Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media

SHAN JIANG, Northeastern University, USA
CHRISTO WILSON, Northeastern University, USA

Misinformation and fact-checking are opposite forces in the news environment: the former creates inaccuracies to mislead people, while the latter provides evidence to rebut the former. These news articles are often posted on social media and attract user engagement in the form of comments. In this paper, we investigate linguistic (especially emotional and topical) signals expressed in user comments in the presence of misinformation and fact-checking. We collect and analyze a dataset of 5,303 social media posts with 2,614,374 user comments from Facebook, Twitter, and YouTube, and associate these posts to fact-check articles from Snopes and PolitiFact for veracity rulings (i.e., from true to false). We find that linguistic signals in user comments vary significantly with the veracity of posts, e.g., we observe more misinformation-awareness signals and extensive emoji and swear word usage with falser posts. We further show that these signals can help to detect misinformation. In addition, we find that while there are signals indicating positive effects after fact-checking, there are also signals indicating potential "backfire" effects.[1]

CCS Concepts: • **Human-centered computing** → Social media; • **Information systems** → Sentiment analysis; • **Social and professional topics** → Hate speech; Political speech;

Additional Key Words and Phrases: misinformation; "fake news"; fact-checking; social media; social computing

ACM Reference Format:
Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 82 (November 2018), 23 pages. https://doi.org/10.1145/3274351

1 INTRODUCTION

Misinformation takes many forms, ranging from unintentional poor journalism to deliberate hoaxes, propaganda [11, 62, 98, 117], disinformation [51, 117], and recently (and controversially) "fake news" [20, 123]. Such misinformation (broadly construed) has spread wildly since the beginning of the 2016 US presidential election cycle [53, 54]. Specifically, researchers have estimated that approximately one in four Americans visited a "fake news" website during the election [37].

Social media was and remains ground-zero for the misinformation epidemic. For example, Facebook and Twitter have banned hundreds of pages and tens of thousands of accounts, respectively, linked to the Russian Internet Research Agency for generating and promoting misinformation [92, 107]. Misinformation continues to be posted on social media by politicians, partisan

[1] Please kindly note that swear and offensive words from user comments appear in quotes and examples throughout this paper. In keeping with the norms established by prior work [12, 16], we decide not to moderate these words.

Authors' addresses: Shan Jiang, Northeastern University, USA, sjiang@ccs.neu.edu; Christo Wilson, Northeastern University, USA, cbw@ccs.neu.edu.

2573-0142/2018/11-ART82 $15.00
https://doi.org/10.1145/3274351

pundits, and even ordinary users [119]. This misinformation is viewed by millions of users [1], who sometimes comment on and share it, thus increasing its engagement and spreading it further.

One critical question is whether and how this misinformation affects individual people. Although scholars debate whether misinformation impacted the outcome of the 2016 US presidential election [1, 37], exposure to misinformation may still harm individuals by promoting partisanship, reducing trust in civic institutions, and discouraging reasoned conversation [10, 32]. Research suggests that people are indeed vulnerable to misinformation because of psychological and sociological predispositions [34, 80, 103, 122]. Furthermore, misinformation often uses inflammatory and sensational language [98, 117, 118], which may alter people's emotions. People's emotions are a core component of how they perceive their political world [66], and sometimes affect their perceived bias of news [124].

The response to misinformation by platforms has been to increase reliance on fact-checking as a means to determine the veracity and correctness of factual statements in social media posts. For example, Facebook and Google have both deployed systems that integrate fact-checking services [19, 36]. Additionally, social media users may post links to facts as a way to independently debunk misinformation. These facts can originate from different sources, ranging from first-hand experiences to scientific studies, and sometimes fact-check articles by specialized journalists. For example, a tweet posted by Donald Trump claiming that Barack Obama was born in Kenya was later fact-checked by both Snopes and PolitiFact and found to be false [25, 68] (see Figure 1). Subsequently, users on Twitter posted comments on Trump's tweet linking to the fact-check articles.

However, this reliance on fact-checking raises a second, parallel question: whether and how people respond to fact-checking itself. Some studies have found that fact-checking has corrective effects on people's beliefs [31, 39, 93, 125], while others found that it has minimal impact [55, 81] and sometimes even "backfires" on its audience [81–83]. In fact, the work of Snopes and PolitiFact has itself become politicized by those who view their work as biased, and this has led to attempts to discredit fact-check articles [78, 100, 105]. Thus, it remains unclear how people react to fact-checking in-the-wild on social media (as opposed to all prior work, which was conducted in-lab).

In this paper, we look at linguistic signals in the presence of misinformation and fact-checking using evidence from user comments on social media. We first collect a dataset of 5,303 social media posts with 2,614,374 user comments from Facebook, Twitter, and YouTube, and associate these posts to fact-check articles from Snopes and PolitiFact for veracity rulings (i.e., from true to false). Then, we build a novel emotional and topical lexicon, named *ComLex*, using a hybrid method of linguistic modeling and human labeling. Finally, we use this lexicon to analyze our dataset and perform statistical hypothesis testing. Overall, we make contributions in the following areas:

- **Misinformation.** We find that linguistic signals in user comments vary significantly with the veracity of posts. As post veracity decreases, social media users express more misinformation-awareness signals, as well as different emotional and topical signals, e.g., extensive use of emojis and swear words, less discussion of concrete topics, and decreased objectivity.
- **Detection.** We show that these linguistic signals in user comments can be used to predict the veracity of social media posts.
- **Fact-checking.** We find that there are signals indicating positive effects after fact-checking, such as more misinformation-awareness and less doubtful signals. However, there are also signals indicating potential "backfire" effects, such as increased swear word usage.
- **Lexicon.** We publish a new context specific and human validated lexicon *ComLex*, along with code and data from our paper,[2] and we present applications of ComLex in related domains.

[2] Available at: https://misinfo.shanjiang.me

The rest of the paper is organized as follows: § 2 presents related work, § 3 describes our data collection process, § 4 introduce our lexicon ComLex and its validation, § 5 and § 6 analyze linguistic signals under misinformation and fact-checking, respectively, and finally § 7 concludes.

2 RELATED WORK

The interdisciplinary nature of misinformation has drawn researchers from different areas, including computer, social, and political science. In this section, we briefly review related work and position our study within three areas: 1) the understanding, measurement, modeling, and algorithmic detection of misinformation; 2) the role of fact-checking on people's beliefs and the backfire effect; 3) lexicon-based linguistic modeling for text analysis.

2.1 Misinformation

Terminology. There is currently no agreement upon terminology across communities for false and inaccurate information. In general, there are two criteria that separate existing terminology: *veracity* and *intentionality* [109]. Some scholars prefer to use "misinformation" to broadly refer to all false and inaccurate information regardless of intent [21, 37, 42, 55, 124], while others prefer the more modern (but polarizing) term "fake news" [1, 53, 54, 109]. Other scholars restrict "misinformation" to unintentional inaccuracies, and use "disinformation" for deliberate deception [51, 123]. "Propaganda" typically refers to intentional and strictly political information [11, 62], although its veracity may vary from untruths to true but manipulative information.

In this paper, we adopt the term "misinformation" because it is inclusive and not as politicized as "fake news". The fact-checked social media posts in our dataset exhibit a wide range of veracity, so it would not be appropriate to broadly characterize them as "untrue" or "fake". Some posts even originated as jokes, sarcasm, or satire, but were later disseminated without their original context. Furthermore, we make no attempt to infer to intent of the individuals who authored the posts.

Foundations. The examination of misinformation has a long history of research. The psychological foundations are rooted in people's individual vulnerabilities. One theory that explains susceptibility to misinformation is *naïve realism*, where people tend to believe that their perceptions of reality are accurate, and views that disagree with their perceptions are uninformed, irrational, and biased [33, 103, 122]. Another theory called *confirmation bias* shows that people prefer to accept information that confirms their existing beliefs [80]. Sociological theories including *social identity theory* [15, 113] and *normative influence theory* [6] also suggest that social acceptance and affirmation are essential for people's identity and self-esteem. This causes people to choose "socially safe" options when responding to and spreading information by following the norms of their established ideological groups, regardless of the information veracity. Finally, economic theory posits that "fake news" occurs when a news publishers values short-term expansion of its customer base more than its long-term reputation, coupled with news consumers that prefer information that confirms their preexisting false beliefs [34].

People's vulnerability to misinformation affects their behavior and communication. For example, in-lab experiments have shown that exposure to biased information online [102] may significantly impact voting behavior [22, 23], while naïve information sharing may promote homophilous "echo chambers" of information [21, 57, 58].

In contrast to the above theories, there is a growing body of empirical research on people's ability to identify misinformation. Surveys that have asked people how much trust they place in different news media outlets have found that people do perceive specific outlets as biased (e.g., InfoWars) and thus do not trust news from these sources [49, 71].

Another line of work measured the spread and impact of misinformation, finding that "fake news" spread faster than truthful news [119], and that a large fraction of "fake news" are spread by "bots" [29, 104]. Misinformation is especially (and alarmingly) easy to be spread during crises, because people attempt to complete partial information using their natural sense-making processes [42], although such misinformation can sometimes be self-corrected by the crowd [5].

Given the increasing prevalence of online misinformation, it is worth investigating people's ability or lack thereof to identify misinformation. We aim to further this discussion by looking at linguistic signals in users comments on social media, leading to our first research question is: **RQ1) Are there any misinformation-awareness signals in user comments indicating their ability to identify misinformation?**

Detection. Algorithmically modeling and detecting misinformation is an emerging area of research [109]. These studies are generally divided into two categories. The *first* category analyzes *text content* to assess veracity. Some researchers use the claims included in text to do automatic fact-checking by comparing the consistency and frequency of claims [8, 65], or by attempting to infer a claim from existing knowledge graphs [18, 41, 108, 121, 126]. Others note that fake or hyper-partisan news publishers usually have malicious intent to spread misinformation as widely as possible, which causes them to adopt a writing style that is inflammatory and sensational. Such stylistic features can distinguish false from truthful news [28, 94, 98, 117, 118, 120].

In this study, we ask whether the inflammatory content and sensational writing style that is sometimes characteristic of misinformation affects the emotional and topical signals that people express in their social media comments. Previous research has shown that linguistic signals, e.g., usage of emojis, can be used to infer people's actual emotional states [27, 48, 61, 101, 127]. Thus, our second research question is: **RQ2) Do emotional and topical signals in user comments vary with post veracity?**

The *second* category of detection algorithms leverage *social context* to predict misinformation, i.e., users' different behaviors when reacting to false or truthful news. These behaviors including different stances and discussed topics than the original threads [46, 96, 112], as well as different propagation network structures between fake and truthful news [38, 45]. Recent work has also proposed tools that actively solicit and analyze "flags" on misinformation from users [115].

Different detection models have their own limitations, e.g., content-based models may have difficulty analyzing video posts, active flagging systems require extra labor from users, etc. Since our approach measures the passive response of users and is orthogonal to the content format of posts, we propose our third research question: **RQ3) Can linguistic signals in users comments help to detect misinformation?**

RQ1) to **RQ3)** potentially have substantial implications on the design of social computing systems. If user comments on misinformation significantly deviate from typical conversations (e.g., extensive use of swear words, shown in § 5.1), they could easily deteriorate into trolling [17], harassment [87], or hate speech [77]. Understanding and detecting the linguistic variants present in these comment threads may help when implementing intervention and moderation systems [35, 44].

Datasets. Previous public datasets of misinformation include *Liar* [120], *BuzzFeedNews* [94], *CREDBANK* [72, 73], etc. These datasets usually contain the textual content of social media posts and news articles along with veracity scores based on domain-specific metrics. Our published dataset is complementary to these efforts, as we focus on analyzing user comments.

2.2 Fact-Checking

Another line of research focuses on the effects of fact-checking. Many in-lab experiments have examined the effects of fact-checking on human behaviors, but unfortunately they reveal drastically different behaviors in different contexts. A fact-check against a false rumor that the flu vaccine gave people the flu significantly reduced people's belief in the rumor, but also reduced some people's willingness to vaccinate because of side effects [82, 83]. However, later research failed to duplicate the results [39]. This phenomenon is called the "backfire" effect, where attempting to intervene against misinformation only entrenches the original, false belief further [81].

Even without the backfire effect, there are several experiments that found that fact-checking has limited corrective effects [55, 81]. However, others found that people are willing to accept fact-checking even when the information challenges their ideological commitments [31, 93, 125].

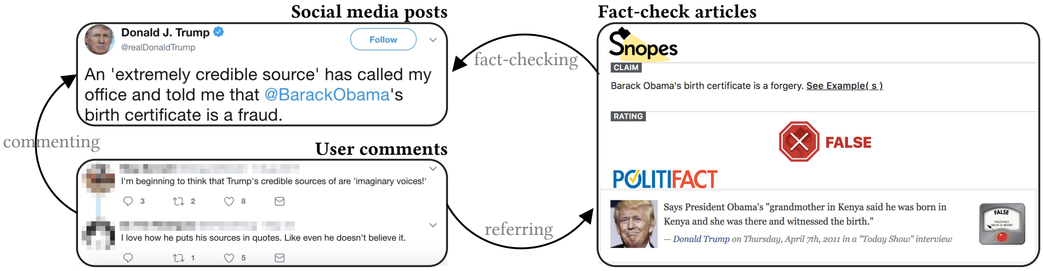
Major fact-checking organizations include Snopes [69], Politifact [106], and FactCheck.org [43]. These websites use facts and evidence to determine the veracity and correctness of factual claims in news articles, political speeches, social media posts, etc. In general, their verdicts have a very high degree of agreement with each other [3, 4]. However, the corrective effects of these websites has not been investigated in detail. Previous research has shown that fact-check articles posted on social media are likely to get more exposure when shared by a friend instead of strangers [40, 67], but that including biographical information about the author of the fact-check in the article itself reduces the effectiveness [33]. On online platforms, alert messages and tags that warn users to the presence of misinformation can help reduce the influence of this content [23, 90].

Building on this line of work, we propose two research questions: **RQ4) After a social media post is fact-checked, are there any linguistic signals in comments indicating positive effects?** and **RQ5) After a post is fact-checked, are there any linguistic signals in comments indicating backfire effects?** Although related questions were investigated in previous research using in-lab experiments, our context and approach are different. Context is an important variable when examining the effect of fact-checking, as studies under different conditions often generate different results that cannot be generalized. Furthermore, recent studies have proposed integrating fact-checking results [50] or bias warnings [23] into social computing systems. **RQ4)** and **RQ5)** speak directly to the feasibility of such proposals.

2.3 Linguistic Modeling

Using dictionary categories is one of the traditional ways to perform computational analysis of text corpora [47, 85]. Originally, these techniques focused on *sentiment analysis*, with only positive and negative sentiment labels on words. Over time, researchers built more fine-grained lexicons for more sophisticated emotions and topics.

There are several existing lexicons that are commonly used to perform text analysis. The most extensively used and validated lexicon is Linguistic Inquiry and Word Count (LIWC) [88, 114], which contains both emotional, topical, and syntactic categories. An alternative for LIWC is Empath [26], which is an automatically generated lexicon that was later manually validated by crowdsourced workers. Empath has strong correlations with LIWC within overlapping categories of words. NRC Word-Emotion Association Lexicon (EmoLex) [75, 76] is another human curated lexicon that is structured around Plutchik's wheel of emotions [91]; it includes eight primary emotions (anger, anticipation, joy, trust, fear, surprise, sadness, and disgust) and two additional classes for all positive and negative emotions. Other lexicons include the General Inquirer (GI) [110] which has more topics than LIWC but fewer emotions, and Affective Norms for English Words (ANEW) [13] and SentiWordNet [7, 24] which have more emotions than LIWC.



An 'extremely credible source' has called my office and told me that @BarackObama's birth certificate is a fraud.

[redacted] I'm beginning to think that Trump's credible sources of are 'imaginary voices!'

[redacted] I love how he puts his sources in quotes. Like even he doesn't believe it.

Barack Obama's birth certificate is a forgery. See Example(s)

Says President Obama's "grandmother in Kenya said he was born in Kenya and she was there and witnessed the birth."

— *Donald Trump on Thursday, April 7th, 2011 in a "Today Show" interview*

Fig. 1. **Illustration of social media and fact-checking website interaction.** Following the publication of a post on Twitter, Facebook, YouTube, etc., Snopes and PolitiFact fact-check it and rate its veracity. Meanwhile, users comment on the post and sometimes refer to fact-check articles once they are released.

Although the psychological foundations of the above lexicons are solid, they are extracted from general text, and usually do not perform well when analyzing text from specific contexts [56]. In the case of social media, existing lexicons such as NRC Hashtag Emotion Lexicon (HashEmo) [74] and others [9, 59] are mostly automatically generated and not manually validated.

An alternative approach to perform text analysis is to *learn* a lexicon for a specific domain. Recently, one extensively used method is to learn vector representations of word embeddings [63, 70, 89] and then use unsupervised learning to cluster words [26]. This methods has been used by some studies in the misinformation domain to analyze stylistic features of articles [98, 118].

We found existing lexicons to be insufficient for our research because they only offered a limited number of word categories. Therefore we constructed a new context-specific lexicon called *ComLex* with emotional and topical categories for user comments on fact-checked social media posts. Additionally, we use EmoLex and LIWC throughout our study as supporting evidence to validate our findings. We also present a performance evaluation between lexicons in terms of predictive ability in § 4.3.

3 DATA

Our study is based on a dataset of user comments on social media posts that have been fact-checked. In the section, we discuss how we collected this dataset and give an overview of the data.

3.1 Data Collection

The interaction between social media and fact-checking websites is shown in Figure 1. Politicians, news organizations, or other individuals publish posts on social media websites such as Twitter, Facebook, YouTube, etc. Some of these posts are selected for fact-checking by specialized journalists at websites such as Snopes and PolitiFact, who then publish articles containing evidence for or against the claims and reasoning within the posts, as well as a veracity ruling for the posts. Meanwhile, users may comment on the posts, which sometimes refer to the fact-check articles.

To gather this data (i.e., posts and their associated comments and fact-check articles), we use the fact-checking websites Politifact and Snopes as starting points. We choose PolitiFact and Snopes because **a)** they are both confirmed by the International Fact-Checking Network (IFCN) to be non-partisan, fair, and transparent fact-checking agencies; and **b)** they list their sources and rulings in a structured format that is easy to automatically parse. We crawled all the fact-check articles from Politifact and Snopes, and then filtered this set down to articles that point specifically to social media posts on Facebook, Twitter, and YouTube (e.g., the one from Figure 1). We extracted the

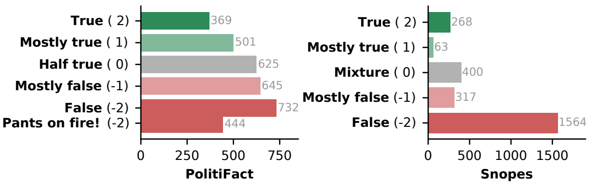


Fig. 2. **Distribution of veracity for posts from PolitiFact and Snopes and mapping to ordinal values.** We map textual descriptions of veracity to ordinal values. We ignore descriptions that cannot be categorized such as *full flop, half flip, no flip* from PolitiFact and *legend, outdated, unproven, undetermined, research in progress, miscaptioned, misattributed, correct attribution, not applicable, etc.* from Snopes.

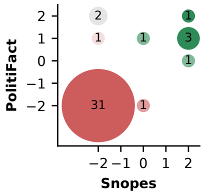


Fig. 3. **Veracity of posts fact-checked by both PolitiFact and Snopes.** The veracity rulings are strongly correlated ($\rho = 0.671^{***}$).

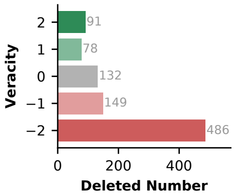


Fig. 4. **Distribution of veracity for deleted posts.** The likelihood of post deletion is negatively correlated with the veracity of posts ($r_{pb} = -0.052^{***}$).

unique post ID[3] and veracity rating from these articles. Finally, we used the Facebook, Twitter, and YouTube platform APIs to crawl all of the user comments on the fact-checked posts by leveraging their unique IDs.

3.2 Data Overview

Overall, we collected 14,184 fact-check articles from Politifact and 11,345 from Snopes, spanning from their founding to January 9, 2018. After filtered out all articles whose sources were not from Facebook, Twitter, or YouTube, our dataset contained 1,103 social media posts from Facebook, 2,573 from Twitter, and 2,753 YouTube videos.

Note that PolitiFact and Snopes have different ruling criterion and therefore different textual descriptions for post veracity. To make them comparable, we translated their descriptions to a scale from -2 to 2 using the mapping shown in Figure 2. We view *pants on fire!* and *false* as -2 for PolitiFact, and ignore descriptions that cannot be categorized such as *full flop, half flip, no flip* from PolitiFact and *legend, outdated, unproven, undetermined, research in progress, miscaptioned, misattributed, correct attribution, not applicable,* etc. from Snopes. After mapping and removing descriptions that cannot be categorized, we kept 5,303 posts. 41 of 5,303 (0.77%) of the mapped posts were checked by both PolitiFact and Snopes, and their veracity rulings from the two websites are strongly correlated (Spearman $\rho = 0.671^{***}$) as shown in Figure 3, which is consistent with previous observations [3, 4].

Finally, we collected user comments on the 5,303 fact-checked social media posts using their respective platform APIs. We note that 1,659 (31%) of the posts were no longer available because they were either deleted by the platform or by their authors, of which 1,364 (82%) had veracity ≤ 0. This finding may be attributable to platforms' efforts to fight misinformation [30, 111]. In addition, there were 757 posts with zero comments. From the remaining posts we collected 1,672,687 comments from Facebook, 113,687 from Twitter, and 828,000 from YouTube.

Before moving on, we take a deeper look at the deleted posts. The distribution of their veracity is shown in Figure 4. We observe that the likelihood of post deletion increases significantly as veracity decreases (Point Biserial $r_{pb} = -0.052^{***}$). This means that, overall, our dataset is missing some deeply misleading and/or untrue posts and their associated comments. These omissions will make our model under-estimate the effect of misinformation and fact-checking. Therefore, our statistics

[3] Although the post ID formats for Facebook, Twitter, and YouTube are not the same, they are all structured and relatively easy to automatically parse.

should be viewed as conservative lower bounds on the linguistic variants in user comments in the presence misinformation and fact-checking.

3.3 Ethics

We were careful to obey standard ethical practices during our data collection. We only used official APIs from the social networks to collect data, we did not make any "sock puppet" accounts, and we rate limited our crawlers. All of the posts and associated comments are publicly accessible, and our dataset does not contain any posts or comments that were deleted or hidden by their authors prior to our crawl in January 2018. The datasets that we plan to publish are fully anonymized, i.e., all user IDs are removed.

4 MODEL

Using the collected dataset, we build a new lexicon called *ComLex* based on the corpus of user comments. In this section, we discuss how we constructed the lexicon, and then present three complementary validation tests based on (1) human raters, (2) comparisons with two representative lexicons from prior work, and (3) re-evaluation of datasets used in prior work.

4.1 A Lexicon of User Comments: ComLex

We generate ComLex using a combination of learning word embeddings and unsupervised clustering. We first build a corpus of user comments by applying standard text preprocessing techniques using *NLTK* [60], including tokenization, case-folding, and lemmatization. Importantly, we choose not to remove punctuation and non-letter symbols because such symbols may carry meanings for our task, such as exclamation "!" and smile ":)". This also allow us to keep emojis, which are important "words" for our analysis because they enable users' to express emotional signals, sometimes even more significantly than with text [2, 27]. In addition, we replaced all URLs that link to a Snopes or PolitiFact webpages with the special tokens *snopesref* or *politifactref*. This enables us to group all fact-checked posts from Snopes and PolitiFact together, respectively, and later learn their semantics.

Next, we learn word embeddings from the clean corpus, i.e., transform words into vector space to provide numerical representations of each word in the corpus. To do this, we use *gensim* [99] to learn *Word2Vec* [70] representations, and use a 100-dimension vector to represent each word. To avoid noise, we only kept words that appear ≥ 100 times in the corpus. Subsequently, we apply spectral clustering [79] to divide our vectors into 300 disjoint clusters, with each cluster contains words with similar semantics. Finally, we manually examined each cluster and provide a suitable name and additional descriptive information for it. The final, labeled clusters of words are ComLex.

For each cluster in a given lexicon (e.g., ComLex, EmoLex, or LIWC), we compute a statistic for each user comment based on the word frequencies in each cluster. We then normalize these statistics by the total word frequencies in a cluster. Our analytical sections mainly focus on the statistics from ComLex, but we also provide results from EmoLex and LIWC as support.

4.2 Human Evaluation

To validate the robustness of our lexicon, we designed a survey that included two rating questions:

> **Rating 1: How closely, in terms of semantics, are words in each cluster related to each other?** Please provide a rating from 1 to 5 with 1 being not related and 5 being extremely related for each word cluster. *e.g., "apple, banana, peach, grape, cherry" should be considered extremely related (5) since they are all fruits; "apple, sky, happy, tomorrow, birds" should be considered not related (1).*
>
> **Rating 2: How accurately do the name and additional information describe the word cluster?** Please provide a rating from 1 to 5 with 1 being not accurate and 5 being extremely accurate for each word cluster. *e.g., "fruit" should be considered extremely accurate (5) for a cluster of apple, banana, peach, grape, cherry; "weather" should be considered not accurate (1).*

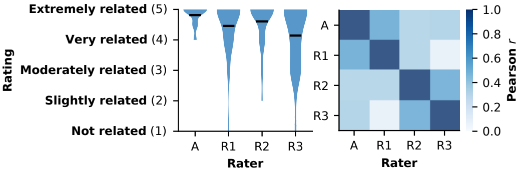


Fig. 5. **Survey results of rating 1.** The left violin plot shows the rating distribution of four raters and the right heatmap shows the inter-rater correlations. Words in clusters are rated on average above "very related" ($\overline{\mu} = 4.506$) with moderate inter-rater agreement ($\overline{r} = 0.531$).

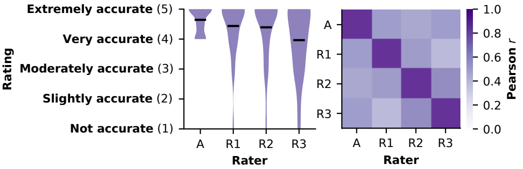


Fig. 6. **Survey results of rating 2.** The left violin plot shows the rating distribution of four raters and the right heatmap shows the inter-rater correlations. Cluster names and additional information are rated on average above "very accurate" ($\overline{\mu} = 4.359$) with strong inter-rater agreement ($\overline{r} = 0.675$).

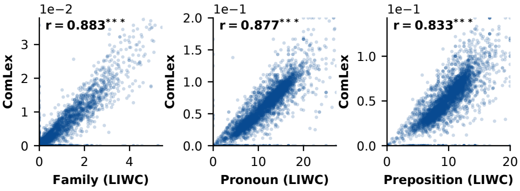


Fig. 7. **Comparison with LIWC.** Each scatter plot shows the correlation of ComLex and LIWC for a similar word cluster. Selected clusters including *family* ($r = 0.883^{***}$), *pronoun* ($r = 0.877^{***}$) and *preposition* ($r = 0.833^{***}$) show very strong correlation.

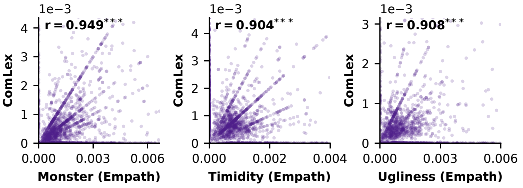


Fig. 8. **Comparison with Empath.** Each scatter plot shows the correlation of ComLex and Empath for a similar word cluster. Selected clusters including *monster* ($r = 0.949^{***}$), *timidity* ($r = 0.904^{***}$) and *ugliness* ($r = 0.908^{***}$) show very strong correlation.

Each question asked for a rating on 5-point Likert scale, with descriptive adverbs chosen from [14]. The authors (*A* in Figures 5 and 6) took the survey first and gave ratings for all learned clusters. We then keep only the top 56 of 300 (18.7%) clusters with ratings ≥4 for both questions. After this filtering process, we then asked three independent raters (*R1*, *R2*, and *R3*) to take the survey to rate the remaining 56 clusters to ensure semantic closeness and accurate cluster names.

Figure 5 shows the results of the first survey question. The violin plot shows the distribution of four raters, among which the authors gave the highest average rating ($\mu_A = 4.814$) and R3 gave the lowest ($\mu_{R3} = 4.143$). Overall, words in clusters are rated above "very related" on average (mean average $\overline{\mu} = 4.506$), and the difference in μ among raters is significant (Kruskal-Wallis $H = 11.3^{*}$). The heatmap shows the inter-rater agreement represented by Pearson correlation, demonstrating moderate agreement among the raters on average (mean Pearson $\overline{r} = 0.531$). Figure 6 shows the results of the second survey question. As shown in the violin plot, the authors gave the highest average rating ($\mu_A = 4.643$) and R3 gave the lowest ($\mu_{R3} = 3.964$). Overall, cluster names and additional information are rated above "very accurate" on average ($\overline{\mu} = 4.359$), and the difference in μ among raters is significant ($H = 10.8^{*}$). As shown in the heatmap, the raters are strongly agreed with each other on average ($\overline{r} = 0.675$). These results show that ComLex is perceived as valid by humans.

4.3 Comparison with LIWC and Empath

Next, we compare ComLex with two existing lexicons: LIWC and Empath. LIWC is arguably the most extensively used lexicon, while Empath is generated in a similar manner to ComLex. We pair the statistics of user comments mapped using these lexicons and then select similar clusters to compare their correlation.

Table 1. **Application of ComLex on other datasets.** The upper part of the table shows the performance of ComLex at detecting deception in hotel reviews. It outperforms human judges, GI, and LIWC, but is not as accurate as learned unigrams. The lower part of the table shows the performance of ComLex at detecting sentiment of movie reviews. It outperforms human judges and is nearly as accurate as learned unigrams.

Dataset	Lexicon	Model	Accuracy*
Hotel reviews [84]	Human judges		56.9% - 61.9%
	GI	SVM	73.0%
	LIWC		76.8%
	ComLex		**81.4%**
	Learned unigrams		88.4%
Movie reviews [86]	Human judges		58.0% - 69.0%
	ComLex	SVM	**72.3%**
	Learned unigrams		72.8%

**All accuracy data are drawn from the original papers [84, 86] except for ComLex.*

Figure 7 shows the comparison with LIWC. Each scatter plot shows the correlation of a similar word cluster between ComLex and LIWC. ComLex shows very strong correlation with LIWC in similar clusters such as *family* (Pearson $r = 0.883^{***}$), *pronoun* ($r = 0.877^{***}$) and *preposition* ($r = 0.833^{***}$). Figure 8 shows the comparison with Empath. Again, ComLex shows very strong correlation with Empath in similar clusters such as*monster* ($r = 0.949^{***}$), *timidity* ($r = 0.904^{***}$) and *ugliness* ($r = 0.908^{***}$). This step shows that statistics derived from ComLex and LIWC/Empath are similar for overlapping word categories.

4.4 Applications

Lastly, we test ComLex on previously released datasets to evaluate the generality and performance of ComLex when applied to related domains. In the following experiments, we run ComLex datasets of hotel and movie reviews, respectively, and build predictive models to evaluate the performance of ComLex. To compare our accuracy with the ones reported in the original papers, we adopt the same learning model (Support Vector Machine, SVM), and report the same evaluation metric (accuracy). Note that the datasets we choose have balanced binary labels, therefore accuracy is a reasonable metric for evaluation.

The first application uses a hotel dataset of 800 positive reviews [84], of which half are truthful reviews from TripAdvisor and half are deceptive reviews from Amazon Mechanical Turk. The task is to predict whether a review is truthful or deceptive. The original paper reported the accuracy of three human judges, the existing GI and LIWC lexicons, and domain-specific learned unigrams. We run 10-fold cross validation using vectors mapped by ComLex and report our results in Table 1. We see that ComLex outperforms the human judges, GI, and LIWC, but not the learned unigrams.

The second application uses a movie dataset of 1,400 reviews [86], of which half are labeled as positive and half as negative. The task is to predict whether a review is positive or negative. The original paper reported the accuracy of three human judges and domain-specific learned unigrams. We run 10-fold cross validation using vectors mapped by ComLex and report our results in Table 1. Again, ComLex outperforms the human judges, and is nearly as accurate as the learned unigrams.

ComLex is generated using our dataset of user comments specifically on misinformation, yet it is essentially a lexicon of user comments in general, and leverages comments from multiple sources, i.e., Facebook, Twitter, and YouTube. This step demonstrates that ComLex can be broadly and flexibly applied to other related domains with reasonable performance.

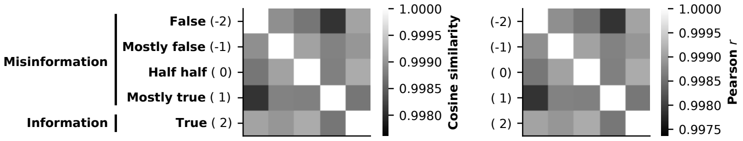


Fig. 9. **Similarity matrix over veracity.** Heatmaps shows the similarity matrix over veracity using cosine similarity and Pearson correlation respectively. Using both measures, clear patterns of decreasing similarity are visible from -2 to 1, but the trend does not hold for 2.

5 LINGUISTIC SIGNALS UNDER MISINFORMATION

In this section, we focus on linguistic signals in the presence of misinformation. Using our dataset and ComLex, we analyze how linguistic signals vary versus the veracity of social media posts. Considering that fact-checking articles may be a strong confounding variable in this analysis, we only examine user comments that were generated *before* the post was fact-checked.

Before we analyze specific linguistic clusters, we first take a look at the overall linguistic similarity between user comments on posts of varying veracity. To do this, we group user comments by the veracity (-2 to 2) of the post and compute the mean of all vectors in that veracity group. We then compute the cosine similarity and Pearson's r between different veracity groups.

As shown in Figure 9, there is a clear pattern from *false* (-2) to *mostly true* (-1): users' comments are self-identical (1.0), and the similarity gradually decrease as the comparisons become more distant (e.g., *false* versus *mostly true*). However, this pattern does not hold for comments on posts whose veracity is *true* (2). This observation holds regardless of whether cosine similarity or Pearson correlation is used to compute distance. This motivates us to split our research questions into different experiments by looking at the *degree* of misinformation and the *existence* of misinformation separately. In the following sections, we will first look at how linguistic signals vary versus the *degree* of misinformation by analyzing user comments from posts rated from -2 to 1, and then looking at how linguistic signals vary versus the *existence* of misinformation by comparing posts rated 2 to those rated < 2.

5.1 Linguistic Signals versus Degree of Misinformation

In this section we examine the whether there are differences in the emotional and topical signals expressed in user comments based on the degree of misinformation in the original post. We perform Spearman correlation tests between each word cluster's normalized frequency and each veracity value, and report significant results of ρ in Figure 10.

With regards to **RQ1**, we observed that **the usage likelihoods for several word clusters that express misinformation-awareness are negatively correlated with veracity.** These clusters include verbs that describe fakes (*fake*, *mislead*, *fabricate*, etc., $\rho = -0.087^{***}$), and nouns for very fake content (*hoax*, *scam*, *conspiracy*, etc., $\rho = -0.045^{*}$) and somewhat fake content (*propaganda*, *rumor*, *distortion*, etc., $\rho = -0.046^{*}$), e.g., "this is fake news", "this is brainwash propaganda", etc. This means social media users are more likely to use these misinformation-aware words when commenting on posts that are ultimately proven to have low veracity. Combining these word clusters together, their mean values increase from 0.0025 to 0.0033 as veracity decreases from 1 to -2, i.e., on average, each word that identifies misinformation has a 9.7% greater chance of appearing in each user comments with one decrement in veracity. This observation is, in a different direction, supported by EmoLex where **trust declines as misinformation increases**. We observe positive correlations between veracity and word clusters that express *trust* (*accountable*, *lawful*, *scientific*, etc., $\rho = 0.063^{**}$). This means people are less likely to express trust when commenting on posts that are ultimately shown to have low veracity. In terms of effect size, the mean value of the trust

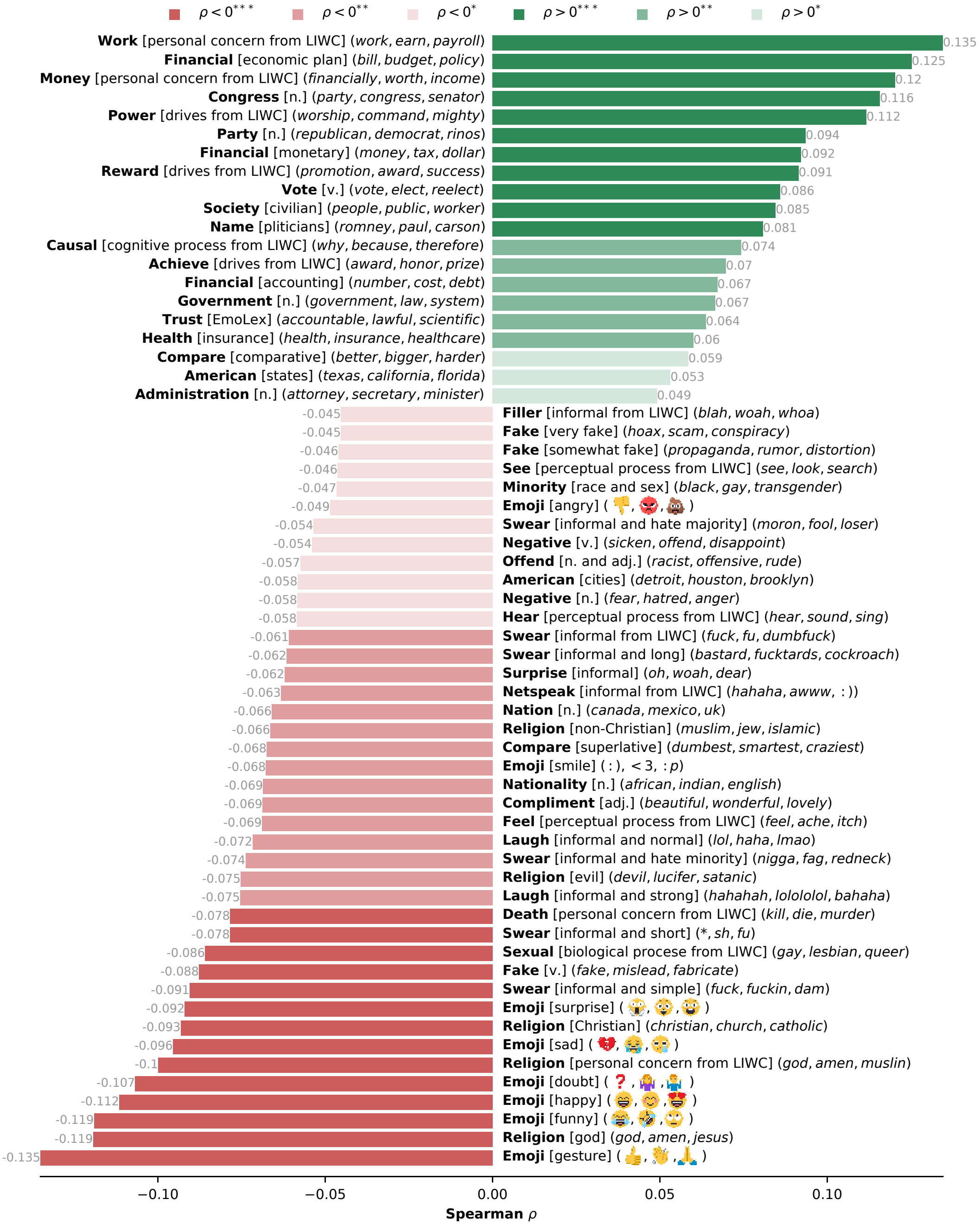


Fig. 10. **Linguistic signals versus degree of misinformation.** Clusters with significance ρ are plotted, ranked by the sign and strength of correlation. A positive ρ indicates that the statistic increases with veracity, and vice versa. Clusters are labeled in the figure using the format: **name** [additional information] (*three example words*).

category decreases from 0.0554 to 0.053 as veracity decreases from 1 to -2, i.e., on average, people are using 1.4% less of these words with each single decrement of veracity.

The first evidence that supports **RQ2** is that **the usage of emojis increases as misinformation increases**. We observe significant negative correlations for eight clusters of emoji, including *gesture* (👍, 👋, 🙏, etc., $\rho = -0.135^{***}$), *funny* (😂, 🤣, 🙄, etc., $\rho = -0.119^{***}$), *happy* (😄, 😊, 😍, etc., $\rho = -0.112^{***}$), *question* (❓, 🤷‍♀️, 🤷‍♂️, etc., $\rho = -0.107^{***}$), *sad* (💔, 😭, 😟, etc., $\rho = -0.096^{***}$), *surprise* (😱, 😳, 😨, etc., $\rho = -0.092^{***}$), and *angry* (👎, 😡, 💩, etc., $\rho = -0.049^{*}$), e.g., “so ridiculous 😂😂”, “really? 😭”, “i smell bull 💩”, etc. This means people are more likely to use these emoji when commenting on posts that are ultimately proven to have low veracity. Combining these emoji clusters together, their mean values increase from 0.0015 to 0.005 as veracity decreases from 1 to -2, i.e., users are 49.4% more likely to use emojis with each single decrement in veracity value. Given the popularity of emojis [27, 61], we view them as important proxies for people’s actual emotional state [48, 101, 127] when confronted with misinformation.

The second evidence that supports **RQ2** is that **the usage of swear words increases as misinformation increases**. We observe significant negative correlations for five clusters of swear words, including popular swear words (*fuck*, etc., $\rho = -0.091^{***}$), shortened or moderated swear words (*, *fu*, etc., $\rho = -0.078^{***}$), hateful terms against minority groups ($\rho = -0.074^{**}$), long and complicated swears (*bastard*, etc., $\rho = -0.062^{**}$), and belittling words (*moron*, *fool*, *loser*, etc., $\rho = -0.054^{*}$). This means people are more likely to swear or use hateful terms towards other users (including the author of the post) when commenting on posts that are eventually found to have low veracity. Combining these swear clusters together, their mean values increases from 0.0034 to 0.0046 as veracity decreases to -2, i.e., on average, users are using 16.3% more swear words with one decrement in veracity value. This observation is further supported by LIWC’s swear word category (*fuck*, etc., $\rho = -0.061^{**}$). People associate swear words with their own emotional states, and these words affect the emotional states of others [97]. In our data, we observe an increasing amount of people using negative or offensive words in comments as veracity decreases and swear words increase. This includes negative correlations with a cluster of negative verbs (*sicken*, *offend*, *disappoint*, etc., $\rho = -0.054^{*}$) and another of offensive nouns and adjectives (*racist*, *offensive*, *rude*, etc., $\rho = -0.057^{*}$) with an effect size of 3.7%.

The third evidence that supports **RQ2** is that **discussion of concrete topics declines as misinformation increases**. We observe significant positive correlations for 12 clusters of words about concrete political topics, including financial clusters about economic plans (*bill*, *budget*, *policy*, etc., $\rho = 0.125^{***}$) and monetary issues (*money*, *tax*, *dollar*, etc., $\rho = 0.092^{***}$), and clusters about congress (*party*, *congress*, *senator*, etc., $\rho = 0.116^{***}$), party (*republican*, *democrat*, *rinos*, etc., $\rho = 0.094^{***}$), voting (*vote*, *elect*, *reelect*, etc., $\rho = 0.086^{***}$), society (*people*, *public*, *worker*, etc., $\rho = 0.085^{***}$), government (*government*, *law*, *system*, etc., $\rho = 0.067^{**}$), health (*health*, *insurance*, *healthcare*, etc., $\rho = 0.06^{**}$), administration (*attorney*, *secretary*, *minister*, etc., $\rho = 0.049^{*}$), and references to states ($\rho = 0.053^{*}$) and politicians ($\rho = 0.081^{**}$). This means people are more likely to talk about concrete topics on posts with higher veracity. Combining these clusters together, their mean value increases from 0.046 to 0.065 as veracity value increases from -2 to 1, i.e., on average, users are 12.2% more likely to use words in these clusters with one increment in veracity value. This observation is supported by LIWC’s word categories involving concrete topics, including *work* ($\rho = 0.135^{***}$), *money* ($\rho = 0.120^{***}$), *power* ($\rho = 0.091^{***}$), and *achieve* ($\rho = 0.07^{**}$).

The fourth evidence that supports **RQ2** is that **objectivity declines as misinformation increases**. We observe that users are more likely to use superlatives (*dumbest*, *smartest*, *craziest*, etc., $\rho = -0.068^{**}$), e.g., “dumbest thing i’ve seen today”, with an effect size of 25.5% with each single decrement in veracity value. At the same time, we observe that users are less likely to use comparatives (*better*, *bigger*, *harder*, etc., $\rho = 0.059^{*}$), e.g., “she would do better”, with an effect size

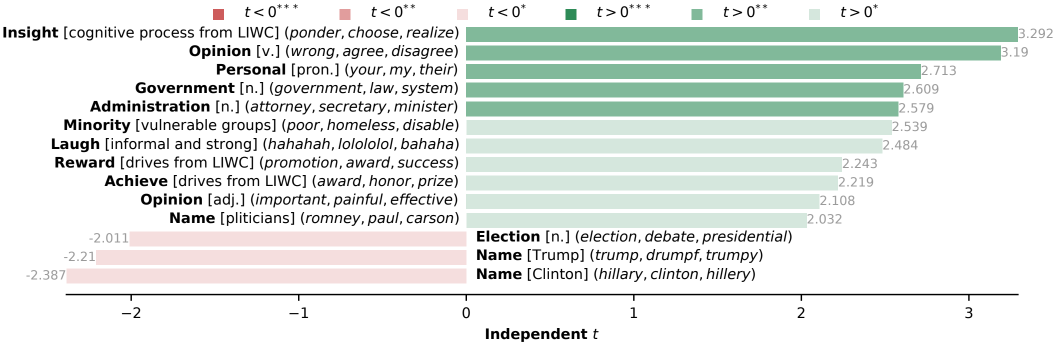


Fig. 11. **Linguistic signals versus existence of misinformation.** Clusters with significance independent t are plotted, ranked by the sign and strength of difference. A positive t indicates that the mean of the statistic for accurate information is higher than misinformation, and vice versa.

of 15.6% with each single decrement in veracity value. This observation is also supported by LIWC, where people use less causal inference (*why*, *because*, *therefore*, etc., $\rho = 0.074^{*}$) as misinformation increases. This implies that subjectivity increases and objectivity decreases as the veracity of the underlying post decreases. The relationship between subjectivity and objectivity has long been studied within the context of people's emotional states in sociology [52].

5.2 Linguistic Signals versus Existence of Misinformation

We now look at the differences in the emotional and topical signals of user comments in relation to the existence of misinformation (i.e., posts with veracity value 2 versus posts with value < 2). We report statistically significant independent t in Figure 11. Our findings are similar to § 5.1, such as an increased likelihood of discussion about concrete political topics on true posts. This includes *government* ($t = 2.609^{**}$), *administration* ($t = 2.579^{**}$) and *minority* ($t = 2.539^{*}$). In terms of effect size, combining these clusters together, their mean is 0.0074 for misinformation and 0.01 for true posts, which represents a 35.1% difference. Similarly, we also observe that concrete topical categories from LIWC such as reward ($t = 2.243^{*}$) and achieve ($t = 2.219^{*}$) are significant.

Another supporting evidence for **RQ2** is **the increased likelihood of personal opinions on true posts**. We observe that users are more likely to express their opinions in a concrete manner, including opinionated adjectives (*important*, *painful*, *effective*, etc., $t = 2.108^{*}$), and personal opinions (*wrong*, *agree*, *disagree*, etc., $t = 3.190^{**}$), e.g., "this is important", "i agree with you", etc. These two clusters have a mean of 0.0036 for misinformation and 0.0049 for true posts, which represents a 36.1% difference. This is also supported by LIWC in its *insight* category, which is a subset of cognitive process ($t = 3.292^{**}$).

We also found that users are 43.1% less likely to mention the election ($t = -2.011^{*}$), Trump ($t = -2.210^{*}$), and Clinton ($t = -2.387^{*}$) when commenting on true posts. One possible explanation for this is that true posts invite discussion of more original and substantive topics, versus 2016 election coverage itself which was polarizing and prone to misinformation [1, 37].

5.3 Detection

We now explore **RQ3**: whether user comments can be used to predict the veracity of posts. We built predictive models using the vectors of user comments as input to predict the veracity value of associated posts. We apply both linear models such as linear regressions and linear SVM, as well as non-linear neural network models. We implement these models using mapped word vectors from

Table 2. **Results of detection models.** Spearman's ρ, LRAP, and LRL are reported for each model and lexicon. All models using any lexicon perform better than random guessing. ComLex performs better than EmoLex and LIWC. Using ComLex, regulated models perform better than the unregulated model, and non-linear models perform better than linear models.

	EmoLex			LIWC			ComLex		
Model	ρ	**LRAP**	**LRL**	ρ	**LRAP**	**LRL**	ρ	**LRAP**	**LRL**
Random	0.009	0.298	0.877	0.009	0.298	0.877	0.009	0.298	0.877
Linear regression	0.096	0.332	0.835	0.100	0.336	0.830	0.160	0.351	0.811
Linear regression (L1)	0.094	0.332	0.835	0.123	0.336	0.830	0.194	0.344	0.820
Linear regression (L2)	0.095	0.332	0.835	0.110	0.337	0.829	0.193	0.344	0.820
SVM (linear kernel, L1)	0.082	0.332	0.835	0.124	0.339	0.826	0.197	0.360	0.800
Linear SVM (linear kernel, L2)	0.095	0.332	0.835	0.107	0.337	0.828	0.189	0.346	0.817
Neural network (50 nodes)	0.063	0.331	0.836	0.136	0.345	0.819	0.207	0.360	0.800
Neural network (100 nodes)	0.073	0.332	0.835	0.136	0.343	0.821	0.213	0.373	0.784
Neural network (500 nodes)	0.078	0.334	0.832	0.117	0.346	0.817	**0.216**	0.388	0.765
Neural network (1000 nodes)	0.081	0.331	0.836	0.116	0.343	0.821	0.214	**0.392**	**0.760**

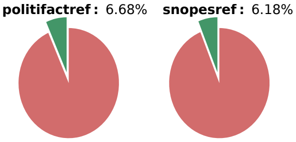


Fig. 12. **Percentage of politifactref and snopesref.** Each pie chart shows the percentage of posts that contains *politifactref* or *snopesref* over all posts checked by the website.

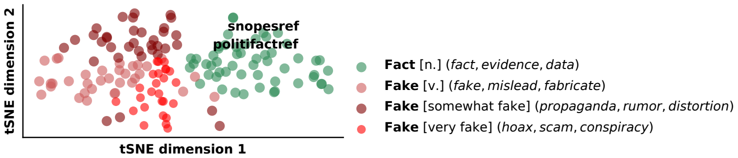


Fig. 13. **Semantics of politifactref and snopesref.** The learned embedding, which encodes the semantics of *politifactref* or *snopesref*, is plotted along with other words in *fact* and three *fake* clusters. Dimensions are reduced from 100 to 2 using t-SNE. References to PolitiFact and Snopes carry similar semantics as other words expressing *fact* in the right part of the figure, as oppose to words expressing *fake* in the left part of the figure.

EmoLex, LIWC and ComLex to compare the efficacy of the three lexicons. We use 10-fold cross validation and report the mean of metrics in Table 2.

We report three different metrics for the predicted ranking label (i.e., veracity from -2 to 2). Spearman's ρ measures the ranking correlation between the predicted veracity and the ground truth with a range of -1 to 1, with a random guess of 0 and a perfect prediction of 1. Label Ranking Average Precision (LRAP) reflects the ability to give more accurate veracity for each post, with a perfect prediction of 1. Finally, Label Ranking Loss (LRL) measures the incorrectly ordered pairs of predicted veracity and ground truth, with a perfect score of 0. These metrics are used for multi-class ranking prediction in prior work [95, 116].

As shown in Table 2, using any lexicon and any model yields better results than a random guess, which means emotional and topical signals can help to determine the veracity of posts. We also find that ComLex results in better performance than EmoLex and LIWC under any model, which aligns with previous findings that domain-learned lexicons perform better than lexicons designed for general writing [56]. When looking at the results of ComLex, we observe that regressions with L1 or L2 regulators perform better than models without regulation, and non-linear models perform better than linear models.

6 LINGUISTIC SIGNALS UNDER FACT-CHECKING

In this section, we focus on linguistic signals in the presence of fact-checking and analyze how they vary in users' comments. To motivate this analysis, we first examine the prevalence and semantics

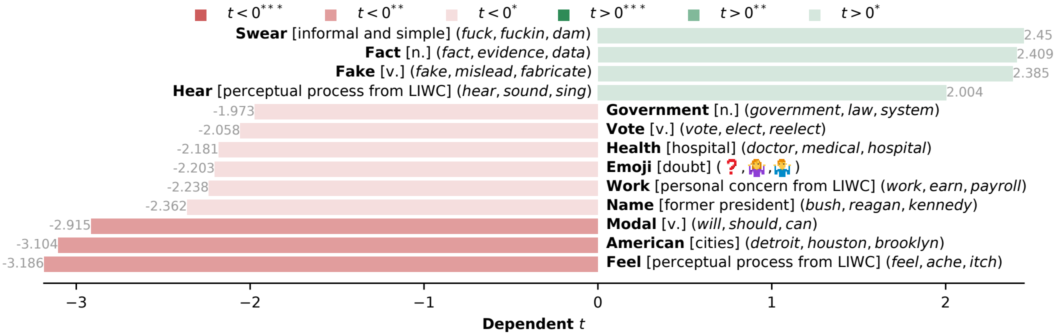


Fig. 14. **Linguistic signals before and after fact-checking.** Clusters with significance dependent t are plotted, ranked by the sign and strength of difference. A positive t indicates that the mean of the statistic is higher after fact-checking than before, and vice versa.

of references to fact-check articles. Note that we replaced any reference to PolitiFact and Snopes in user comments with special tokens *politifactref* and *snopesref*, respectively. We use these tokens for our analysis.

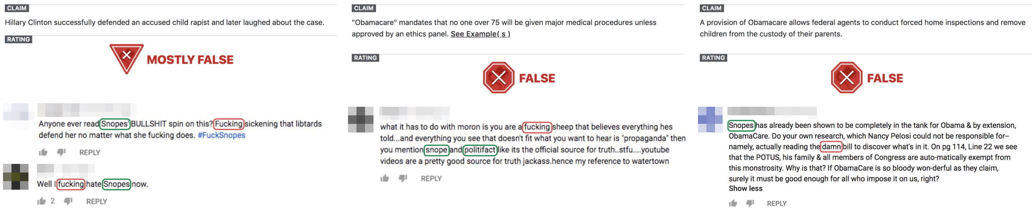
Figure 12 shows the prevalence of *politifactref* and *snopesref*. For all posts that were fact-checked by PolitiFact, 6.68% of them have at least one comment that mentioned PolitiFact. The number for Snopes is similar at 6.18%. This gives us an overview of the prevalence of direct references to PolitiFact and Snopes in the user comments.

Figure 13 shows the semantics of *politifactref* and *snopesref*. As before, we use a 100-dimensional vector to represent the semantics of each word. To visualize the proximity of word semantics, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) [64] to reduce the dimensionality of each vector to 2-dimensional space. As shown in the figure, references to Snopes and PolitiFact have very similar semantics to words in the *fact* cluster (e.g., *fact*, *evidence*, *data*, *non-partisan*, etc.) as oppose to the words in three misinformation clusters (e.g., *fake*, *propaganda*, *hoax*, etc.). Also note that we observed references to other factual sources such as Wikipedia, Pew, Factcheck.org, etc. in the *fact* cluster, which suggests that within the context of user comments on social media, fact-checking websites and general purpose non-partisan websites are afforded a similar degree of trust by users.

6.1 Linguistic Signals before and after Fact-Checking

We now analyze how linguistic signals in users' comments vary before and after fact-checking. To do this, we split user comments into two groups (those written before a fact-check article was available for the given post, and those written after) and use them to perform dependent t tests. Figure 14 highlights the significant ($p < 0.05$) differences in emotions and topics before and after fact-checking for ComLex clusters.

The first evidence that supports **RQ4** is that **the usage likelihoods of several word clusters that express misinformation-awareness increase after a fact-check article is available**. The evidence for this claim includes an increase in factual references (*fact*, *evidence*, *data*, etc., $t = 2.409^{*}$) and verbs expressing deceit (*fake*, *mislead*, *fabricate*, etc., $t = 2.385^{*}$). Comments such as "check *snopesref* for the fact" and "according to *snopesref*, this is fake news" appear more frequently after the publication of fact-check articles. These two clusters have a mean of 0.0028 before fact-checking and 0.0042 after, which represents a 50% difference. This result suggests that social media users are aware of fact-checks, once they become available, and that this increases the likelihood of



CLAIM

Hillary Clinton successfully defended an accused child rapist and later laughed about the case.

RATING

MOSTLY FALSE

Anyone ever read Snopes' BULLSHIT spin on this? Fucking sickening that libtards defend her no matter what she fucking does. #FuckSnopes

REPLY

Well I fucking hate Snopes now.

2 REPLY

CLAIM

"Obamacare" mandates that no one over 75 will be given major medical procedures unless approved by an ethics panel. See Example(s)

RATING

FALSE

what it has to do with moron is you are a fucking sheep that believes everything hes told...and everything you see that doesn't fit what you want to hear is "propaganda" then you mention snope and politifact like its the official source for truth..stfu....youtube videos are a pretty good source for truth jackass.hence my reference to watertown

REPLY

CLAIM

A provision of Obamacare allows federal agents to conduct forced home inspections and remove children from the custody of their parents.

RATING

FALSE

Snopes has already been shown to be completely in the tank for Obama & by extension, ObamaCare. Do your own research, which Nancy Pelosi could not be responsible for--namely, actually reading the damn bill to discover what's in it. On pg 114, Line 22 we see that the POTUS, his family & all members of Congress are auto-matically exempt from this monstrosity. Why is that? If ObamaCare is so bloody won-derful as they claim, surely it must be good enough for all who impose it on us, right?
Show less

REPLY

Fig. 15. **Examples of backfire.** Three examples are given that include the post veracity from a fact-check articles (top) and selected user comments indicating backfire effects (bottom). Words in green blocks (i.e., Snopes, PolitiFact) are identified as reference to fact-checking website, while words in red blocks (i.e., fuck, damn) are mapped in the *swear* word cluster.

rational statements. This observation holds for the subset of posts with comments that explicitly link to PolitiFact or Snopes, e.g., the green boxes in Figure 12 (*fake*, $t = 2.224^*$; *fact*, $t = 2.441^*$).

The second evidence the supports **RQ4** is that **the usage likelihoods of word clusters expressing doubt decrease after a fact-check article is available**. Users' certainty increases after fact-checking, and this is reflected in the decreasing probability of using doubtful emojis (❓, 🤷‍♀️, 🤷‍♂️, etc., $t = -2.203^*$), which have a mean of 0.0005 before fact-checking and 0.00025 after, which represents a 100% difference. Questions such as "is that true❓" and "is this a joke? 🤷‍♀️" appear more frequently before the publication of corresponding fact-check articles.

The supporting evidences for backfire effects in **RQ5** is **the increase in swear words after a fact-check article is published**. We observe more swear word usage after fact-checking (*fuck*, *fuckin*, *dam*, etc., $t = 2.450^*$). In terms of effect size, the mean probability of this cluster is 0.0011 before and 0.0015 after fact-checking, which represents a 36.4% difference. However, we note that only one of five swear word clusters had significant differences before and after fact-checking, which suggests that the backfire effect in comments may be limited. Furthermore, we caution that the use of swear words is, at best, an indirect indicator of backfire: it suggests an increase in negative emotion from some users, and previous lab experiments have shown that this is symptomatic of a stubborn individual clinging to their original false beliefs [124].

Figure 15 shows three examples of backfire. Each presents the post veracity from a fact-check article and selected user comments exemplifying backfire effects. In all three examples, users referred to fact-checking websites and used swear words to expressed their dissatisfaction. These backfire comments also tend to express doubt about the fact-checker themselves because the users perceive them to be biased and unreliable sources [78, 100, 105]. Note that these examples criticized Snopes or PolitiFact in whole rather than referring to individual fact-check articles.

7 CONCLUSION

In this paper, we analyzed how misinformation and fact-checking from PolitiFact and Snopes affect user comments on three major social media platforms (Facebook, Twitter, and YouTube). Our dataset includes 5,303 fact-checked social media posts with 2,614,374 associated user comments. Overall, we found supporting evidence for five research questions:

- **RQ1) Are there any misinformation-awareness signals in user comments indicating their ability to identify misinformation?** We observe that people are more likely to use words showing awareness of misinformation, and are less likely to express trust, as veracity decreases, as shown in Figure 10.
- **RQ2) Do emotional and topical signals in user comments vary with post veracity?** We observe significant emotional and topical changes as veracity decreases, including more

extensive use of emojis and swear words, and less discussion of concrete topics and less objectivity, as shown in Figure 10.

- **RQ3) Can linguistic signals in users comments help to detect misinformation?** We build predictive models to label the veracity of social media posts. Our models easily outperformed a random guessing baseline, however the correlation between predicted value and ground-truth is limited, as shown in Table 2.
- **RQ4) After a social media post is fact-checked, are there any linguistic signals in comments indicating positive effects?** We observe significant increases in misinformation-awareness and factual references after fact-checking articles are available, accompanied by decreases in people's doubtful emotions, as shown in Figure 14.
- **RQ5) After a post is fact-checked, are there any linguistic signals in comments indicating backfire effects?** We observe more swear word usage after fact-checking articles are available, as shown in Figure 14. Examples in Figure 15 show this is linked to backfire effects.

These observations demonstrate the positive impacts of fact-checkers on social media commentary, as well as some of the pitfalls. Furthermore, we hope these observations can help social media users understand how they are affected by misinformation. Finally, our findings may help social computing system designers design more effective interventions and moderation strategies against misinformation.

Limitations and Future Work. There are several limitations of our work. *First*, we only focus on fact-check articles on social media posts with a comments sections. This ignores other fact-check articles that originate from news media or forums, but are then later shared and discussed on social media. Future work should investigate if and how these different dissemination channels alter our findings.

Second, our predictive models only demonstrated weak predictability, thus implying that the content of users' comments are not very strong signals for detecting misinformation. It is noteworthy that previous models that try to predict news veracity based on datasets that are *manually* labeled also cannot achieve high accuracy [94, 120]. In contrast, studies that use URL list-based labeling can achieve high accuracy [98, 118], yet such studies face the dilemma of explaining whether the models really learned to discern the veracity of *individual posts*, or just learned how to distinguish misinformation *sources*. Thus, we believe there is still a long way to go towards the development of truly accurate and generalizable algorithmic misinformation detectors.

Third, we only conduct a coarse temporal analysis (before and after the availability of a fact-check article). Specific events (e.g., the 2016 presidential election) may affect some of our results. Future work should delve into this further using focused case studies or time series analysis.

Fourth, our analysis of misinformation focuses on the veracity dimension of misinformation, but ignores the intention dimension. Without examining intent, we cannot compare and contrast user comments across hoaxes, propaganda, rumor, satire, etc. This is potentially a fruitful angle for future work.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF grants IIS-1408345 and IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.

[2] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 579–586.

[3] Michelle A Amazeen. 2015. Revisiting the epistemology of fact-checking. *Critical Review* 27, 1 (2015), 1–22.

[4] Michelle A Amazeen. 2016. Checking the fact-checkers in 2008: Predicting political ad scrutiny and assessing consistency. *Journal of Political Marketing* 15, 4 (2016), 433–464.

[5] Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 155–168.

[6] Solomon E Asch and H Guetzkow. 1951. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men* (1951), 222–236.

[7] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2010)*, Vol. 10. 2200–2204.

[8] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web.. In *IJCAI*, Vol. 7. 2670–2676.

[9] Lee Becker, George Erhart, David Skiba, and Valentine Matula. 2013. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Vol. 2. 333–340.

[10] Nina Berman. 2017. The victims of fake news. (2017). https://www.cjr.org/special_report/fake-news-pizzagate-seth-rich-newtown-sandy-hook.php

[11] Edward L Bernays. 1928. *Propaganda.* Ig publishing.

[12] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW, 24.

[13] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings.* Technical Report. Citeseer.

[14] Sorrel Brown. 2010. Likert scale examples for surveys. (2010). https://www.extension.iastate.edu/documents/anr/likertscaleexamplesforsurveys.pdf

[15] Craig J. Calhoun. 1994. *Social Theory and the Politics of Identity.*

[16] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 31.

[17] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1217–1230.

[18] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.

[19] Josh Constine. 2017. Facebook tries fighting fake news with publisher info button on links. (10 2017). https://techcrunch.com/2017/10/05/facebook-article-information-button/

[20] Nicole A Cooke. 2017. Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The Library Quarterly* 87, 3 (2017), 211–221.

[21] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2015. Echo chambers in the age of misinformation. *arXiv preprint arXiv:1509.00189* (2015).

[22] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.

[23] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM: Human-Computer Interaction* 1, CSCW (11 2017).

[24] Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: a publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. 417–422.

[25] Robert Farley. 2011. Trump said Obama's grandmother caught on tape saying she witnessed his birth in Kenya. (7 2011). http://www.politifact.com/truth-o-meter/statements/2011/apr/07/donald-trump/

donald-trump-says-president-obamas-grandmother-cau/

[26] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4647–4657.

[27] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to pretrain any-domain models for detecting emotion, sentiment and sarcasm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*. Copenhagen, Denmark.

[28] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 171–175.

[29] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[30] Seth Fiegerman. 2017. Facebook, Google, Twitter to fight fake news with 'trust indicators'. (2017). http://money.cnn.com/2017/11/16/technology/tech-trust-indicators/index.html

[31] Kim Fridkin, Patrick J Kenney, and Amanda Wintersieck. 2015. Liar, liar, pants on fire: How fact-checking influences citizens? reactions to negative advertising. *Political Communication* 32, 1 (2015), 127–151.

[32] Uri Friedman. 2017. The real-world consequences of "fake news". (12 2017). https://www.theatlantic.com/international/archive/2017/12/trump-world-fake-news/548888/

[33] R Kelly Garrett, Erik C Nisbet, and Emily K Lynch. 2013. Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication* 63, 4 (2013), 617–637.

[34] Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. 2015. Media bias in the marketplace: Theory. In *Handbook of media economics*. Vol. 1. Elsevier, 623–645.

[35] Eric Gilbert, Cliff Lampe, Alex Leavitt, Katherine Lo, and Lana Yarosh. 2017. Conceptualizing, Creating, & Controlling Constructive and Controversial Comments: A CSCW Research-athon. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 425–430.

[36] Google. 2018. Google fact checks feature. (2018). https://developers.google.com/search/docs/data-types/factcheck

[37] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council* (2018).

[38] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 153–164.

[39] Kathryn Haglin. 2017. The limitations of the backfire effect. *Research & Politics* 4, 3 (2017), 2053168017716547.

[40] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations.. In *ICWSM*.

[41] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. *world* (2015).

[42] Y Linlin Huang, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 969–980.

[43] Brooks Jackson. 2018. FactCheck. (2018). https://www.factcheck.org

[44] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 12.

[45] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 230–239.

[46] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News Verification by Exploiting Conflicting Social Viewpoints in Microblogs.. In *AAAI*. 2972–2978.

[47] Dan Jurafsky and James H Martin. 2014. *Speech and language processing*. Vol. 3. Pearson London:.

[48] Linda K Kaye, Stephanie A Malone, and Helen J Wall. 2017. Emojis: Insights, affordances, and possibilities for psychological science. *Trends in cognitive sciences* 21, 2 (2017), 66–68.

[49] Michael W. Kearney. 2017. Trusting News Project Report. *Reynolds Journalism Institute* (7 2017).

[50] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 1188–1199.

[51] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 591–602.

[52] Barbara Laslett. 1990. Unfeeling knowledge: Emotion and objectivity in the history of sociology. In *Sociological Forum*, Vol. 5. Springer, 413–433.

[53] David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. *Harvard Kennedy School, Shorenstein Center on Media, Politics and Public Policy* 2 (2017).

[54] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[55] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131.

[56] Minglei Li, Qin Lu, and Yunfei Long. 2017. Are Manually Prepared Affective Lexicons Really Useful for Sentiment Analysis. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2. 146–150.

[57] Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now?: mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 184–196.

[58] Q Vera Liao and Wai-Tat Fu. 2014. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2745–2754.

[59] Kar Wai Lim and Wray Buntine. 2014. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 1319–1328.

[60] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Vol. 1. 63–70.

[61] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 770–780.

[62] Cristian Lumezanu, Nick Feamster, and Hans Klein. 2012. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[63] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics, 142–150.

[64] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008), 2579–2605.

[65] Amr Magdy and Nayer Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, 103–110.

[66] George E. Marcus. 2002. The Sentimental Citizen: Emotion in Democratic Politics. *Perspectives on Politics* (2002).

[67] Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2017. Political Fact-Checking on Twitter: When Do Corrections Have an Effect? *Political Communication* (2017), 1–24.

[68] David Mikkelson. 2011. Barack Obama Birth Certificate: Is Barack Obama's birth certificate a forgery? (8 2011). https://www.snopes.com/fact-check/birth-certificate/

[69] David Mikkelson. 2018. Snopes. (2018). https://www.snopes.com

[70] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[71] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. 2014. Political polarization & media habits. *Pew Research Center* 21 (2014).

[72] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations.. In *ICWSM*. 258–267.

[73] Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 126–145.

[74] Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence* 31, 2 (2015), 301–326.

[75] Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, 26–34.

[76] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[77] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 85–94.

[78] NewsBusters. 2018. Don't Believe the Liberal "Fact-Checkers"! (2018). https://www.newsbusters.org/fact-checkers

[79] Andrew Y Ng, Michael I Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*. 849–856.

[80] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175.

[81] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[82] Brendan Nyhan and Jason Reifler. 2015. Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine* 33, 3 (2015), 459–464.

[83] Brendan Nyhan, Jason Reifler, and Peter A Ubel. 2013. The hazards of correcting myths about health care reform. *Medical care* 51, 2 (2013), 127–132.

[84] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. Association for Computational Linguistics, 309–319.

[85] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1–2 (2008), 1–135.

[86] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol. 10. Association for Computational Linguistics, 79–86.

[87] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*. ACM, 369–374.

[88] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001).

[89] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[90] Gordon Pennycook and David G Rand. 2017. Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy. (2017).

[91] Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approches to emotion* (1984), 197–219.

[92] Ben Popken. 2018. Twitter deleted 200,000 Russian troll tweets. (2 2018). https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

[93] Ethan Porter, Thomas J Wood, and David Kirby. 2018. Sex Trafficking, Russian Infiltration, Birth Certificates, and Pedophilia: A Survey Experiment Correcting Fake News. *Journal of Experimental Political Science* (2018), 1–6.

[94] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A Stylometric Inquiry into Hyperpartisan and Fake News. *arXiv preprint arXiv:1702.05638* (2017).

[95] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 729–740.

[96] 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*. Association for Computational Linguistics, 1589–1599.

[97] Lin Qiu, Han Lin, Jonathan Ramsay, and Fang Yang. 2012. You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality* 46, 6 (2012), 710–718.

[98] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2931–2937.

[99] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50.

[100] Valerie Richardson. 2018. Conservative project seeks to fact-check the fact-checkers accused of liberal bias. (3 2018). https://www.washingtontimes.com/news/2018/mar/27/conservative-project-seeks-fact-check-fact-checker/

[101] Monica A Riordan. 2017. Emojis as tools for emotion work: Communicating affect in text messages. *Journal of Language and Social Psychology* 36, 5 (2017), 549–567.

[102] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM: Human-Computer Interaction* 2, CSCW (11 2018).

[103] Robert J Robinson, Dacher Keltner, Andrew Ward, and Lee Ross. 1995. Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology* 68, 3 (1995), 404.

[104] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592* (2017).

[105] Matt Shapiro. 2016. Running The Data On PolitiFact Shows Bias Against Conservatives. (12 2016). http://thefederalist.com/2016/12/16/running-data-politifact-shows-bias-conservatives/

[106] Aaron Sharockman. 2018. Politifact. (2018). http://www.politifact.com

[107] Sonam Sheth. 2018. Facebook takes down over 200 accounts and pages run by the IRA, a notorious Russian troll farm. (4 2018). http://www.businessinsider.com/facebook-removes-accounts-pages-tied-to-russia-internet-research-agency-2018-4

[108] Baoxu Shi and Tim Weninger. 2016. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 101–102.

[109] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[110] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis. (1966).

[111] Ray Suarez and Kerry Flynn. 2017. Facebook, Twitter issue policy changes to manage fake news and hate speech. (2017). https://www.npr.org/2017/12/24/573333371/facebook-twitter-issue-policy-changes-to-manage-fake-news-and-hate-speech

[112] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).

[113] Henri Tafjel and John C Turner. 1986. The social identity theory of intergroup behavior. *Psychology of intergroup relations* (1986), 7–24.

[114] 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.

[115] Sebastian Tschiatschek, Adish Singla, Manuel Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake News Detection in Social Networks via Crowd Signals. In *The 2018 Web Conference Companion (WWW 2018 Companion)*. Lyon, France.

[116] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2009. Mining multi-label data. In *Data mining and knowledge discovery handbook*. Springer, 667–685.

[117] Svitlana Volkova and Jin Yea Jang. 2018. Misleading or Falsification? Inferring Deceptive Strategies and Types in Online News and Social Media. In *The 2018 Web Conference Companion (WWW 2018 Companion)*. Lyon, France.

[118] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 647–653.

[119] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[120] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 422–426.

[121] Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant Document Discovery for Fact-Checking Articles. In *The 2018 Web Conference Companion (WWW 2018 Companion)*. Lyon, France.

[122] Andrew Ward, L Ross, E Reed, E Turiel, and T Brown. 1997. Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge* (1997), 103–135.

[123] Claire Wardle. 2017. Fake news. It's complicated. *First Draft News* (2017).

[124] Brian E Weeks. 2015. Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication* 65, 4 (2015), 699–719.

[125] Thomas Wood and Ethan Porter. 2016. The elusive backfire effect: mass attitudes? steadfast factual adherence. *Political Behavior* (2016), 1–29.

[126] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment* 7, 7 (2014), 589–600.

[127] Amy X Zhang, Michele Igo, Marc Facciotti, and David Karger. 2017. Using Student Annotated Hashtags and Emojis to Collect Nuanced Affective States. In *Proceedings of the Fourth (2017) ACM Conference on Learning@Scale*. ACM, 319–322.

Received April 2018; revised June 2018; accepted September 2018