A Parsimonious Language Model of Social Media Credibility Across Disparate Events

Tanushree Mitra
tmitra3@gatech.edu

Graham P. Wright
gwrong@gatech.edu

Eric Gilbert
gilbert@cc.gatech.edu

School of Interactive Computing & GVU Center
Georgia Institute of Technology
Atlanta, GA, USA

ABSTRACT

Social media has increasingly become central to the way billions of people experience news and events, often bypassing journalists—the traditional gatekeepers of breaking news. Naturally, this casts doubt on the credibility of information found on social media. Here we ask: Can the language captured in unfolding Twitter events provide information about the event's credibility? By examining the first large-scale, systematically-tracked credibility corpus of public Twitter messages (66M messages corresponding to 1,377 real-world events over a span of three months), and identifying 15 theoretically grounded linguistic dimensions, we present a parsimonious model that maps language cues to perceived levels of credibility. While not deployable as a standalone model for credibility assessment at present, our 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.

Author Keywords

Credibility; linguistic markers; social media; events; Twitter; rumor; misinformation; language; evidentiality

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Asynchronous interaction, Web-based interaction.

INTRODUCTION

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 above, 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 paper 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 [8, 38, 56, 84], 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 [74]. Our analysis overcomes this sampling bias by adopting a corpus called CREDBANK—a large dataset of 1,377 social media event streams, varying in event types and distributed over 66 million twitter messages along with their corresponding crowd-sourced credibility annotations [47]. The massive dataset was constructed by iteratively tracking billions of public Twitter posts over a period of three months, computationally summarizing those into event streams followed by capturing expert-level human judgment at the time the event unfolded. It contains, for example, both objections to red cards thrown soccer matches, as well as the emergence of Ebola in West Africa.

Merging the data from CREDBANK with linguistic scholarship, we built a statistical model to predict perceived credibility from language. Our model takes 15 theoretically driven

CSCW'17, February 25-March 01, 2017, Portland, OR, USA.

ACM 978-1-4503-4335-0/17/03...$15.00.
DOI: http://dx.doi.org/10.1145/2998181.2998351

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. 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, our results show that the language used by millions of people on Twitter has considerable information about an event's perceived credibility.

After identifying the linguistic categories that most powerfully signal credibility, we reflect on the relative importance of specific phrases within these categories. Looking closely at our predictive phrases, we find that expressions indicating ambiguity, such as *confusing* and *disbelief* and words indicating assessment of a situation (e.g., *scrutinize*, *ponder*) were associated with lower perceived credibility. On the other hand, words indicating positive emotion and signaling agreement were correlated with higher perceived credibility. Surprisingly, words indicating positive sentiment but mocking the impracticality of the situation (e.g., *ha*, *grins*, *joking*) were associated with lower credibility. Other intuitively effective signals of lower credibility were the appearance of hedge words in message responses (e.g., *certain level*, *suspects*), while the presence of affirmative booster words such as *undeniable* was associated with higher perceived credibility.

Our findings are based on a parsimonious model with theoretically-selected language variables as its key input. We did this so that our results would be likely to hold over time and across a changing social media landscape. Moreover, the linguistic correlates revealed by our study can be useful in a wide range of computer mediated communication (CMC) applications—detecting and reasoning about the certainty of information, automating factuality judgments, computing reliability standards for an ongoing social media story, or even predicting a forthcoming rumor (with substantial error bounds, at the moment).

The remainder of our paper is organized as follows. After surveying related work, we present our theoretically-grounded language measures for assessing event credibility, describe the statistical framework our study employed and discuss our model's performance. Lastly, we present our results and discuss their implications. Appendices containing implementation details follow the body of the paper.

RELATED WORK

Over the past few years, with social media's emergence as a prominent news source [24] and concomitant concerns as to the quality of information presented therein, researchers of socio-technical systems have become increasingly interested in studying social media credibility. Below we summarize three lines of work within this area and situate our study with respect to them.

Perspectives on Credibility Perceptions

The study of credibility is highly interdisciplinary and scholars from different fields bring diverse perspectives to the definition of credibility [18, 57]. Credibility has been defined as believability [20], trust [31], reliability [65], accuracy [19], objectivity [15] and several other related concepts [30, 66]. It has also been defined in terms of characteristics of persuasive sources, characteristics of the message structure and content, and perceptions of the media [45]. While some studies have focussed 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 [18]. 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 [9, 18]. Thus, despite decades of scholarly research on credibility, a single clear definition is yet to arise [30]. 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 [18, 20, 57]. 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 [57]). Following in their footsteps, when we constructed CREDBANK, we focusing on accuracy as a facet of information quality and instructed raters to rate the accuracy of social media events during the credibility assessment phase. Our current work based on CREDBANK also follows the information science approach and treats credibility perception as a characteristic of information quality.

One key component of credibility judgments that we 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 involve multiple layers of source [71]. 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 [71, 72]. As Sundar [71] 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 [49, 85]. 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 [68]. Motivated by these findings, our focus in the current study is on linguistic 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.

Social Media and Credibility

While individuals increasingly rely on online social networks to share diverse types of information quickly, without recourse to established official sources, modern online social networks like Facebook and Twitter are neutral with respect to information quality [22]. Moreover, quality compromises can occur as a consequence of spam content [25], stealth advertising [51, 56, 70], and propagation of rumor and misinformation [11, 17, 35, 44]. Thus, assessing the credibility of social media information has attracted the attention of many social media researchers. Scholars have studied specific events that were subjects of misinformation, such as the spread of rumors during the 2011 Great East Japan earthquake [41], the 2013 Boston marathon bombings [42], the 2014 Sydney siege event [1], and rumor dynamics in a Chinese microblogging community [40]. Taken together, these studies suggest information content analysis as vital towards understanding the role of misinformation in social media. However, their findings have been limited in scope since they are based on a few selected instances of misinformation analyzed after the event's occurrence. To arrive at a more holistic understanding of the interplay between language content and information credibility in an unfolding social media event, our study went beyond examining specific events and explored the importance of language in a large corpus of disparate events with in-situ credibility annotations.

A parallel trend of work within this domain is developing the capability to predict the credibility of information communicated through social media; for example building classifiers to detect factuality of information on Twitter [8], predicting credibility level of tweets [27, 55], automatically classifying rumor stances expressed in crisis events [82], or detecting controversial information from inquiry phrases [84]. These studies have found that linguistic features are one of the top predictors of whether information is credible or not. In particular, expressions of anxiety, uncertainty and sentiment have been noted as useful signals of credibility [8, 69, 84]. However, the intricacies involved in linguistic expressions affecting information credibility are still largely unknown. Therefore, drawing on the initial results of this line of study, we augmented our language model to include predictors representing expressions of anxiety, uncertainty and sentiment.

Event Factuality and Language

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 [62]. 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 [54]. The MPQA Opinion corpus includes annotations regarding the degree of factuality of expressions [78]. 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 [62]. 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. Following in the footsteps of this rich body of corpus based factuality analysis, our study 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 [47].

Another noteworthy work in this area is Rubin's theoretical framework for identifying certainty in texts [60]. 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 us 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.

METHOD

To search for language cues indicating credibility, we employed data from the CREDBANK corpus [47]. The CREDBANK corpus was built by iteratively tracking millions of public Twitter posts using Twitter's Streaming API followed by routing tweet streams to Amazon Mechanical Turkers to obtain credibility annotations. The annotations were collected on a 5-point Likert scale ranging from "Certainly Inaccurate [-2]" to "Certainly Accurate [+2]". To ensure that the collected annotations were of the same standard as expert level judgments, multiple controlled experiments were performed before finalizing the strategy best suited for obtaining high quality annotations. The corpus covers 1,377 events reported on Twitter between October 2014 and February 2015, their corresponding public tweets (a total of 66M messages) and their associated credibility ratings.

Our study's unit of analysis was an individual event and the perceived credibility level of its reportage on Twitter. Our measurement of perceived credibility level was based on the number of annotators that rated the event's reportage as "Certainly Accurate". More formally, for each event, we found the proportion of annotations (P_{ca}) rating the reportage as

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

Table 1: 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.

"Certainly Accurate".

$$P_{ca} = \frac{\textit{``Certainly Accurate'' ratings for an event}}{\textit{Total ratings for that event}}$$

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, we placed P_{ca} into four classes that cover a range of values. We 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 1 shows the credibility classes and the number of events in each class. Table 2 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, we compared classes generated by the P_{ca} method to those obtained via a data-driven classification technique (refer Appendix for details). We found a high degree of agreement between the P_{ca}-based and data-driven classification approaches. We 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.

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"). We 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 [63] and others advocating for probabilistic modeling of differing annotator judgments so as to capture their inherent uncertainty [14]. By selecting a proportion-based (P_{ca}) ordinal scale, we 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.

Predictive Variables: Linguistic Measures

To detect linguistic strategies corresponding to credibility assessment, we compiled several language-based measures after reviewing the principles underlying factuality judgments and veracity assessments [14, 36, 37, 64]. Building on lexical and computational insights, we identified 15 linguistic measures as potential predictors of perceived credibility level. Using standard methods from computational linguistics, we incorporated these measures as features in our statistical model (discussed shortly). Specifically, we used specialized lexicons designed to operationalize language-based measures. Below we justify our choice of each measure as a potential credibility marker.

Modality: Modality is an expression of an individual's "subjective attitude" [7] and "psychological stance" [46] 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 [14, 60]. 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, we measured the modality expressed in an event's reportage by using Sauri et al's list of modal words [62], which have been successfully used in prior research on veridicality assessment [13, 64, 68]. By counting the occurrences of each modal word in an event reportage, we incorporated them as input features of our statistical model. We followed the same technique for other lexicon-based measures.

Subjectivity: Subjectivity is used to express opinions and evaluations [2, 79]. Hence, detecting the presence of subjectivity can differentiate opinions from factual information (often called objective reporting) [2, 77, 79]. Prior research has shown that knowledge of subjective language can be useful in analyzing objectivity in news reporting [77] and in recognizing certainty in textual information [60]. Drawing on these prior works, we hypothesized that subjectivity can provide meaningful signals for credibility assessment and used OpinionFinder's subjectivity lexicon comprising 8,222 words [80].

Hedges: Hedges refer to terms "whose job is to make things more or less fuzzy" [39]. They are often used to express lack of commitment to the truth value of a claim or to display skepticism and caution [32]. People who are uncertain about a topic tend to use such tentative language [73]. Work on certainty categorization in newspaper articles found that hedges were used to classify statements into low or moderate levels of certainty [60], thus demonstrating the intrinsic connection between hedges and expressions of certainty – a concept closely related to credibility assessment. Hence, we included hedges as potential credibility markers of an event's reportage. To measure hedges, we used two sets of lexicons signaling tentative language: 1) list of hedge words from Hyland [34] and 2) tentative words from the LIWC dictionary [73].

Evidentiality: Evidentials are recognized as a means of expressing the degree of reliability of reported information [3,

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 #sydneysiege	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 #oscars 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
#charliehebdo #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 2: 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. [47] 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.

60]. 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 [60, 62]. Thus the choice of these attributive predicates can express the level of commitment in the reported information [3], or indicate a speaker's evidential stance or even express the level of factuality in events [62]. 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 [60]. Recent studies have shown that evidentiality predicates can affect credibility perceptions of quoted content in journalistic tweets [68]. 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* [29]. 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 [13]. Hence we include negation as a potential credibility marker. We measure it by using a lexicon of negation particles from the De Facto lexicon—a factuality profiler for event mentions in texts [62].

Exclusions and Conjunctions: Both exclusions and conjunctions are components for reasoning [73]. While exclusion words like *but, either, except* are useful in determining if something belongs to a category [73], 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 [29, 52]. 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, we 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 [5, 59]. Since apprehensive statements typically manifest anxiety in the context of information transmission [4], we 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 [4, 5]. To measure the extent of emotions expressed in event specific tweets, we used LIWC's comprehensive list of positive and negative emotion words [73].

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 [33]. Hypothesizing that booster words can be useful credibility markers, we used the list of "booster" words compiled by Hyland [34] and the list of "certainty" words listed in the LIWC dictionary [73] 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, we 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 [14]. 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 [68]. Based on these studies, we 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, we 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 [50]. 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 [33, 34]. 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 [5, 4, 67]. These studies suggest the importance of asking questions in the face of uncertainty. Hence, we propose inclusion of questions as a potential indicator of perceived credibility level. We 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 [8]. Hence we include the count of hashtag terms in tweets associated with an event as a potential credibility marker.

Predictive Variables: Controls

We 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.

Model Limitations

Like any other statistical model building technique, a limitation with our approach is that we 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.

Statistical Technique

Our 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, we 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 3 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 we 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 we 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 [23], family relationships from Facebook messages [6] and in mapping sociocultural identity in tweets [16].

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 $\alpha \in [0, 0.1, 0.5, 1]$, we selected a parsimonious model with $\alpha = 1$. We 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, we included only control variables so as to measure their explanatory power. Next, we 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 4 outlines our iterative model building process. We first added features corresponding to all the original tweets in our dataset. For example, for a feature phrase such as *wow* from our positive emotion lexicon, we 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. We 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, we adopted the simplest representation – raw counts of linguistic features – as our model's independent variables.

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 3: 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 we 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.

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 we included all linguistic measures except *subjectivity*. We 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 enquiring tweets carry useful signals in assessing the certainty of information [84]. 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. We did not repeat the process for retweets because retweets essentially re-iterate what the original poster said. Instead, we simply add retweet count, number of words and average length of retweets to act as control variables to our model.

RESULTS

Model Fit Comparison

We 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.

[1] `https://cran.r-project.org/web/packages/glmnetcr`

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. We also observed a significant reduction in deviance when we 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, we term this omnibus model as our language classifier and report its accuracy in the next section.

The bottom half of Table 4 also lists model fits per linguistic class measure, i.e., how the model performed when we 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. We 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 1 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 βs were indicative of an event with lower perceived credibility. The thickness of the arcs in Figure 1 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. We observe that color saturation inverts for original and reply tweets along a few linguistic categories, such as booster, hedges, anxiety and the emotion categories. We interpret these results in our Discussion section.

How did our control variables perform? We found that the only control variables with non-zero positive β weights were: *average retweet length* (β = 0.25), *average reply length* (β = 0.18). Controls with non-zero negative βs were: *number of retweets* (β = –0.27), *average original tweet length* (β = –0.14), *number of words in retweets* (β = –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*. We found that, with the exception of proportion of quotations in original tweets (β = –0.097), the non-lexicon based features lacked predictive power (β = 0).

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 4: 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 grey).

Model Accuracy

We computed the performance of our language classifier according to four accuracy measurement schemes; the appendix 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 5 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.

	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 5: 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. We 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 6a 2). Level-1 Weight$_{0.25}$ gives partial credit of 0.25 if the classification is incorrect by one level only (Table 6b), 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 6c. Our language classifier significantly outperforms both the baselines (McNemar's test, $p < 10^{-16}$).

	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 6: 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).

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 6(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 6(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 6(c)).

We 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. We 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. We 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 classifier accuracy was significantly higher compared to both the baseline classifiers ($p < 10^{-16}$). Table 5 shows the precision, recall and F1 measures for each credibility class under different accuracy measurement schemes as well as the overall accuracy under each scheme.

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 [61].

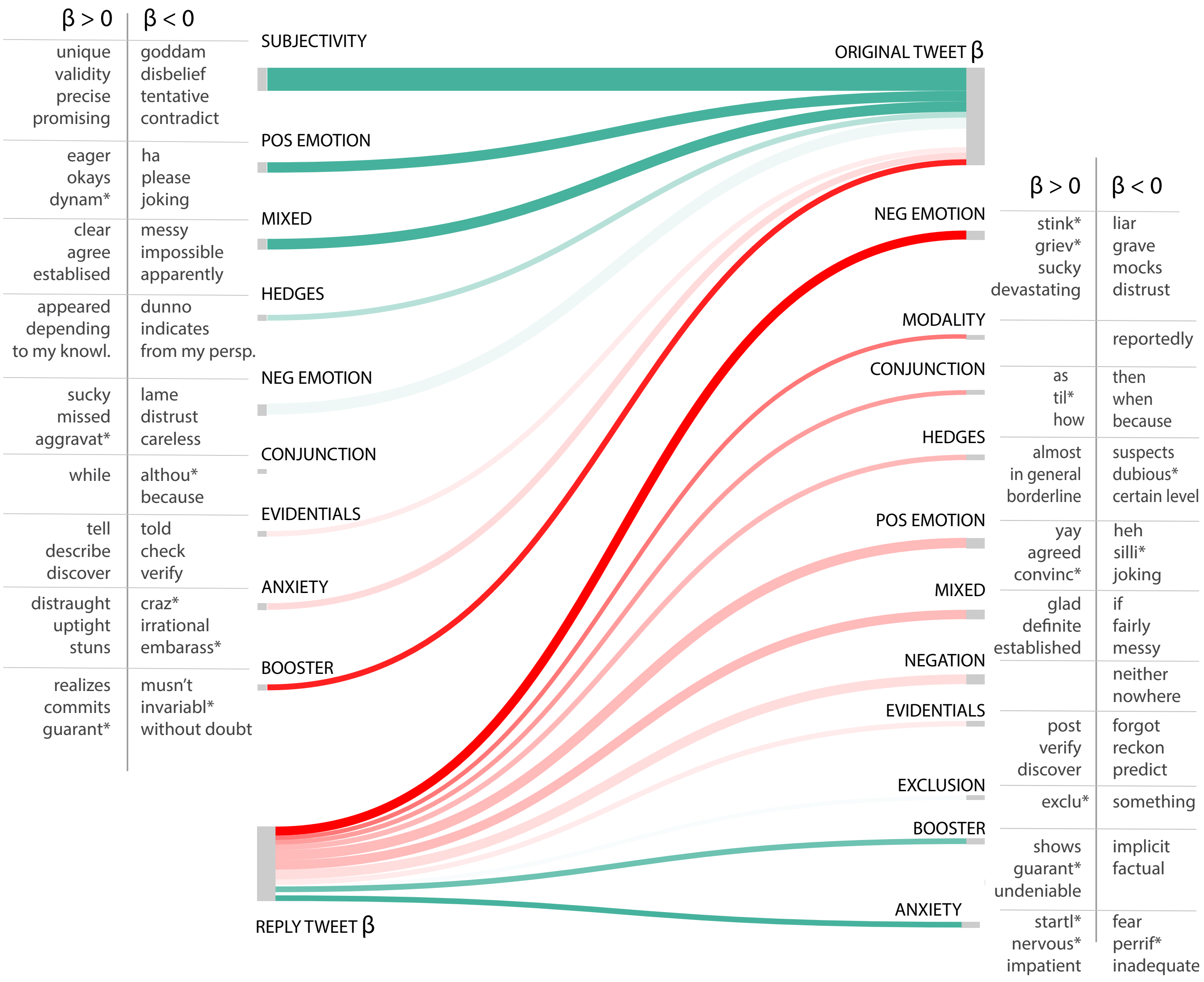


Figure 1: The predictive power of the linguistic measures from the omnibus model. A measure's weight is proportional to the deviance of the corresponding linguistic category (Table 4 lists the deviance numbers). The color saturation corresponds to the difference in the absolute values of positive and 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 βs is higher than the sum of positive βs, while the converse is true for higher concentration of green. The diagram also lists the top predictive phrases in each linguistic measure.

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 [68].

Taking a closer look at our most predictive words in each lexicon, we found a striking view of words and phrases signaling perceived credibility level of social media event reportages (see Figure 1 for an overview). Table 7 and 8 list the top predictive words in each linguistic category. Below we 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 awe 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

Subjecitvity	$\beta > 0$	Subjecitvity	$\beta > 0$	Subjecitvity	$\beta < 0$	Subjecitvity	$\beta < 0$
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 7: 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.

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 [75].

We 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 [59].

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 [26]. Thus event reportages with lower informational content would be perceived as less credible. Other correlates of negative βs 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 [58, 84].

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, we 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 1) 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, we 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*$^{*}_{[\beta=0.2]}$, *troubl*$^{*}_{[\beta=0.20]}$, *misses*$_{[\beta=0.17]}$, *heartbroke*$_{[\beta=0.12]}$, *sobbed*$_{[\beta=0.04]}$, *weep*$^{*}_{[\beta=0.02]}$, *fail*$^{*}_{[\beta=0.75]}$ *0.02, defeat*$_{[\beta=0.02]}$) . We 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*$_{[\beta=0.20]}$, *brillian**$_{[\beta=0.19]}$, *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 [60, 62].

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

Hedges paint a different picture. Hedge words in original tweets conveying information uncertainty (*appeared*$_{[\beta=0.26]}$, *halfass*$^{*}_{[\beta=0.13]}$, *to my knowledge*$_{[\beta=0.12]}$, *tends to*$_{[\beta=0.02]}$) or qualifying claims with conditions (*depending*$_{[\beta=0.23]}$ *contingen*$^{*}_{[\beta=0.14]}$) 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*$_{[\beta=-0.16]}$, *dubious*$^{*}_{[\beta=-0.12]}$, *suspects*$_{[\beta=-0.08]}$ 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*$_{[\beta=0.28]}$, *borderline*$_{[\beta=0.27]}$, *in general*$_{[\beta=0.09]}$, *fuzz*$^{*}_{[\beta=0.02]}$).

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*$_{[\beta=0.14]}$, *express*$_{[\beta=0.06]}$, *describe*$_{[\beta=0.05]}$ in originals, *declare*$_{[\beta=0.22]}$, *post*$_{[\beta=0.02]}$, *according*$_{[\beta=0.05]}$ in replies), fact checking (*verify*$_{[\beta=0.05]}$, *assert*$_{[\beta=0.18]}$ in replies) and knowledge acquisition (*discover* in both replies and original tweets). In contrast, evidentials correlating with lower levels of credibility indicated loosing knowledge (*forget*$_{[\beta=-0.50]}$ in replies), expressing uncertainty (*reckon*$_{[\beta=-0.03]}$, *predict*$_{[\beta=-0.01]}$ in replies) and fact checking in originals (*check*$_{[\beta=-0.09]}$, *verify*$_{[\beta=-0.02]}$). *told*$_{[\beta=-0.13]}$, 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, we 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 (*craz*$^{*}_{[\beta=0.11]}$, *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

[1] A 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.*

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			brillian*	0.19		
wonderf*	0.03			fantastic*	0.18		
				wonderf*	0.06		
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**	$\beta > 0$	**Exclusion**	$\beta < 0$
				exclu*	0.12	something*	-0.09
Booster	$\beta > 0$	**Booster**	$\beta < 0$	**Booster**	$\beta > 0$	**Booster**	$\beta < 0$
commits	0.16	without doubt	-0.25	undeniable	0.36	implicit*	-0.15
guarant*	0.07	invariab*	-0.13	shows	0.23	total	-0.02
realizes	0.02	mustn't	-0.06	guarant*	0.05	factual*	-0.02
Anxiety	$\beta > 0$	**Anxiety**	$\beta < 0$	**Anxiety**	$\beta > 0$	**Anxiety**	$\beta < 0$
distraught	0.19	vulnerab*	-0.34	distress*	0.24	fear	-0.15
uptight	0.05	craz*	-0.19	miser*	0.15	petrif*	-0.12
scaring	0.01	uncontrol*	-0.08	startl*	0.08	inadequa.	-0.10
stuns	0.05	turmoil	-0.04	nervous*	0.04	desperat*	-0.02
		irrational*	-0.03	impatien*	0.03	shaki*	-0.01
		embarrass*	-0.02			**Modality**	$\beta < 0$
						reportedly	-0.53
Evidentials	$\beta > 0$	**Evidentials**	$\beta < 0$	**Evidentials**	$\beta > 0$	**Evidentials**	$\beta < 0$
tell	0.14	told	-0.13	verify	0.05	predict	-0.01
describe	0.05	check	-0.09	post	0.02	reckon	-0.03
discover	0.00	verify	-0.02	discover	0.01	forgot	-0.49
Conjunction	$\beta > 0$	**Conjunction**	$\beta < 0$	**Conjunction**	$\beta > 0$	**Conjunction**	$\beta < 0$
while	0.47	because	-0.09	as	0.14	then	-0.37
		altho	0.00	how	0.06	when	-0.18
				til	0.04	because	-0.07
Mixed	$\beta > 0$	**Mixed**	$\beta < 0$	**Mixed**	$\beta > 0$	**Mixed**	$\beta < 0$
established	0.23	impossible	-0.17	glad	0.59	fairly	-0.19
clear	0.18	messy	-0.14	definite	0.06	messy	-0.12
agree	0.08	apparently	-0.11	established	0.04	if	-0.10
						fun	-0.08
						hilarious	-0.06

Table 8: The top predictive phrases per linguistic category associated with higher ($\beta > 0$) and lower ($\beta < 0$) levels of perceived credibility are listed. Phrases corresponding to the original tweets are on the left while those corresponding to the replies are on the right. All phrases are significant at the 0.001 level.

threatening situations rumors are aimed at relieving tensions of anxiety [5].

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, we 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.

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 [83, 84]. 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 we know a great deal about the relationship between language and sentiments or language and opinion, we know very little about how people perceive credibility of events in textual conversations. By studying social media credibility through a linguistically well-grounded model, we 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.

Design Implications

We 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 we 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. We 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 [28, 38], popularity of the post (measured by the number retweets or replies) [28] or the network structure of content diffusion [8, 38, 56]. While useful, these features can only be collected after the content (whether accurate or not) have disseminated for a while [84]. 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.

CONCLUSION

In this work we uncover 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, we unfold ways in which social media text carry signals of information credibility. We hope our work motivates 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.

ACKNOWLEDGMENTS

This research was supported by DARPA-W911NF-12-1-0043 and by Army Research Lab-W911NF-15-1-0280. We would like to thank our comp.social lab at GeorgiaTech for their valuable comments on early drafts of the paper. The final paper was greatly improved by the advice from the anonymous reviewers, thank you.

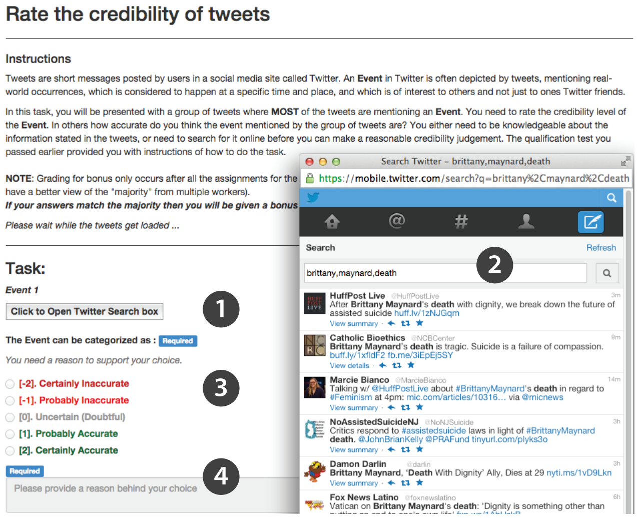
APPENDIX: CORPUS DETAILS

This section provides a brief summary of the CREDBANK corpus on which our paper is based. The CREDBANK corpus was constructed by combining machine computation with crowd-sourced judgments of human annotators. It's construction followed a sequence of phases:

1. Streaming Tweets and Preprocessing: Twitter's Streaming API was used to iteratively collect a continuous 1% sample of all global tweets. Every group of million streaming tweets was filtered to contain only English tweets, followed by spam removal, tokenization using a Twitter specific tokenizer [53] and a sophisticated multi-stage stop word removal step.

2. Detecting Event Candidates: The next phase involved automatic detection of events from social media streams. After carefully considering various approaches for event detection, ranging from key-word based methods, bursty term analysis techniques to topic modeling based methods, Mitra et al. [47] opted for topic models since topic models can learn term co-occurrences and unlike keyword based techniques do not make a priori assumption about what constitutes an event.

3. Event annotation: To eliminate the detection of potential false positives using a purely computational event detection approach, candidate events from the previous step were sent through a human annotation pipeline. Ten independent human raters from Amazon Mechanical Turk (AMT) judged whether a topic relates to a real-world news event or not. The majority agreement was selected as the final annotation, thus separating the event-specific topics from the non-event topics in a reliable manner.



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

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

Required

Please provide a reason behind your choice

Search Twitter – brittany,maynard,death

https://mobile.twitter.com/search?q=brittany%2Cmaynard%2Cdeath

Search Refresh

brittany,maynard,death

HuffPost Live @HuffPostLive 3m
After **Brittany Maynard**'s **death** with dignity, we break down the future of assisted suicide huff.lv/1zNJGqm
View summary

Catholic Bioethics @NCBCenter 9m
Brittany Maynard's **death** is tragic. Suicide is a failure of compassion. buff.ly/1xfldF2 fb.me/3iEpEj5SY
View details

Marcie Bianco @MarcieBianco 14m
Talking w/ @HuffPostLive about #BrittanyMaynard's **death** in regard to #Feminism at 4pm: mic.com/articles/10316… via @micnews
View summary

NoAssistedSuicideNJ @NoNJSuicide 3h
Critics respond to #assistedsuicide laws in light of #BrittanyMaynard **death**. @JohnBrianKelly @PRAFund tinyurl.com/plyks3o
View summary

Damon Darlin @darlin 3h
Brittany Maynard, '**Death** With Dignity' Ally, Dies at 29 nyti.ms/1vD9Lkn
View summary

Fox News Latino @foxnewslatino 6h
Vatican on **Brittany Maynard**'s **death**: 'Dignity is something other than

Figure 2: 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. Figure has been reproduced from Mitra et al. [47].

3. Credibility Assessment: This phase involved three primary steps as outlined below:

Determining the credibility scale: Informed by work done by the linguistic community on 'Event Factuality', the credibility scale was designed as an interaction between two dimensions: Polarity, which differentiates among 'Accurate', 'Inaccurate', and 'Uncertain', and Degree of certainty which distinguishes among 'Certainly','Probably' and 'Uncertain', leading to a 5-point Likert scale annotation scheme.

Determining number of independent Turk ratings for high quality annotation: To determine the number of Turker responses required to closely approximate an event's credibility perception to that of an expert's credibility assessment, the CREDBANK system was piloted over a span of 5 days collecting and annotating 50 events by both Turkers and expert annotators (university research librarians). The pilot study was followed by computing correlation statistics between Turker mean responses and expert mean responses while varying the count of independent Turker ratings per event. The correlation maximized at 30 ratings leading to the decision of collecting 30 annotations per event.

Credibility assessment task: The credibility assessment task framework was designed to ensure that the collected credibility ratings is of high quality. Multiple controlled experiments were performed before finalizing the strategy best suited for obtaining quality annotations [48]. Turkers were first selectively screened and trained via a qualification test. Screened workers were then directed to a task interface as shown in Figure 2. Turkers were asked to categorize an event's credibility after reading through a stream of real-time tweets related to an event topic. They were instructed to either be knowledgeable on the event topic or search online before making their credibility judgments. The task design thus closely resembles how individuals would search Twitter to find information related to an event.

4. Collecting Event Streams: The final phase used Twitter's search API to collect all tweets specific to the event topic.

Overall, this iterative framework resulted in a natural experimental setup where the credibility of social media information was being tracked soon after it gained collective attention.

APPENDIX: ACCURACY MEASUREMENT

This section describes the mathematical implementation of our accuracy metrics. The most common way to access accuracy of a multi-class classification task is based on building 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.

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

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 6(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. [10], we employ an alternative measure defined directly in the confusion matrix. Table 6(b) and (c) displays our additional weighted confusion matrices. The off-diagonals of these matrices can be in the range of (0, $\cdots$, 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.

APPENDIX:VALIDATING CREDIBILITY CLASSIFICATION

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

Generating data-driven credibility classes

We used hierarchical agglomerative clustering (HAC) [43] 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 [12, 81]. We used the Euclidean distance similarity metric and Ward's fusion strategy for merging [76]. The choice of this strategy minimizes the within-cluster variance thus maximizing within-group similarity [76].

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, we need a metric to compare two clusterings of the same dataset. In other words, we need to measure how often both clustering methods classify the same set of observations as members of the same cluster. We borrow a technique proposed by Tibshirani et al. [21]. Let $P_{clust} = \{x_{1_{c_1}}, x_{2_{c_1}}, x_{3_{c_2}}, \cdots, x_{n_{c_4}}\}$ denote the cluster labels from P_{ca} based classification and $H_{clust} = \{x_{1_{h_1}}, x_{2_{h_3}}, x_{3_{h_4}}, \cdots, x_{n_{h_3}}\}$ the labels from HAC-based classification of the same dataset D of n observations. Here, $x_{i_{c_j}}$ denotes that the i^{th} observation belongs to cluster c_j as per the P_{ca} classification and $x_{i_{h_j}}$ denotes that the i^{th} observation belongs to cluster h_j as per the HAC classification. We see that $x_{1_{c_1}}$ and $x_{2_{c_1}}$ belong to the same cluster. Such pairs are called "co-members". While ($x_{1_{c_1}}$, $x_{2_{c_1}}$) are co-members as per P_{ca} classification, ($x_{2_{h_3}}, x_{n_{h_3}}$) are co-members from HAC classification. For each clustering method, we first compute all pairwise co-membership of all pairs of observations belonging to the same cluster. Next we 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. We obtain a fairly high R of 0.774 denoting high agreement between our P_{ca} based and HAC-based clustering approaches. We 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.

REFERENCES

1. Ahmer Arif, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, and Emma S Spiro. 2016. 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*. ACM, 466–477.

2. Ann Banfield. 1982. Unspeakable sentences. (1982).

3. Sabine Bergler, Monia Doandes, Christine Gerard, and René Witte. 2004. Attributions. *Exploring Attitude and Affect in Text: Theories and Applications, Technical Report SS-04-07* (2004), 16–19.

4. Prashant Bordia and Nicholas DiFonzo. 2004. Problem solving in social interactions on the Internet: Rumor as social cognition. *Social Psychology Quarterly* 67, 1 (2004), 33–49.

5. Prashant Bordia and Ralph L Rosnow. 1998. Rumor Rest Stops on the Information Highway Transmission Patterns in a Computer-Mediated Rumor Chain. *Human Communication Research* 25, 2 (1998), 163–179.

6. Moira Burke, Lada Adamic, and Karyn Marciniak. 2013. Families on Facebook. In *Seventh International AAAI Conference on Weblogs and Social Media*.

7. Joan Bybee, Revere Perkins, and William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. University of Chicago Press.

8. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.

9. Steven H Chaffee. 1982. Mass media and interpersonal channels: Competitive, convergent, or complementary. *Inter/media: Interpersonal communication in a media world* (1982), 57–77.

10. Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4 (1968), 213.

11. M Corcoran. 2009. Death by cliff plunge, with a push from twitter. *The New York Times* (2009).

12. Riley Crane and Didier Sornette. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* 105, 41 (2008), 15649–15653.

13. M de Marneffe, Christopher D Manning, and Christopher Potts. 2011. Veridicality and utterance understanding. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*. IEEE, 430–437.

14. Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational linguistics* 38, 2 (2012), 301–333.

15. Jaap J Dijkstra, Wim BG Liebrand, and Ellen Timminga. 1998. Persuasiveness of expert systems. *Behaviour & Information Technology* 17, 3 (1998), 155–163.

16. Jacob Eisenstein, Noah A Smith, and Eric P Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. ACL, 1365–1374.

17. Golnaz Esfandiari. 2010. The Twitter Devolution. *Foreign Policy* 7 (2010), 2010.

18. Andrew J Flanagin and Miriam J Metzger. 2008. Digital media and youth: Unparalleled opportunity and unprecedented responsibility. *Digital media, youth, and credibility* (2008), 5–27.

19. BJ Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. 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*. ACM, 1–15.

20. BJ Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 80–87.

21. Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33, 1 (2010), 1.

22. Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor Cascades. In *Eighth International AAAI Conference on Weblogs and Social Media*.

23. Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1037–1046.

24. Jeffrey Gottfried and Elisa shearer. 2016. News Use Across Social Media Platforms 2016. *Pew Research* (2016). Retrieved Oct 16, 2016 from `www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016`.

25. R Grover. 2011. Ad. ly: The Art of Advertising on Twitter. *Businessweek, January* 6 (2011).

26. Aditi Gupta and Ponnurangam Kumaraguru. 2012. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*. ACM, 2.

27. Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.

28. Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. 2013. $1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter. In *eCrime Researchers Summit, 2013*. IEEE, 1–12.

29. Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes* 45, 1 (2007), 1–23.

30. Brian Hilligoss and Soo Young Rieh. 2008. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management* 44, 4 (2008), 1467–1484.

31. Carl I Hovland, Irving L Janis, and Harold H Kelley. 1953. Communication and persuasion; psychological studies of opinion change. (1953).

32. Ken Hyland. 1998. *Hedging in scientific research articles*. Vol. 54. John Benjamins Publishing.

33. Ken Hyland. 2002. Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics* 34, 8 (2002), 1091–1112.

34. Ken Hyland. 2005. *Metadiscourse: Exploring interaction in writing*. Wiley Online Library.

35. C Kanalley. 2011. Facebook Shutting Down Rumor Goes Viral: Site Said to be Ending March 15, 2011. *The Huffington Post* (2011).

36. Lauri Karttunen and Annie Zaenen. 2005. Veridicity. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

37. Paul-Kiparsky Kiparsky. 1971. Carol (1970),"Fact". *Bierwisch, Manfred-Heidolph, Karl Erich (a cura di), Progress in Linguistics (A Collection of Papers), The Hague, Mouton* (1971), 143–173.

38. Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 1103–1108.

39. George Lakoff. 1975. Hedges: a study in meaning criteria and the logic of fuzzy concepts. In *contemporary Research in Philosophical Logic and Linguistic semantics*. Springer, 221–271.

40. Qinying Liao and Lei Shi. 2013. She gets a sports car from our donation: rumor transmission in a chinese microblogging community. In *Proc. CSCW*.

41. Fang Liu, Andrew Burton-Jones, and Dongming Xu. 2014. Rumor on Social Media in Disasters: Extending Transmission to Retransmission. In *Proceedings of the Pacific Asia Conference on Information Systems*.

42. Jim Maddock, Kate Starbird, Haneen Al-Hassani, Daniel E Sandoval, Mania Orand, and Robert M Mason. 2015. Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures. (2015).

43. Oded Maimon and Lior Rokach. 2005. *Data mining and knowledge discovery handbook*. Vol. 2. Springer.

44. Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter Under Crisis: Can we trust what we RT?. In *Proceedings of the first workshop on social media analytics*. ACM, 71–79.

45. Miriam J Metzger, Andrew J Flanagin, Keren Eyal, Daisy R Lemus, and Robert M McCann. 2003. Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the International Communication Association* 27, 1 (2003), 293–335.

46. Terence F Mitchell and Shāhir ?asan. 1994. *Modality, Mood, and Aspect in Spoken Arabic: With Special Reference to Egypt and the Levant*. Vol. 11. Routledge.

47. Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations. In *Ninth International AAAI Conference on Web and Social Media*.

48. Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing Person-and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1345–1354.

49. Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 441–450.

50. Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. 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*. ACM, 1739–1748.

51. Eni Mustafaraj and P Takis Metaxas. 2010. From obscurity to prominence in minutes: Political speech and real-time search. *Web Science* (2010).

52. Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29, 5 (2003), 665–675.

53. Brendan O'Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter.. In *Fourth International AAAI Conference on Weblogs and Social Media*.

54. James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and others. 2003. The timebank corpus. In *Corpus linguistics*, Vol. 2003.

55. Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1589–1599.

56. Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 249–252.

57. Soo Young Rieh and David R Danielson. 2007. Credibility: A multidisciplinary framework. *Annual review of information science and technology* 41, 1 (2007), 307–364.

58. Ralph L Rosnow. 1988. Rumor as communication: A contextualist approach. *Journal of Communication* 38, 1 (1988), 12–28.

59. Ralph L Rosnow. 1991. Inside rumor: A personal journey. *American Psychologist* 46, 5 (1991), 484.

60. Victoria L Rubin, Elizabeth D Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications*. Springer, 61–76.

61. Nadine B Sarter and David D Woods. 1991. Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology* 1, 1 (1991), 45–57.

62. Roser Sauri. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. Dissertation. Waltham, MA, USA. Advisor(s) Pustejovsky, James.

63. Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language resources and evaluation* 43, 3 (2009), 227–268.

64. Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics* 38, 2 (2012), 261–299.

65. Linda Schamber. 1991. Users' Criteria for Evaluation in a Multimedia Environment.. In *Proceedings of the ASIS Annual Meeting*, Vol. 28. ERIC, 126–33.

66. Charles C Self. 1996. Credibility. *An integrated approach to communication theory and research* 1 (1996), 421–441.

67. Tamotsu Shibutani. 1966. *Improvised news: A sociological study of rumor*. Ardent Media.

68. Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. In *Proc. ACL (short papers)*.

69. Kate Starbird, Emma Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindhuja Narasimhan. 2016. Could This Be True?: I Think So! Expressed Uncertainty in Online Rumoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 360–371.

70. D Sullivan. 2009. Twitter's Real Time Spam Problem. *Search Engine Land* (2009).

71. S Shyam Sundar. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility* 73100 (2008).

72. S Shyam Sundar and Clifford Nass. 2000. Source orientation in human-computer interaction programmer, networker, or independent social actor. *Communication research* 27, 6 (2000), 683–703.

73. Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

74. Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*.

75. Amos Tversky and Daniel Kahneman. 1975. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*. Springer, 141–162.

76. Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.

77. Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. 2001. A corpus study of evaluative and speculative language. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*. ACL, 1–10.

78. Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39, 2-3 (2005), 165–210.

79. Janyce M Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics* 20, 2 (1994), 233–287.

80. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. ACL, 347–354.

81. Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 177–186.

82. Li Zeng, Kate Starbird, and Emma S Spiro. 2016. # Unconfirmed: Classifying Rumor Stance in Crisis-Related Social Media Messages. In *Tenth International AAAI Conference on Web and Social Media*.

83. Zhe Zhao and Qiaozhu Mei. 2013. Questions about questions: An empirical analysis of information needs on Twitter. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1545–1556.

84. Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 1395–1405.

85. Arkaitz Zubiaga and Heng Ji. 2014. Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining* 4, 1 (2014), 1–12.