Modeling Factuality Judgments in Social Media Text

Sandeep Soni **Tanushree Mitra** **Eric Gilbert** **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
soni.sandeepb@gmail.com, {tmitra3,gilbert,jeisenst}@cc.gatech.edu

Abstract

How do journalists mark quoted content as certain or uncertain, and how do readers interpret these signals? Predicates such as *thinks*, *claims*, and *admits* offer a range of options for framing quoted content according to the author's own perceptions of its credibility. We gather a new dataset of direct and indirect quotes from Twitter, and obtain annotations of the perceived certainty of the quoted statements. We then compare the ability of linguistic and extra-linguistic features to predict readers' assessment of the certainty of quoted content. We see that readers are indeed influenced by such framing devices — and we find no evidence that they consider other factors, such as the source, journalist, or the content itself. In addition, we examine the impact of specific framing devices on perceptions of credibility.

1 Introduction

Contemporary journalism is increasingly conducted through social media services like Twitter (Lotan et al., 2011; Hermida et al., 2012). As events unfold, journalists and political commentators use quotes — often indirect — to convey potentially uncertain information and claims from their sources and informants, e.g.,

Figure 1: Indirect quotations in Twitter

A key pragmatic goal of such messages is to convey the provenance and uncertainty of the quoted content. In some cases, the author may also introduce their own perspective (Lin et al., 2006) through the use of framing (Greene and Resnik, 2009). For instance, consider the use of the word *claims* in Figure 1, which conveys the author's doubt about the indirectly quoted content.

Detecting and reasoning about the certainty of propositional content has been identified as a key task for information extraction, and is now supported by the FactBank corpus of annotations for newstext (Saurí and Pustejovsky, 2009). However, less is known about this phenomenon in social media — a domain whose endemic uncertainty makes proper treatment of factuality even more crucial (Morris et al., 2012). Successful automation of factuality judgments could help to detect online rumors (Qazvinian et al., 2011), and might enable new applications, such as the computation of reliability ratings for ongoing stories.

This paper investigates how linguistic resources and extra-linguistic factors affect perceptions of the certainty of quoted information in Twitter. We present a new dataset of Twitter messages that use FactBank predicates (e.g., *claim*, *say*, *insist*) to scope the claims of named entity sources. This dataset was annotated by Mechanical Turk workers who gave ratings for the factuality of the scoped claims in each Twitter message. This enables us to build a predictive model of the factuality annotations, with the goal of determining the full set of relevant factors, including the predicate, the source, the journalist, and the content of the claim itself. However, we find that these extra-linguistic factors do not predict readers' factuality judgments, suggesting that the journalist's own framing plays a decisive role in the credibility of the information being conveyed. We explore the specific linguistic feature that affect factuality judgments, and compare our findings with previously-proposed groupings of factuality-related predicates.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 415–420,
Baltimore, Maryland, USA, June 23-25 2014.

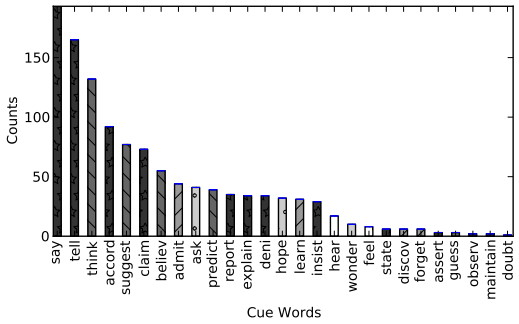

Figure 2: Count of cue words in our dataset. Each word is patterned according to its group, as shown in Figure 3.

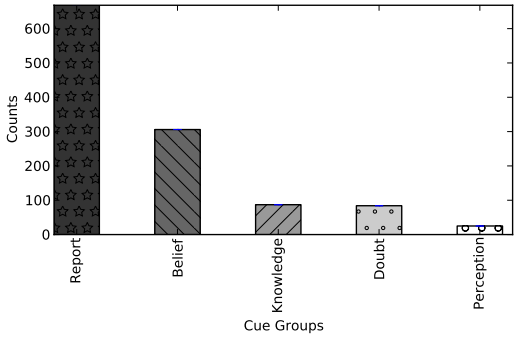

Figure 3: Count of cue groups in our dataset

2 Text data

We gathered a dataset of Twitter messages from 103 professional journalists and bloggers who work in the field of American Politics.[1] Tweets were gathered using Twitter's streaming API, extracting the complete permissible timeline up to February 23, 2014. A total of 959,754 tweets were gathered, and most were written in early 2014.

Our interest in this text is specifically in quoted content — including "indirect" quotes, which may include paraphrased quotations, as in the examples in Figure 1. While labeled datasets for such quotes have been created (O'Keefe et al., 2012; Pareti, 2012), these are not freely available at present. In any case, the relevance of these datasets to Twitter text is currently unproven. Therefore, rather than train a supervised model to detect quotations, we apply a simple dependency-based heuristic.

- We focus on tweets that contain any member of a list of source-introducing predicates (we borrow the terminology of Pareti (2012) and call this the CUE). Our complete list — shown in Table 1 — was selected mainly from the examples presented by Saurí and Pustejovsky (2012), but with reference also to Saurí's (2008) dissertation for cues that are common in Twitter. The Porter Stemmer is applied to match inflections, e.g. *denies/denied*; for irregular cases not handled by the Porter Stemmer (e.g., *forget/forgot*), we include both forms. We use the CMU Twitter Part-of-Speech Tagger (Owoputi et al., 2013) to select only instances in the verb sense. Figure 2 shows the distribution of the cues and Figure 3 shows the distribution of the cue groups. For cues that appear in multiple groups, we chose the most common group.
- We run the Stanford Dependency parser to obtain labeled dependencies (De Marneffe et al., 2006), requiring that the cue has outgoing edges of the type NSUBJ (noun subject) and CCOMP (clausal complement). The subtree headed by the modifier of the CCOMP relation is considered the **claim**; the subtree headed by the modifier of the NSUBJ relation is considered the **source**. See Figure 4 for an example.
- We use a combination of regular expressions and dependency rules to capture expressions of the type "CLAIM, *according to* SOURCE." Specifically, the PCOMP path from *according* is searched for the pattern `according to *`. The text that matches the * is the source and the remaining text other than the source is taken as the claim.
- Finally, we restrict consideration to tweets in which the source contains a named entity or twitter username. This eliminates expressions of personal belief such as *I doubt Obama will win*, as well as anonymous sources such as *Team sources report that Lebron has demanded a trade to New York.* Investigating the factuality judgments formed in response to such tweets is clearly an important problem for future research, but is outside the scope of this paper.

This heuristic pipeline may miss many relevant tweets, but since the overall volume is high, we

Report	*say*, *report*, *tell*, *told*, *observe*, *state*, *accord*, *insist*, *assert*, *claim*, *maintain*, *explain*, *deny*
Knowledge	*learn*, *admit*, *discover*, *forget*, *forgot*
Belief	*think*, *thought*, *predict*, *suggest*, *guess*, *believe*
Doubt	*doubt*, *wonder*, *ask*, *hope*
Perception	*sense*, *hear*, *feel*

Table 1: Lemmas of source-introducing predicates (cues) and groups (Saurí, 2008).

[1] We used the website `http://muckrack.com`.

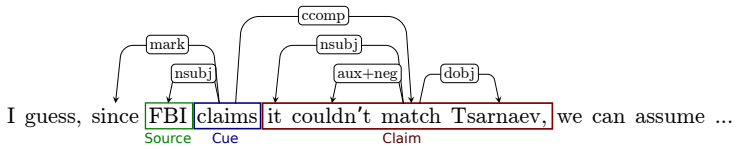

Figure 4: Dependency parse of an example message, with claim, source, and cue.

Total journalists	443
Total U.S. political journalists	103
Total tweets	959754
Tweets with cues	172706
Tweets with source and claims	40615
Total tweets annotated	1265
Unique sources in annotated dataset	766
Unigrams in annotated dataset	1345

Table 2: Count Statistics of the entire data collected and the annotated dataset

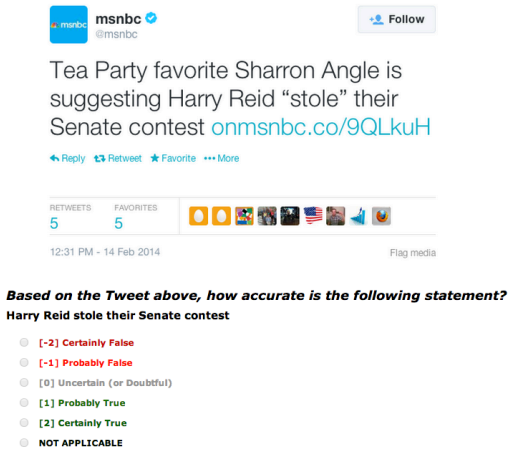

Figure 5: Turk annotation interface

prioritize precision. The resulting dataset is summarized in Table 2.

3 Annotation

We used Amazon Mechanical Turk (AMT) to collect ratings of claims. AMT has been widely used by the NLP community to collect data (Snow et al., 2008), with "best practices" defined to help requesters best design Turk jobs (Callison-Burch and Dredze, 2010). We followed these guidelines to perform pilot experiments to test the instruction set and the quality of responses. Based on the pilot study we designed Human Intelligence Tasks (HITs) to annotate 1265 claims.

Each HIT contained a batch of ten tweets and rewarded $0.10 per hit. To ensure quality control we required the Turkers to have at least 85% hit approval rating and to reside in the United States, because the Twitter messages in our dataset were related to American politics. For each tweet, we obtained five independent ratings from Turkers satisfying the above qualifications. The ratings were based on a 5-point Likert scale ranging from "[-2] Certainly False" to "[2] Certainly True" and allowing for "[0] Uncertain". We also allowed for "Not Applicable" option to capture ratings where the Turkers did not have sufficient knowledge about the statement or if the statement was not really a claim. Figure 6 shows the set of instructions provided to the Turkers, and Figure 5 illustrates the annotation interface.[2]

We excluded tweets for which three or more Turkers gave a rating of "Not Applicable," leaving us with a dataset of 1170 tweets. Within this set, the average variance per tweet (excluding "Not Applicable" ratings) was 0.585.

4 Modeling factuality judgments

Having obtained a corpus of factuality ratings, we now model the factors that drive these ratings.

4.1 Predictive accuracy

First, we attempt to determine the impact of various predictive features on rater judgments of factuality. We consider the following features:

- **Cue word**: after stemming
- **Cue word group**: as given in Table 1
- **Source**: represented by the named entity or username in the source field (see Figure 4)
- **Journalist**: represented by their Twitter ID
- **Claim**: represented by a bag-of-words vector from the claim field (Figure 4)

These features are used as predictors in a series of linear ridge regressions, where the dependent variable is the mean certainty rating. We throw out tweets that were rated as "not applicable" by a majority of raters, but otherwise ignore "not applicable" ratings of the remaining tweets. The goal of these regressions is to determine which features are predictive of raters' factuality judgments. The ridge regression regularization parameter was tuned via cross-validation in the training set. We used the bootstrap to obtain multiple training/test

[2]The data is available at `https://www.github.com/jacobeisenstein/twitter-certainty`.

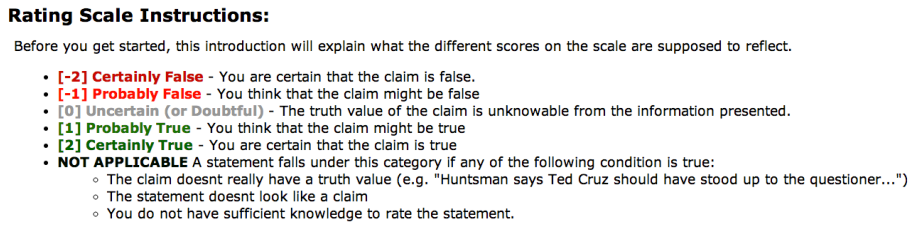

Figure 6: User instructions for the annotation task

Features	Error
Baseline	.442
Cue word	.404*
Cue word group	.42
Source	.447
Journalist	.444
Claim	.476
Cue word + cue word group	.404*
All features	.420

Table 3: Linear regression error rates for each feature group. * indicates improvement over the baseline at $p < .05$.

splits (70% training), which were used for significance testing.

Table 3 reports mean average error for each feature group, as well as a baseline that simply reports the mean rating across the training set. Each accuracy was compared with the baseline using a paired z-test. Only the cue word features pass this test at $p < .05$. The other features do not help, even in combination with the cue word.

While these findings must be interpreted with caution, they suggest that readers — at least, Mechanical Turk workers — use relatively little independent judgment to assess the validity of quoted text that they encounter on Twitter. Of course, richer linguistic models, more advanced machine learning, or experiments with more carefully-selected readers might offer a different view. But the results at hand are most compatible with the conclusion that readers base their assessments of factuality only on the framing provided by the journalist who reports the quote.

4.2 Cue words and cue groups

Given the importance of cue words as a signal for factuality, we want to assess the factuality judgments induced by each cue. A second question is whether proposed groupings of cue words into groups cohere with such perceptions. Saurí (2008) describes several classes of source-introducing predicates, which indicate how the source relates to the quoted claim. These classes are summarized in Table 1, along with frequently-occuring cues from our corpus. We rely on FactBank to assign the cue words to classes; the only word not covered by FactBank was *sense*, which we placed in predicates of perception.

We performed another set of linear regressions, again using the mean certainty rating as the dependent variable. In this case, there was no training/test split, so confidence intervals on the resulting parameters are computed using the analytic closed form. We performed two such regressions: first using only the individual cues as predictors, and then using only the cue groups. Results are shown in Figures 7 and 8; Figure 7 includes only cues which appear at least ten times, although all cues were included in the regression.

The cues that give the highest factuality coefficients are *learn* and *admit*, which are labeled as predicates of knowledge. These cues carry a substantial amount of framing, as they purport to describe the private mental state of the source. The word *admit* often applies to statements that are perceived as damaging to the source, such as *Bill Gates admits Control-Alt-Delete was a mistake*; since there can be no self-interest behind such statements, they may be perceived as more likely to be true.

Several of the cues with the lowest factuality coefficients are predicates of belief: *suggest*, *predict* and *think*. The words *suggest*, *think*, and *believe* also purport to describe the private mental state of the source, but their framing function is the opposite of the predicates of knowledge: they imply that it is important to mark the claim as the source's belief, and not a widely-accepted fact. For example, *Mubarak clearly believes he has the military leadership's support*.

A third group of interest are the predicates of report, which have widely-varying certainty coefficients. The cues *according*, *report*, *say*, and *tell*

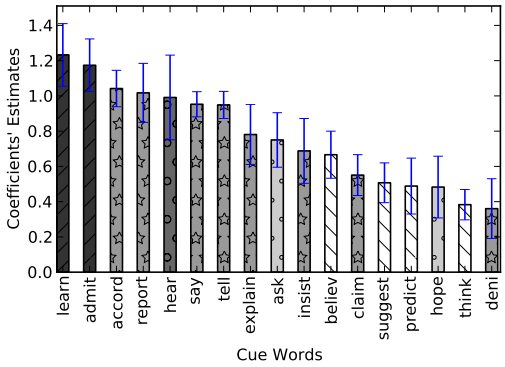

Figure 7: Linear regression coefficients for frequently-occurring cue words. Each word is patterned according to its group, shown in Figure 8.

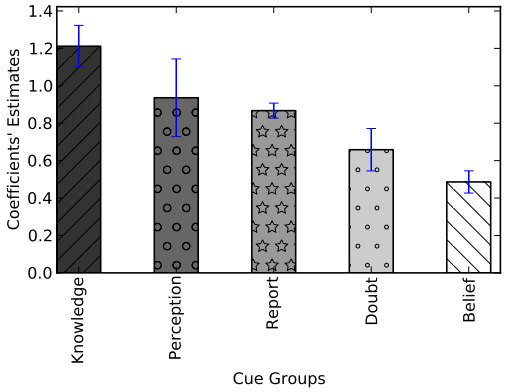

Figure 8: Linear regression coefficients for cue word group.

are strongly predictive of certainty, but the cues *claim* and *deny* convey uncertainty. Both *according* and *report* are often used in conjunction with impersonal and institutional sources, e.g., *Cuccinelli trails McAuliffe by 24 points , according to a new poll*. In contrast, *insist*, *claim*, and *deny* imply that there is uncertainty about the quoted statement, e.g., *Christie insists that Fort Lee Mayor was never on my radar*. In this case, the fact that the predicate indicates a report is not enough to determine the framing: different sorts of reports carry radically different perceptions of factuality.

5 Related work

Factuality and Veridicality The creation of FactBank (Saurí and Pustejovsky, 2009) has enabled recent work on the factuality (or "veridicality") of event mentions in text. Saurí and Pustejovsky (2012) propose a two-dimensional factuality annotation scheme, including polarity and certainty; they then build a classifier to predict annotations of factuality from statements in FactBank. Their work on source-introducing predicates provides part of the foundation for this research, which focuses on quoted statements in social media text. de Marneffe et al. (2012) conduct an empirical evaluation of FactBank ratings from Mechanical Turk workers, finding a high degree of disagreement between raters. They also construct a statistical model to predict these ratings. We are unaware of prior work comparing the contribution of linguistic and extra-linguistic predictors (e.g., source and journalist features) for factuality ratings. This prior work also does not measure the impact of individual cues and cue classes on assessment of factuality.

Credibility in social media Recent work in the area of computational social science focuses on understanding credibility cues on Twitter. Such studies have found that users express concern over the credibility of tweets belonging to certain topics (politics, news, emergency). By manipulating several features of a tweet, Morris et al. (2012) found that in addition to content, users often use additional markers while assessing the tweet credibility, such as the user name of the source. The search for reliable signals of information credibility in social media has led to the construction of automatic classifiers to identify credible tweets (Castillo et al., 2011). However, this prior work has not explored the *linguistic* basis of factuality judgments, which we show to depend on framing devices such as cue words.

6 Conclusion

Perceptions of the factuality of quoted content are influenced by the cue words used to introduce them, while extra-linguistic factors, such as the source and the author, did not appear to be relevant in our experiments. This result is obtained from real tweets written by journalists; a natural counterpart study would be to experimentally manipulate this framing to see if the same perceptions apply. Another future direction would be to test whether the deployment of cue words as framing devices reflects the ideology of the journalist. We are also interested to group multiple instances of the same quote (Leskovec et al., 2009), and examine how its framing varies across different news outlets and over time.

Acknowledgments: This research was supported by DARPA-W911NF-12-1-0043 and by a Computational Journalism research award from Google. We thank the reviewers for their helpful feedback.

References

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

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

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Marie C. de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Comput. Linguist.*, 38(2):301–333, June.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado, June. Association for Computational Linguistics.

Alfred Hermida, Seth C Lewis, and Rodrigo Zamith. 2012. Sourcing the arab spring: A case study of andy carvins sources during the tunisian and egyptian revolutions. In *international symposium on online journalism, Austin, TX, April*, pages 20–21.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 109–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, et al. 2011. The arab spring—the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication*, 5:31.

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*, pages 441–450. ACM.

Tim O'Keefe, Silvia Pareti, James R Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799. Association for Computational Linguistics.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

Silvia Pareti. 2012. A database of attribution relations. In *LREC*, pages 3213–3217.

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*, pages 1589–1599. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

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

Roser Saurí. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.