained information separating “Low” and “Perfect”, and “Moderate” and “Perfect” credibility classes. These results indicate that *peak attention* is not a useful signal for event credibility. To ensure that these ratio-based, collective attention measures described above are not sensitive to event duration, which can affect the denominator (cumulative volume), I performed pairwise Wilcoxon Rank Sum test comparisons of event duration across the credibility classes. I found no significant difference, indicating that event duration does not skew the collective attention metrics for a particular credibility class. Moreover, to ensure that the collective attention metrics are independent observations over time—a criteria necessary for the validity of the statistical analysis—I performed Ljung-Box Q (LBQ) tests [108]. I was able to reject null hypothesis for every LBQ tests; thus confirming that the collective attention measures for both message and people volume are independent over time.

6.4 Discussion

By investigating the most comprehensive large-scale longitudinal credibility corpus constructed to date, I was able to test the relationship between an event’s perceived credibility level and the temporal dynamics of its collective attention. According to the findings of this study, moments of renewed collective attention are associated with event reportage marked by decreased levels of perceived credibility. Do frequent peaks in collective attention lead to lower perceived credibility? Or do reduced levels of credibility spark the continued interest in the event? The current study cannot establish the causal direction of this relation. However, I was able to establish that the persistence of collective attention peaks is a reliable temporal signature for an event’s perceived credibility level. Moreover, an advantage of

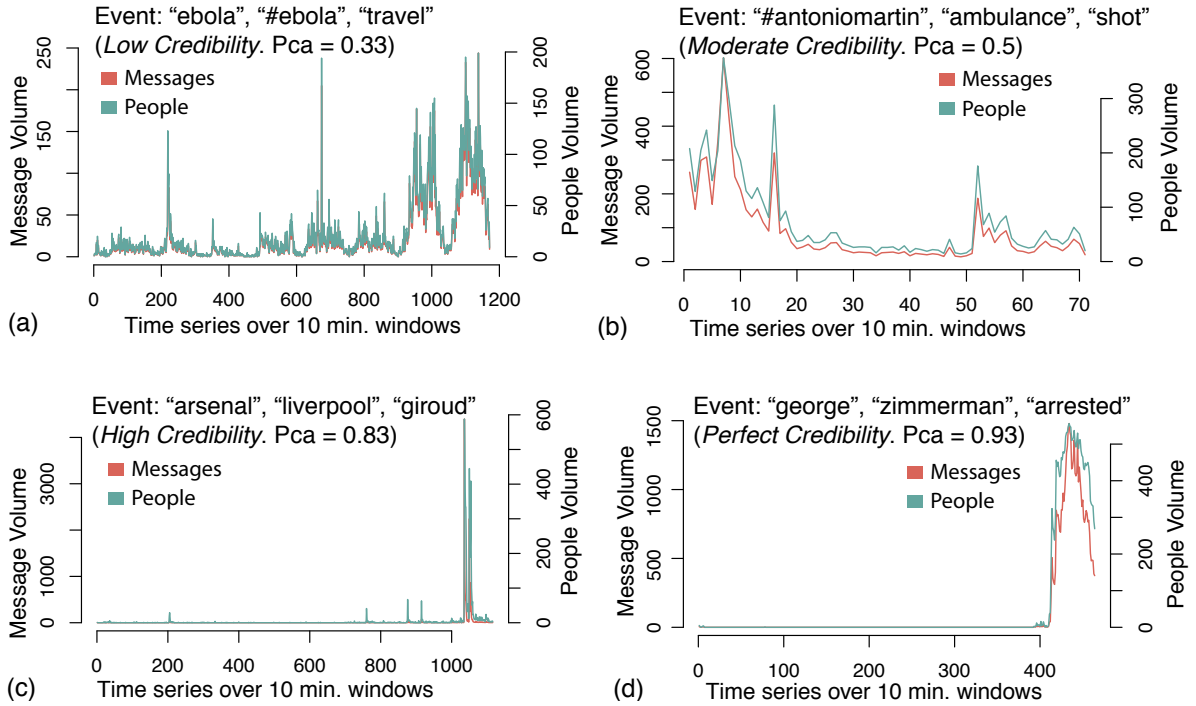


Figure 18: Time series of collective attention metrics (message volume and unique people volume) for example events in each credibility class. The examples show representative behavior of collective attention metrics in each credibility class. While events in all four classes are marked by peak attention with respect to both message and people volume, events in the low and moderate credibility classes exhibit multiple minor peaks, signifying that persistent attention is characteristic of lower credible social media events.

viewing these phenomena through the lens of a fundamental property of human activity, such as time, is that the resultant findings are likely to hold irrespective of the platform (e.g., Twitter) hosting the collective human attention directed toward real-world events.

Note that by using a simple proportion based classification technique, I identified robust and scalable credibility classes; hence it is also potentially applicable to other online settings where user’s collective attention drives popularity of content. Moreover, by using simple and interpretable parameters computed on times series of minute-wise user and message attention, I revealed vital temporal indicators associated with information credibility. Contrary to other sophisticated methods which require the estimation of power-law exponents for unraveling collective attention dynamics, or the calculation of costly correlations between activity time series, the parameters employed here can easily be computed in a scalable way. Although

Table 12: Pairwise statistical significance after Wilcoxon Rank Sum tests. P, H, M, L correspond to Perfect, High, Moderate and Low credibility classes. The top half of the diagonal corresponds to *message volume*, while bottom half shows pairwise differences in *people volume*. ns stands for non-significance.

	P	H	M	L
P		0.00012**	5.4e-12****	4.6e-11****
H	0.002639*		0.001703*	3.6e-06****
M	1.0e-08****	0.004221**		0.00547*
L	8.4e-11****	1.5e-06****	0.001981**	

(a) Minor peak attention pairwise statistical differences

	P	H	M	L
P		ns	0.00322*	ns
H	ns		ns	ns
M	0.00093****	ns		ns
L	0.00045****	ns	ns	

(b) Peak attention pairwise statistical differences

devoid of any predictive power, these measures can support the discovery of collective attention patterns in large-scale records of human activity.

On the basis of these results, I envision that organizations struggling to handle the propagation of online misinformation [109] can harness the temporality of collective attention to predict the level of credibility. We may be able to subsequently design interventions aimed at controlling the spread of false information or cautioning social media users to be skeptical about an evolving topic’s veracity, ultimately raising an individual’s capacity to assess credibility of information. Imagine a news reporting tool which shows social media discussions highlighting areas which witnessed multiple minor peaks of human activity, or think of a fact-checking system that compares temporal regions of high minor peak attentions to those with fewer attention peaks, or consider temporal tagging of scientific discourse or medical records emphasizing areas that garnered intermittent temporal popularity. I foresee that these findings can lead to a new class of such temporally aware systems which underscore degrees of information uncertainty based on temporal signals of collective attention. Finally, this

study has practical implications in the field of computational social science where inferences about human social behavior are based on reports of online interactions [32] and trusting the credibility of those reports is crucial for any downstream analysis. For example, imagine a health researcher investigating the spread of Ebola via social media reports or a financial trader gauging market volatility based on breaking news and citizen reports on social media; veracity of those reports will affect the subsequent inferences.

CHAPTER VII

LIMITATION, FUTURE WORK & CONCLUSIONS

We live in a media saturated era, where we constantly witness news and information from a wide variety of sources. To this end, social media with its ability to easily transmit information has empowered citizens in unprecedented ways. It has democratized the production and distribution of news and information at a scale which was unimaginable a few decades ago. This unparalleled information access, presents its own unique set of challenges. Instances of misinformation, disinformation, inaccurate reporting, fake news, phony news releases and a host of other information sharing malpractices are prevalent in social media platforms [131]. While it is beyond the scope of this thesis to unravel every layer and sub-layer of the misinformation ecosystem, I have approached the problem domain from the lens of information accuracy. My definition of accuracy purposely refrains from alluding to “epistemic truth”, rather I treat information accuracy as a socially driven construct. More concretely, in my study of social media credibility, I consider a stream of social post to be credible (or not) based on the collective human judgments of accuracy.

To ensure that collective judgments from a crowd of non-experts is as good as judgments made by expert fact-checkers, I deployed an exhaustive set of controlled experiments. My experiments compared the performance of non-experts and experts across a variety of tasks of varying levels of difficulty (including credibility assessment task). I deployed the experiments with different state-of-the-art crowd-sourcing strategies in order to finally converge to the best strategy for obtaining high-quality collective crowd annotations, that is, annotations which are at par with expert fact-checker judgments.

I used the learnings from this study to subsequently obtain event annotations (that is determining whether a topic on Twitter is a news topic or not) and credibility assessments

(that is determining the level of accuracy of a newsworthy topic on Twitter). My next goal was to design a framework to systematically track “all” events happening on Twitter (the specific social media platform used in this thesis) along with their level of accuracy. This was a necessary step so as to overcome *sampling bias* otherwise present in studies focused on investigating specific instances of popular rumors or post hoc investigation of prominent events with known disputed information. To address this challenge, I had to sift through massive amounts of social media posts, followed by a labor intensive task of content evaluation for credibility assessment. The effort resulted in a unique framework combining machine and human computation in a systematic way. Running the framework for more than three months, ultimately led to the creation of CREDBANK – the first large-scale social media corpus with associated credibility annotations.

One of the most basic questions which I could answer with CREDBANK’s dataset is the following: *how much misinformation is present in Twitter?* I found that roughly 24% of events in the global tweet stream cannot be perceived as credible. This number seemed rather large, and alludes to the ever increasing problem of social media turning to a “cesspool” of inaccurate information.

But is there any glimmer of hope? While social media’s ease of information flow is a problem, it can also be a solution. When information with dubious claims are shared in our social network, some of our social connections also question that information, or express doubt by using *hedge* words or laugh at the irrationality of statements or even out-rightly deny the claims. Can we determine the credibility of information using these linguistic signals? What about the speed with which people pay attention to certain newsworthy topics? Are there systematic differences in the way a low versus a high credible event trend on social media. In Chapters 5 and 6, I unravel these questions. In Chapter 5, I uncovered words and phrases which are associated with low versus high credible events. By identifying 15 theoretically grounded linguistic dimensions spread across thousands of words and phrases, I developed a parsimonious model working on millions of tweets

corresponding to thousands of events and their corresponding credibility annotations. The model maps language cues to perceived levels of credibility. While not deployable as a standalone model for credibility assessment at present, these results show that certain linguistic categories and their associated phrases are strong predictors surrounding disparate social media events. In other words, the language used by millions of people on Twitter has considerable information about an event's credibility. For example, hedge words and positive emotion words are associated with lower credibility. In Chapter 6, I presented the dynamics of collective attention and its relation to information credibility by analyzing the temporal patterns of millions of tweets from the CREDBANK corpus. I performed multiple statistical comparison tests over parameters computed on the time series of collective attention of messages and distinct users. Although simple, this approach provides fundamental insights about collective attention and information credibility that would otherwise be missed by more complicated predictive analysis methods.

7.1 Limitations

There are many possible ways to model credibility. I have considered only the linguistic and temporal constructs. We could do more and we could do better. What about the structural signals? As I pointed out earlier, I faced a practical challenge in fetching the social graphs of every user who are present in my CREDBANK dataset. But with more resources and access to Twitter's firehose, it is possible to include additional structural signals.

Do the results generalize? I have taken a deep dive to study credibility in one specific social medium – Twitter. I have no way to empirically answer the question of generalizability. How does credibility look like in Facebook? What happens when the platform is Reddit? One way to find answers to these questions will be to repeat the studies in multiple platforms. I envision that the methods from my thesis can lay the groundwork for these subsequent studies.

What happens when Twitter changes its design? Will the results still be valid? Social

systems continually change in design. In 2016, Twitter adopted an algorithmic timeline, forcing a non-chronological selection of tweets in a user’s feed. I built CREDBANK before this major design change. This change will likely alter the way users are obtaining news and information through Twitter. Removing chronological ordering changes the very definition of “current” news and “breaking” stories. To what extent will our findings be resilient to design changes? I think this is a question which every social media researcher struggles with and needs significant future work.

7.2 Future Research Directions

There are multiple directions that can stem from this work. Throughout my work, I have focused on understanding and modeling credibility. One can take this understanding to create design elements that address the misinformation ecosystem. Here I present a few scenarios.

7.2.1 Promoting reflective news reading

Imagine you are going through your Twitter feed and the system highlights tweets where people have expressed uncertainty. A naive version of this design would be a scenario where all the hedge words in the Twitter replies (or Facebook comments) are highlighted. Recall, that our language model, showed that hedging in replies was associated with lower credible events. My position here is that if we want to make progress on addressing the problem of misinformation and in shaking people from their preexisting ideological echo chambers, we should try to support reflective and deliberative styles of interactions, which are controlled by the user and which are more simply integrated with existing social media platforms. This serves three purposes. First, the user does not get the sense that her ideologies are being explicitly challenged. She can use the tool the way she wants. Second, by piggybacking off of existing social media platforms, we can leverage users’ existing social environments on Twitter and Facebook to explore novel interactions around news reading habits and nudge them towards more informed information reading habits. Third, by integrating with existing

social media platforms, we lower any overhead on the part of the user to learn how to use our tool.

7.2.2 Ordering newsfeed by information quality

Biased information, misleading facts or outright rumors are inherently sensational in nature. They easily pick up a lot of steam and in the process amplify the damage caused by inaccurate reporting. But the more attention a topic gets (irrespective of whether it is accurate or fake), the better it is for social media companies. It directly feeds into their revenue model. But this indifference to information quality has pushed our society to the so-called “post-truth” era. What if current social-feed algorithms bubble up high-quality information, followed by things which are disputed and questioned? This design suggestion seems counter-intuitive – why not show disputed information first so that people fact-check? This is a valid suggestion, but is it the right approach, because as soon as disputed information is made visible, the non-factual information also starts to get more attention.

7.2.3 Bias aware social systems

While pilot testing CREDBANK’s pipeline, I found that when making credibility judgments, crowd worker’s might have certain biases in their decision making process. By aggregating responses from a large crowd, I ensured that the overall annotation is not effected by these one-off biases. There are still limitations of this approach. It is almost impossible to completely remove bias from any social decision making process. But can we at least make users aware of their underlying biases? When users are interacting with their social media feeds, their individual biases are constantly guiding their judgments. Current social media systems do not provide any indication to the user about his/her underlying biases. Are you only reading news from non-credible sources? Are you only reading left leaning or right leaning news? Are you retweeting or sharing information which have been questioned by others, perhaps even by people outside your immediate social connections or people with different ideological views? An algorithm can determine if the information source

is disputed. Based on the analysis I presented in chapters 5 and 6, the algorithm can also predict the level of accuracy. I envision these signals can live as a tool tip on the user interface, or as a floating window for a short time, allowing users to dismiss it if they don't care.

While I have outlined three design spaces, we can use computationally inferred credibility in multiple domains (e.g. search engines to mark if the content is disputed, questionable or well-established). We can design better models which can predict credibility levels with higher accuracy. We can also include additional signals, such as, structural features or event type (breaking vs. ongoing), event topic (sports vs. politics), or interactions among these individual features. I leave that to future work.

REFERENCES

- [1] “Cleantweets.” <https://addons.mozilla.org/en-US/firefox/addon/clean-tweets/>.
- [2] “Sec charges: False tweets sent two stocks reeling in market manipulation.” <https://www.sec.gov/news/pressrelease/2015-254.html>. Accessed: 2017-04-04.
- [3] “Spam words by wordpress.” http://codex.wordpress.org/Spam_Words.
- [4] ADAR, E., ZHANG, L., ADAMIC, L. A., and LUKOSE, R. M., “Implicit structure and the dynamics of blogspace,” in *Workshop on the weblogging ecosystem*, vol. 13, pp. 16989–16995, 2004.
- [5] ALLAN, J., “Introduction to topic detection and tracking,” in *Topic detection and tracking*, pp. 1–16, Springer, 2002.
- [6] ALSUMAIT, L., BARBARÁ, D., and DOMENICONI, C., “On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking,” in *Proc. ICDM*, 2008.
- [7] ANDRÉ, P., KITTUR, A., and DOW, S. P., “Crowd synthesis: Extracting categories and clusters from complex data,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 989–998, ACM, 2014.
- [8] ARIF, A., SHANAHAN, K., CHOU, F.-J., DOSOUTO, Y., STARBIRD, K., and SPIRO, E. S., “How information snowballs: Exploring the role of exposure in online rumor propagation,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 466–477, ACM, 2016.
- [9] ASUR, S., HUBERMAN, B. A., SZABO, G., and WANG, C., “Trends in social media: Persistence and decay,” *Available at SSRN Scholarly Paper ID 1755748, Social Science Research Network.*, 2011.
- [10] BAKSHY, E., HOFMAN, J. M., MASON, W. A., and WATTS, D. J., “Everyone’s an influencer: quantifying influence on twitter,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 65–74, ACM, 2011.
- [11] BANFIELD, A., “Unspeakable sentences,” 1982.
- [12] BARABASI, A.-L., “The origin of bursts and heavy tails in human dynamics,” *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.
- [13] BASS, M., “Frank,.” *A New Product Growth for Model Consumer Durables*, vol. 50, pp. 1825–1832, 1969.

- [14] BECKER, H., NAAMAN, M., and GRAVANO, L., “Beyond trending topics: Real-world event identification on twitter,” in *Proc. ICWSM*, 2011.
- [15] BERGLER, S., DOANDES, M., GERARD, C., and WITTE, R., “Attributions,” *Exploring Attitude and Affect in Text: Theories and Applications, Technical Report SS-04-07*, pp. 16–19, 2004.
- [16] BERINSKY, A. J., HUBER, G. A., and LENZ, G. S., “Evaluating online labor markets for experimental research: Amazon.com’s mechanical turk,” *Political Analysis*, vol. 20, no. 3, pp. 351–368, 2012.
- [17] BLEI, D. M., NG, A. Y., and JORDAN, M. I., “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
- [18] BLEI, D. M., NG, A. Y., and JORDAN, M. I., “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [19] BORDIA, P. and DiFONZO, N., “Problem solving in social interactions on the internet: Rumor as social cognition,” *Social Psychology Quarterly*, vol. 67, no. 1, pp. 33–49, 2004.
- [20] BORDIA, P. and ROSNOW, R. L., “Rumor rest stops on the information highway transmission patterns in a computer-mediated rumor chain,” *Human Communication Research*, vol. 25, no. 2, pp. 163–179, 1998.
- [21] BURKE, M., ADAMIC, L., and MARCINIAK, K., “Families on facebook,” in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.
- [22] BYBEE, J., PERKINS, R., and PAGLIUCA, W., *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. University of Chicago Press, 1994.
- [23] CASTILLO, C., MENDOZA, M., and POBLETE, B., “Information credibility on twitter,” in *Proceedings of the 20th international conference on World wide web*, pp. 675–684, ACM, 2011.
- [24] CASTILLO, C., MENDOZA, M., and POBLETE, B., “Information credibility on twitter,” in *Proc. WWW*, 2011.
- [25] CAUMONT, A., “12 trends shaping digital news. <http://www.pewresearch.org/fact-tank/2013/10/16/12-trends-shaping-digital-news/>. Accessed June 6, 2014.,” *Pew Research*, 2013.
- [26] CHAFFEE, S. H., “Mass media and interpersonal channels: Competitive, convergent, or complementary,” *Inter/media: Interpersonal communication in a media world*, pp. 57–77, 1982.
- [27] CHANG, J., BOYD-GRABER, J. L., GERRISH, S., WANG, C., and BLEI, D. M., “Reading tea leaves: How humans interpret topic models.,” in *Nips*, vol. 31, pp. 1–9, 2009.

- [28] CHU, Z., GIANVECCHIO, S., WANG, H., and JAJODIA, S., “Who is tweeting on twitter: human, bot, or cyborg?,” in *Proc. ACSAC*, pp. 21–30, ACM, 2010.
- [29] CHU, Z., GIANVECCHIO, S., WANG, H., and JAJODIA, S., “Detecting automation of twitter accounts: Are you a human, bot, or cyborg?,” *IEEE Trans. Dependable Sec. Comput.*, vol. 9, no. 6, pp. 811–824, 2012.
- [30] COHEN, J., “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.,” *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.
- [31] CORCORAN, M., “Death by cliff plunge, with a push from twitter,” *The New York Times*, 2009.
- [32] COUNTS, S., DE CHOUDHURY, M., DIESNER, J., GILBERT, E., GONZALEZ, M., KEEGAN, B., NAAMAN, M., and WALLACH, H., “Computational social science: CSCW in the social media era,” in *Proc. CSCW Companion publication*, pp. 105–108, ACM, 2014.
- [33] CRANE, R. and SORNETTE, D., “Robust dynamic classes revealed by measuring the response function of a social system,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 41, pp. 15649–15653, 2008.
- [34] CRUMP, M. J., McDONNELL, J. V., and GURECKIS, T. M., “Evaluating amazon’s mechanical turk as a tool for experimental behavioral research,” *PloS one*, vol. 8, no. 3, p. e57410, 2013.
- [35] CULOTTA, A., “Towards detecting influenza epidemics by analyzing twitter messages,” in *Proceedings of the first workshop on social media analytics*, 2010.
- [36] DALAL, N. and TRIGGS, B., “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.
- [37] DE MARNEFFE, M., MANNING, C. D., and POTTS, C., “Veridicality and utterance understanding,” in *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pp. 430–437, IEEE, 2011.
- [38] DE MARNEFFE, M.-C., MANNING, C. D., and POTTS, C., “Did it happen? the pragmatic complexity of veridicality assessment,” *Computational linguistics*, vol. 38, no. 2, pp. 301–333, 2012.
- [39] DEL VICARIO, M., BESSI, A., ZOLLO, F., PETRONI, F., SCALA, A., CALDARELLI, G., STANLEY, H. E., and QUATTROCIOCCI, W., “The spreading of misinformation online,” *PNAS*, vol. 113, no. 3, pp. 554–559, 2016.
- [40] DIJKSTRA, J. J., LIEBRAND, W. B., and TIMMINGA, E., “Persuasiveness of expert systems,” *Behaviour & Information Technology*, vol. 17, no. 3, pp. 155–163, 1998.

- [41] DOWNS, J. S., HOLBROOK, M. B., SHENG, S., and CRANOR, L. F., “Are your participants gaming the system?: screening mechanical turk workers,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2399–2402, ACM, 2010.
- [42] DUNN, O. J., “Estimation of the medians for dependent variables,” *The Annals of Mathematical Statistics*, pp. 192–197, 1959.
- [43] EISENSTEIN, J., SMITH, N. A., and XING, E. P., “Discovering sociolinguistic associations with structured sparsity,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 1365–1374, ACL, 2011.
- [44] ESFANDIARI, G., “The twitter devolution,” *Foreign Policy*, vol. 7, p. 2010, 2010.
- [45] FLANAGIN, A. J. and METZGER, M. J., “Digital media and youth: Unparalleled opportunity and unprecedented responsibility,” *Digital media, youth, and credibility*, pp. 5–27, 2008.
- [46] FOGG, B., SOOHOO, C., DANIELSON, D. R., MARABLE, L., STANFORD, J., and TAUBER, E. R., “How do users evaluate the credibility of web sites?: a study with over 2,500 participants,” in *Proceedings of the 2003 conference on Designing for user experiences*, pp. 1–15, ACM, 2003.
- [47] FOGG, B. and TSENG, H., “The elements of computer credibility,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 80–87, ACM, 1999.
- [48] FRIEDMAN, J., HASTIE, T., and TIBSHIRANI, R., “Regularization paths for generalized linear models via coordinate descent,” *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.
- [49] FRIGGERI, A., ADAMIC, L., ECKLES, D., and CHENG, J., “Rumor cascades,” in *Proc. ICWSM*, 2014.
- [50] FRIZELL, S., “3 Muslim Students Murdered in North Carolina.” time.com/3704759/muslim-students-murdered-chapel-hill, 2015.
- [51] GILBERT, E., “Phrases that signal workplace hierarchy,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 1037–1046, ACM, 2012.
- [52] GILBERT, E., “What if we ask a different question?: social inferences create product ratings faster,” in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 2759–2762, ACM, 2014.
- [53] GOEL, S., HOFMAN, J. M., LAHAIE, S., PENNOCK, D. M., and WATTS, D. J., “Predicting consumer behavior with web search,” *PNAS*, vol. 107, no. 41, pp. 17486–17490, 2010.

- [54] GOLDMAN, A. and BLANCHARD, T., “Social epistemology,” in *The Stanford Encyclopedia of Philosophy* (ZALTA, E. N., ed.), Metaphysics Research Lab, Stanford University, winter 2016 ed., 2016.
- [55] GOTTFRIED, J. and SHEARER, E., “News Use Across Social Media Platforms 2016. <http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/>. Accessed April 3, 2017.,” *Pew Research*, 2016.
- [56] GRANOVETTER, M. S., “The strength of weak ties,” *American journal of sociology*, pp. 1360–1380, 1973.
- [57] GRIFFITHS, T. L. and STEYVERS, M., “Finding scientific topics,” *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.
- [58] GROVER, R., “Ad. ly: The art of advertising on twitter,” *Businessweek, January*, vol. 6, 2011.
- [59] GRUHL, D., GUHA, R., LIBEN-NOWELL, D., and TOMKINS, A., “Information diffusion through blogspace,” in *Proceedings of the 13th international conference on World Wide Web*, pp. 491–501, ACM, 2004.
- [60] GUPTA, A. and KUMARAGURU, P., “Credibility ranking of tweets during high impact events,” in *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, p. 2, ACM, 2012.
- [61] GUPTA, A., LAMBA, H., and KUMARAGURU, P., “\$1.00 per rt# bostonmarathon# pray-forboston: Analyzing fake content on twitter,” in *eCrime Researchers Summit, 2013*, pp. 1–12, IEEE, 2013.
- [62] GUPTA, A., LAMBA, H., KUMARAGURU, P., and JOSHI, A., “Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy,” in *Proc. WWW companion*, 2013.
- [63] HAIGHT, F. A. and HAIGHT, F. A., “Handbook of the poisson distribution,” 1967.
- [64] HANCOCK, J. T., CURRY, L. E., GOORHA, S., and WOODWORTH, M., “On lying and being lied to: A linguistic analysis of deception in computer-mediated communication,” *Discourse Processes*, vol. 45, no. 1, pp. 1–23, 2007.
- [65] HART, S. G. and STAVELAND, L. E., “Development of nasa-tlx (task load index): Results of empirical and theoretical research,” *Advances in psychology*, vol. 52, pp. 139–183, 1988.
- [66] HEER, J. and BOSTOCK, M., “Crowdsourcing graphical perception: using mechanical turk to assess visualization design,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 203–212, ACM, 2010.

- [67] HILLIGOSS, B. and RIEH, S. Y., “Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context,” *Information Processing & Management*, vol. 44, no. 4, pp. 1467–1484, 2008.
- [68] HOFFMAN, M., BACH, F. R., and BLEI, D. M., “Online learning for latent dirichlet allocation,” in *advances in neural information processing systems*, pp. 856–864, 2010.
- [69] HOVLAND, C. I., JANIS, I. L., and KELLEY, H. H., “Communication and persuasion; psychological studies of opinion change.” 1953.
- [70] HU, M., LIU, S., WEI, F., WU, Y., STASKO, J., and MA, K.-L., “Breaking news on twitter,” in *Proc. CHI*, pp. 2751–2754, ACM, 2012.
- [71] HUANG, S.-W. and FU, W.-T., “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*, pp. 621–630, ACM, 2013.
- [72] HUTTO, C. J. and GILBERT, E., “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International AAI Conference on Weblogs and Social Media*, 2014.
- [73] HYLAND, K., *Hedging in scientific research articles*, vol. 54. John Benjamins Publishing, 1998.
- [74] HYLAND, K., “Authority and invisibility: Authorial identity in academic writing,” *Journal of pragmatics*, vol. 34, no. 8, pp. 1091–1112, 2002.
- [75] HYLAND, K., *Metadiscourse: Exploring interaction in writing*. A&C Black, 2005.
- [76] IPEIROTIS, P. G., “Analyzing the amazon mechanical turk marketplace,” *XRDS: Crossroads, The ACM Magazine for Students*, vol. 17, no. 2, pp. 16–21, 2010.
- [77] IPEIROTIS, P. G., PROVOST, F., and WANG, J., “Quality management on amazon mechanical turk,” in *Proceedings of the ACM SIGKDD workshop on human computation*, pp. 64–67, ACM, 2010.
- [78] JOHNSON-LAIRD, P., GAWRONSKI, B., and STRACK, F., “Mental models and consistency,” *Cognitive consistency: A fundamental principle in social cognition*, pp. 225–244, 2012.
- [79] KANALLEY, C., “Facebook shutting down rumor goes viral: Site said to be ending march 15, 2011,” *The Huffington Post*, 2011.
- [80] KARTTUNEN, L. and ZAENEN, A., “Veridicity,” *Annotating, extracting and reasoning about time and events*, no. 05151.

- [81] KAZAI, G., KAMPS, J., and MILIC-FRAYLING, N., “Worker types and personality traits in crowdsourcing relevance labels,” in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1941–1944, ACM, 2011.
- [82] KIM, R. K. and LEVINE, T. R., “The effect of suspicion on deception detection accuracy: Optimal level or opposing effects?,” *Communication Reports*, vol. 24, no. 2, pp. 51–62, 2011.
- [83] KIPARSKY, P.-K., “Carol (1970), ‘fact’,” *Bierwisch, Manfred-Heidolph, Karl Erich (a cura di), Progress in Linguistics (A Collection of Papers), The Hague, Mouton*, pp. 143–173, 1971.
- [84] KITTUR, A., CHI, E. H., and SUH, B., “Crowdsourcing user studies with mechanical turk,” in *Proc. CHI*, 2008.
- [85] KITTUR, A., CHI, E. H., and SUH, B., “Crowdsourcing user studies with mechanical turk,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 453–456, ACM, 2008.
- [86] KOVACH, B. and ROSENSTIEL, T., *The elements of journalism: What newspeople should know and the public should expect*. Three Rivers Press, 2014.
- [87] KRIPLEAN, T., BONNAR, C., BORNING, A., KINNEY, B., and GILL, B., “Integrating on-demand fact-checking with public dialogue,” in *Proc. CSCW*, 2014.
- [88] KRIPLEAN, T., BONNAR, C., BORNING, A., KINNEY, B., and GILL, B., “Integrating on-demand fact-checking with public dialogue,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 1188–1199, ACM, 2014.
- [89] KWAK, H., LEE, C., PARK, H., and MOON, S., “What is twitter, a social network or a news media?,” in *Proc. WWW*, pp. 591–600, ACM, 2010.
- [90] KWAK, H., LEE, C., PARK, H., and MOON, S., “What is twitter, a social network or a news media?,” in *Proc. WWW*, pp. 591–600, ACM, 2010.
- [91] KWON, S., CHA, M., JUNG, K., CHEN, W., and WANG, Y., “Aspects of rumor spreading on a microblog network,” in *Social Informatics*, pp. 299–308, Springer, 2013.
- [92] KWON, S., CHA, M., JUNG, K., CHEN, W., and WANG, Y., “Prominent features of rumor propagation in online social media,” in *Proc. ICDM*, 2013.
- [93] LAKOFF, G., *Hedges: a study in meaning criteria and the logic of fuzzy concepts*. Springer, 1975.
- [94] LAKOFF, G., *Women, fire, and dangerous things*. University of Chicago press, 2008.

- [95] LASECKI, W. S., GORDON, M., KOUTRA, D., JUNG, M. F., DOW, S. P., and BIGHAM, J. P., “Glance: Rapidly coding behavioral video with the crowd,” in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pp. 551–562, ACM, 2014.
- [96] LATOUR, B. and WOOLGAR, S., *Laboratory life: The construction of scientific facts*. Princeton University Press, 2013.
- [97] LAU, J. H., COLLIER, N., and BALDWIN, T., “On-line trend analysis with topic models:\# twitter trends detection topic model online.,” in *COLING*, pp. 1519–1534, 2012.
- [98] LAU, J. H., COLLIER, N., and BALDWIN, T., “On-line trend analysis with topic models:\# twitter trends detection topic model online.,” in *COLING*, pp. 1519–1534, 2012.
- [99] LEHMANN, J., GONÇALVES, B., RAMASCO, J. J., and CATTUTO, C., “Dynamical classes of collective attention in twitter,” in *Proc. WWW*, pp. 251–260, ACM, 2012.
- [100] LESKOVEC, J., ADAMIC, L. A., and HUBERMAN, B. A., “The dynamics of viral marketing,” *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, p. 5, 2007.
- [101] LESKOVEC, J., BACKSTROM, L., and KLEINBERG, J., “Meme-tracking and the dynamics of the news cycle,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506, ACM, 2009.
- [102] LEWANDOWSKY, S., ECKER, U. K., SEIFERT, C. M., SCHWARZ, N., and COOK, J., “Misinformation and its correction continued influence and successful debiasing,” *Psychological Science in the Public Interest*, vol. 13, no. 3, pp. 106–131, 2012.
- [103] LEWIS, D. D., YANG, Y., ROSE, T. G., and LI, F., “Rcv1: A new benchmark collection for text categorization research,” *The Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.
- [104] LI, C., SUN, A., and DATTA, A., “Twevent: segment-based event detection from tweets,” in *Proc. CIKM*, pp. 155–164, 2012.
- [105] LIAO, Q. and SHI, L., “She gets a sports car from our donation: rumor transmission in a chinese microblogging community,” in *Proc. CSCW*, pp. 587–598, ACM, 2013.
- [106] LITTLE, G., CHILTON, L. B., GOLDMAN, M., and MILLER, R. C., “Exploring iterative and parallel human computation processes,” in *Proceedings of the ACM SIGKDD workshop on human computation*, pp. 68–76, ACM, 2010.
- [107] LIU, F., BURTON-JONES, A., and XU, D., “Rumors on social media in disasters: Extending transmission to retransmission,” in *Proc. PACIS*, 2014.
- [108] LJUNG, G. M. and BOX, G. E., “On a measure of lack of fit in time series models,” *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978.
- [109] LUCKERSON, V., “Fear, Misinformation, and Social Media Complicate Ebola Fight. Time Inc.” time.com/3479254/ebola-social-media/, 2014.

- [110] MACDONALD, P. L. and GARDNER, R. C., “Type i error rate comparisons of post hoc procedures for i j chi-square tables,” *Educational and Psychological Measurement*, vol. 60, no. 5, pp. 735–754, 2000.
- [111] MADDOCK, J., STARBIRD, K., AL-HASSANI, H. J., SANDOVAL, D. E., ORAND, M., and MASON, R. M., “Characterizing online rumoring behavior using multi-dimensional signatures,” in *Proc. CSCW*, pp. 228–241, ACM, 2015.
- [112] MAIMON, O. and ROKACH, L., *Data mining and knowledge discovery handbook*, vol. 2. Springer, 2005.
- [113] MAKREHCHI, M. and KAMEL, M. S., “Automatic extraction of domain-specific stop-words from labeled documents,” in *Advances in information retrieval*, pp. 222–233, Springer, 2008.
- [114] MASON, W. and SURI, S., “Conducting behavioral research on amazon’s mechanical turk,” *Behavior research methods*, vol. 44, no. 1, pp. 1–23, 2012.
- [115] MASON, W. and WATTS, D. J., “Financial incentives and the performance of crowds,” *ACM SigKDD Explorations Newsletter*, vol. 11, no. 2, pp. 100–108, 2010.
- [116] MATHIOUDAKIS, M. and KOUDAS, N., “Twittermonitor: trend detection over the twitter stream,” in *Proc. SIGMOD*, pp. 1155–1158, ACM, 2010.
- [117] MATTHEWS, C., “How does one fake tweet cause a stock market crash?,” <http://business.time.com/2013/04/24/how-does-one-fake-tweet-cause-a-stock-market-crash/>. Accessed: 2017-04-04.
- [118] MENDOZA, M., POBLETE, B., and CASTILLO, C., “Twitter under crisis: Can we trust what we rt?,” in *Proceedings of the first workshop on social media analytics*, pp. 71–79, ACM, 2010.
- [119] METZGER, M. J., FLANAGIN, A. J., EYAL, K., LEMUS, D. R., and McCANN, R. M., “Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment,” *Annals of the International Communication Association*, vol. 27, no. 1, pp. 293–335, 2003.
- [120] MITCHELL, T. F. and ?ASAN, S., *Modality, Mood, and Aspect in Spoken Arabic: With Special Reference to Egypt and the Levant*, vol. 11. Routledge, 1994.
- [121] MITRA, T. and GILBERT, E., “Credbank: A large-scale social media corpus with associated credibility annotations,” in *Proc. ICWSM*, 2015.
- [122] MITRA, T., HUTTO, C., and GILBERT, E., “Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk,” In Submission.
- [123] MORRIS, M. R., COUNTS, S., ROSEWAY, A., HOFF, A., and SCHWARZ, J., “Tweeting is believing?: understanding microblog credibility perceptions,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 441–450, ACM, 2012.

- [124] MORRIS, M. R., COUNTS, S., ROSEWAY, A., HOFF, A., and SCHWARZ, J., “Tweeting is believing?: understanding microblog credibility perceptions,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 441–450, ACM, 2012.
- [125] MORRIS, M. R., TEEVAN, J., and PANOVICH, K., “What do people ask their social networks, and why?: a survey study of status message q&a behavior,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1739–1748, ACM, 2010.
- [126] MUSTAFARAJ, E. and METAXAS, P. T., “From obscurity to prominence in minutes: Political speech and real-time search,” 2010.
- [127] NAAMAN, M., BOASE, J., and LAI, C.-H., “Is it really about me?: message content in social awareness streams,” in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 189–192, ACM, 2010.
- [128] NEWMAN, M. L., PENNEBAKER, J. W., BERRY, D. S., and RICHARDS, J. M., “Lying words: Predicting deception from linguistic styles,” *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.
- [129] NEWMAN, N., “Mainstream media and the distribution of news in the age of social discovery,” *Reuters Institute for the Study of Journalism, University of Oxford*, 2011.
- [130] O’CONNOR, B., KRIEGER, M., and AHN, D., “Tweetmotif: Exploratory search and topic summarization for twitter,” in *Proc. ICWSM*, 2010.
- [131] O’CONNOR, R., “Word of mouse: Credibility, journalism and emerging social media,” 2009.
- [132] OSBORNE, M., PETROVIC, S., MCCREADIE, R., MACDONALD, C., and OUNIS, I., “Bieber no more: First story detection using twitter and wikipedia,” in *Proceedings of the Workshop on Time-aware Information Access. TAIA*, 2012.
- [133] OWOPUTI, O., O’CONNOR, B., DYER, C., GIMPEL, K., SCHNEIDER, N., and SMITH, N. A., “Improved part-of-speech tagging for online conversational text with word clusters,” in *HLT-NAACL*, 2013.
- [134] PAOLACCI, G., CHANDLER, J., and IPEIROTIS, P. G., “Running experiments on amazon mechanical turk,” 2010.
- [135] PAPAGEORGIOU, C. and POGGIO, T., “A trainable system for object detection,” *International journal of computer vision*, vol. 38, no. 1, pp. 15–33, 2000.
- [136] PARSONS, T., “The traditional square of opposition,” 1999.
- [137] PEER, E., VOSGERAU, J., and ACQUISTI, A., “Reputation as a sufficient condition for data quality on amazon mechanical turk,” *Behavior research methods*, vol. 46, no. 4, pp. 1023–1031, 2014.

- [138] PETROVIĆ, S., OSBORNE, M., and LAVRENKO, V., “Streaming first story detection with application to twitter,” in *Proc. HLT-NAACL*, 2010.
- [139] PRELEC, D., “A bayesian truth serum for subjective data,” *science*, vol. 306, no. 5695, pp. 462–466, 2004.
- [140] PRIEDHORSKY, R., CULOTTA, A., and DEL VALLE, S. Y., “Inferring the origin locations of tweets with quantitative confidence,” in *Proc. CSCW*, 2014.
- [141] PUSTEJOVSKY, J., HANKS, P., SAURI, R., SEE, A., GAIZAUSKAS, R., SETZER, A., RADEV, D., SUNDHEIM, B., DAY, D., FERRO, L., and OTHERS, “The timebank corpus,” in *Corpus linguistics*, vol. 2003, p. 40, 2003.
- [142] QAZVINIAN, V., ROSENGREN, E., RADEV, D. R., and MEI, Q., “Rumor has it: Identifying misinformation in microblogs,” in *Proc. EMNLP*, 2011.
- [143] QAZVINIAN, V., ROSENGREN, E., RADEV, D. R., and MEI, Q., “Rumor has it: Identifying misinformation in microblogs,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1589–1599, Association for Computational Linguistics, 2011.
- [144] RASHTCHIAN, C., YOUNG, P., HODOSH, M., and HOCKENMAIER, J., “Collecting image annotations using amazon’s mechanical turk,” in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pp. 139–147, Association for Computational Linguistics, 2010.
- [145] RATKIEWICZ, J., CONOVER, M., MEISS, M., GONÇALVES, B., PATIL, S., FLAMMINI, A., and MENCZER, F., “Truthy: mapping the spread of astroturf in microblog streams,” in *Proceedings of the 20th international conference companion on World wide web*, pp. 249–252, ACM, 2011.
- [146] RATKIEWICZ, J., FORTUNATO, S., FLAMMINI, A., MENCZER, F., and VESPIGNANI, A., “Characterizing and modeling the dynamics of online popularity,” *Physical review letters*, vol. 105, no. 15, p. 158701, 2010.
- [147] RESNICK, P., CARTON, S., PARK, S., SHEN, Y., and ZEFFER, N., “Rumorlens: A system for analyzing the impact of rumors and corrections in social media,” in *Symposium on Computation + Journalism*, 2014.
- [148] RIEH, S. Y. and DANIELSON, D. R., “Credibility: A multidisciplinary framework,” *Annual review of information science and technology*, vol. 41, no. 1, pp. 307–364, 2007.
- [149] ROMERO, D. M., MEEDER, B., and KLEINBERG, J., “Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter,” in *Proc. WWW*, pp. 695–704, ACM, 2011.
- [150] ROSNOW, R. L., “Rumor as communication: A contextualist approach,” *Journal of Communication*, vol. 38, no. 1, pp. 12–28, 1988.

- [151] ROSNOW, R. L., “Inside rumor: A personal journey.,” *American Psychologist*, vol. 46, no. 5, p. 484, 1991.
- [152] RUBIN, K., “The ultimate list of email spam trigger words.” <http://blog.hubspot.com/blog/tabid/6307/bid/30684/The-Ultimate-List-of-Email-SPAM-Trigger-Words.aspx>.
- [153] RUBIN, V. L., LIDDY, E. D., and KANDO, N., “Certainty identification in texts: Categorization model and manual tagging results,” in *Computing attitude and affect in text: Theory and applications*, pp. 61–76, Springer, 2006.
- [154] SAIF, H., FERNÁNDEZ, M., HE, Y., and ALANI, H., “On stopwords, filtering and data sparsity for sentiment analysis of twitter,” in *LREC*, pp. 810–817, 2014.
- [155] SALDAÑA, J., *The coding manual for qualitative researchers*. Sage, 2015.
- [156] SALGANIK, M. J., DODDS, P. S., and WATTS, D. J., “Experimental study of inequality and unpredictability in an artificial cultural market,” *Science*, vol. 311, no. 5762, pp. 854–856, 2006.
- [157] SARTER, N. B. and WOODS, D. D., “Situation awareness: A critical but ill-defined phenomenon,” *The International Journal of Aviation Psychology*, vol. 1, no. 1, pp. 45–57, 1991.
- [158] SASAHARA, K., HIRATA, Y., TOYODA, M., KITSUREGAWA, M., and AIHARA, K., “Quantifying collective attention from tweet stream,” *PloS one*, vol. 8, no. 4, p. e61823, 2013.
- [159] SAURI, R., *A factuality profiler for eventualities in text*. ProQuest, 2008.
- [160] SAURÍ, R., KNIPPEN, R., VERHAGEN, M., and PUSTEJOVSKY, J., “Evita: a robust event recognizer for qa systems,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 700–707, Association for Computational Linguistics, 2005.
- [161] SAURÍ, R. and PUSTEJOVSKY, J., “Factbank: A corpus annotated with event factuality,” *Language resources and evaluation*, vol. 43, no. 3, pp. 227–268, 2009.
- [162] SAURÍ, R. and PUSTEJOVSKY, J., “Are you sure that this happened? assessing the factuality degree of events in text,” *Computational Linguistics*, vol. 38, no. 2, pp. 261–299, 2012.
- [163] SCHAMBER, L., “Users’ criteria for evaluation in a multimedia environment.,” in *Proceedings of the ASIS Annual Meeting*, vol. 28, pp. 126–33, ERIC, 1991.
- [164] SCHWARZ, N., BLESS, H., STRACK, F., KLUMPP, G., RITTENAUER-SCHATKA, H., and SIMONS, A., “Ease of retrieval as information: Another look at the availability heuristic.,” *Journal of Personality and Social psychology*, vol. 61, no. 2, p. 195, 1991.

- [165] SELF, C. C., “Credibility,” *An integrated approach to communication theory and research*, vol. 1, pp. 421–441, 1996.
- [166] SHAW, A. D., HORTON, J. J., and CHEN, D. L., “Designing incentives for inexperienced human raters,” in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 275–284, ACM, 2011.
- [167] SHENG, V. S., PROVOST, F., and IPEIROTIS, P. G., “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*, pp. 614–622, ACM, 2008.
- [168] SHESHADRI, A. and LEASE, M., “Square: A benchmark for research on computing crowd consensus,” in *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.
- [169] SHIBUTANI, T., *Improvised news: A sociological study of rumor*. Ardent Media, 1966.
- [170] SHIVAJAPPA, A. N., “Top 100 spam trigger words and phrases to avoid.” <http://www.leadformix.com/blog/2013/09/top-100-spam-trigger-words-and-phrases-to-avoid/>.
- [171] SHROUT, P. E. and FLEISS, J. L., “Intraclass correlations: uses in assessing rater reliability,” *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.
- [172] SINHA, S., CHATTERJEE, A., CHAKRABORTI, A., and CHAKRABARTI, B. K., *Econophysics: an introduction*. John Wiley & Sons, 2010.
- [173] SNOW, R., O’CONNOR, B., JURAFSKY, D., and NG, A. Y., “Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks,” in *Proceedings of the conference on empirical methods in natural language processing*, pp. 254–263, Association for Computational Linguistics, 2008.
- [174] SONI, S., MITRA, T., GILBERT, E., and EISENSTEIN, J., “Modeling factuality judgments in social media text,” in *ACL (2)*, pp. 415–420, 2014.
- [175] SONI, S., MITRA, T., GILBERT, E., and EISENSTEIN, J., “Modeling factuality judgments in social media text,” in *Proc. ACL*, 2014.
- [176] SOROKIN, A. and FORSYTH, D., “Utility data annotation with amazon mechanical turk,” in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW’08. IEEE Computer Society Conference on*, pp. 1–8, IEEE, 2008.
- [177] STEWART, J., *Calculus: Early transcendentals*. Cengage Learning, 2015.
- [178] SULLIVAN, D., “Twitter’s real time spam problem,” *Search Engine Land*, 2009.
- [179] SUN, Y.-A., ROY, S., and LITTLE, G., “Beyond independent agreement: A tournament selection approach for quality assurance of human computation tasks,” *Human Computation*, vol. 11, p. 11, 2011.

- [180] SUNDAR, S. S., “The main model: A heuristic approach to understanding technology effects on credibility,” *Digital media, youth, and credibility*, vol. 73100, 2008.
- [181] SUNDAR, S. S. and NASS, C., “Source orientation in human-computer interaction programmer, networker, or independent social actor,” *Communication research*, vol. 27, no. 6, pp. 683–703, 2000.
- [182] SUROWIECKI, J., *The wisdom of crowds*. Random House LLC, 2005.
- [183] SUROWIECKI, J., *The wisdom of crowds*. Anchor, 2005.
- [184] SUZUKI, R. and SHIMODAIRA, H., “Pvclust: an r package for assessing the uncertainty in hierarchical clustering,” *Bioinformatics*, vol. 22, no. 12, pp. 1540–1542, 2006.
- [185] TAUSCZIK, Y. R. and PENNEBAKER, J. W., “The psychological meaning of words: Liwc and computerized text analysis methods,” *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.
- [186] TRUMBACH, C. C. and PAYNE, D., “Identifying synonymous concepts in preparation for technology mining,” *Journal of Information Science*, vol. 33, no. 6, pp. 660–677, 2007.
- [187] TUFEKCI, Z., “Big questions for social media big data: Representativeness, validity and other methodological pitfalls,” in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
- [188] TVERSKY, A. and KAHNEMAN, D., “Availability: A heuristic for judging frequency and probability,” *Cognitive psychology*, vol. 5, no. 2, pp. 207–232, 1973.
- [189] TVERSKY, A. and KAHNEMAN, D., “Judgment under uncertainty: Heuristics and biases,” in *Utility, probability, and human decision making*, pp. 141–162, Springer, 1975.
- [190] UGANDER, J., KARRER, B., BACKSTROM, L., and MARLOW, C., “The anatomy of the facebook social graph,” *arXiv preprint arXiv:1111.4503*, 2011.
- [191] WANG, X., ZHAI, C., HU, X., and SPROAT, R., “Mining correlated bursty topic patterns from coordinated text streams,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 784–793, ACM, 2007.
- [192] WANG, Y.-C., KRAUT, R., and LEVINE, J. M., “To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 833–842, ACM, 2012.
- [193] WARD JR, J. H., “Hierarchical grouping to optimize an objective function,” *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.
- [194] WENG, J. and LEE, B.-S., “Event detection in twitter,” in *Proc. ICWSM*, 2011.

- [195] WIEBE, J., BRUCE, R., BELL, M., MARTIN, M., and WILSON, T., “A corpus study of evaluative and speculative language,” in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*, pp. 1–10, Association for Computational Linguistics, 2001.
- [196] WIEBE, J., WILSON, T., and CARDIE, C., “Annotating expressions of opinions and emotions in language,” *Language resources and evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.
- [197] WIEBE, J. M., “Tracking point of view in narrative,” *Computational Linguistics*, vol. 20, no. 2, pp. 233–287, 1994.
- [198] WILLETT, W., HEER, J., and AGRAWALA, M., “Strategies for crowdsourcing social data analysis,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 227–236, ACM, 2012.
- [199] WILSON, T., HOFFMANN, P., SOMASUNDARAN, S., KESSLER, J., WIEBE, J., CHOI, Y., CARDIE, C., RILOFF, E., and PATWARDHAN, S., “Opinionfinder: A system for subjectivity analysis,” in *Proceedings of hlt/emnlp on interactive demonstrations*, pp. 34–35, Association for Computational Linguistics, 2005.
- [200] WILSON, T., WIEBE, J., and HOFFMANN, P., “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 347–354, Association for Computational Linguistics, 2005.
- [201] WU, F. and HUBERMAN, B. A., “Novelty and collective attention,” *PNAS*, vol. 104, no. 45, pp. 17599–17601, 2007.
- [202] WU, Y., ZHOU, C., XIAO, J., KURTHS, J., and SCHELLNHUBER, H. J., “Evidence for a bimodal distribution in human communication,” *PNAS*, vol. 107, no. 44, pp. 18803–18808, 2010.
- [203] YANG, J. and LESKOVEC, J., “Patterns of temporal variation in online media,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 177–186, ACM, 2011.
- [204] YOGATAMA, D., WANG, C., ROUTLEDGE, B. R., SMITH, N. A., and XING, E. P., “Dynamic language models for streaming text,” *TACL*, vol. 2, pp. 181–192, 2014.
- [205] ZANZOTTO, F. M., PENNACCHIOTTI, M., and TSIOUTSIOLIKLIS, K., “Linguistic redundancy in twitter,” in *Proc. EMNLP*, 2011.
- [206] ZENG, L., STARBIRD, K., and SPIRO, E. S., “# unconfirmed: Classifying rumor stance in crisis-related social media messages,” in *Tenth International AAAI Conference on Web and Social Media*, 2016.

- [207] ZHAO, W. X., JIANG, J., WENG, J., HE, J., LIM, E.-P., YAN, H., and LI, X., “Comparing twitter and traditional media using topic models,” in *European Conference on Information Retrieval*, pp. 338–349, Springer, 2011.
- [208] ZHAO, Z. and MEI, Q., “Questions about questions: An empirical analysis of information needs on twitter,” in *Proceedings of the 22nd international conference on World Wide Web*, pp. 1545–1556, ACM, 2013.
- [209] ZHAO, Z., RESNICK, P., and MEI, Q., “Enquiring minds: Early detection of rumors in social media from enquiry posts,” in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1395–1405, ACM, 2015.
- [210] ZIPE, G. K., *Human behavior and the principle of least effort*. addison-wesley press, 1949.
- [211] ZUBIAGA, A. and JI, H., “Tweet, but verify: epistemic study of information verification on twitter,” *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–12, 2014.