Linguistic Cues to Deception: Identifying Political Trolls on Social Media

Aseel Addawood,[*] Adam Badawy,[+] Kristina Lerman,[+] Emilio Ferrara[+]
[*]Department of Informatics, University Illinois at Urbana-Champaign
[+]Information Science Institute, University Southern California
aaddaw2@illinois.edu, {badawy, lerman, ferrarae}@isi.edu

Abstract

The ease with which information can be shared on social media has opened it up to abuse and manipulation. One example of a manipulation campaign that has garnered much attention recently was the alleged Russian interference in the 2016 U.S. elections, with Russia accused of, among other things, using trolls and malicious accounts to spread misinformation and politically biased information. To take an in-depth look at this manipulation campaign, we collected a dataset of 13 million election-related posts shared on Twitter in 2016 by over a million distinct users. This dataset includes accounts associated with the identified Russian trolls as well as users sharing posts in the same time period on a variety of topics around the 2016 elections. To study how these trolls attempted to manipulate public opinion, we identified 49 theoretically grounded linguistic markers of deception and measured their use by troll and non-troll accounts. We show that deceptive language cues can help to accurately identify trolls, with average F1 score of 82% and recall 88%.

Introduction

According to Pew Research Center (Gottfried and Shearer, 2016), two-thirds of Americans get their news from social media. However, even as social media has become a vital source of information for many, it has also become a source of misinformation, hoaxes, and fake news. This is because, unlike traditional news outlets, social media platforms provide little in the way of individual accountability or fact-checking. Misinformation, including conspiracy theories, hoaxes, and rumors, propagate on social media just as readily as factual information. For example, a study showed that when the Ebola crisis broke out in 2014, lies, half-truths, and rumors spread as quickly as accurate information on the Twitter social media platform (Jin et al., 2014).

This issue becomes more prominent when the topic of discussion is related to a highly controversial issue, such as politics, since online users are being exposed to more political content written by ordinary people than ever before. Bakshy, Messing, and Adamic (2015) report that 13% of posts by Facebook users who report their political ideology are political news. Moreover, these posts may not be even generated by humans. Troll accounts and social bots for example, have attempted to manipulate the 2016 U.S. presidential elections by injecting false tweets, or "fake news", in support of or against certain candidates (Pennycook and Rand, 2018). This deceptive, made-up content was shared with millions of Americans, both on Twitter and Facebook, before the 2016 election.

In this paper, we study the language used by Russian trolls during Russia's campaign to interfere in the 2016 US presidential election. Trolls are user accounts whose sole purpose is to sow conflict and deception. In the context of the 2016 elections, their intent was to harm the political process and create distrust in the political system. These trolls were allegedly funded by the Russian government to influence conversations about political issues, with the goal of creating discord and hate among different groups (Gerber and Zavisca, 2016). Stanley Renshon notes that deception in U.S. presidential politics has become more pervasive over the past several decades (Borenstein, 2016). To combat the corrosive influence of online political manipulation, it is important to identify speech that is meant to deceive and mislead. However, the topic of automatic detection of deceptive information has not been widely studied until recently. Our paper addresses this gap with an empirical study of deceptive language used by Russian trolls in their attempts to influence U.S. elections. This may lead to better tools to detect misinformation in the Twitter sphere produced by fake accounts.

Contributions of this work

The focus of our ongoing research is to understand the effects of trolls' interference in the U.S. election. To do so, we plan to answer the following questions:

1. How do trolls insert themselves into political discussions on Twitter? What topics do they discuss?
2. What deceptive linguistic cues do trolls rely upon to generate tweets?
3. Can we automatically detect troll accounts using these deceptive linguistic cues?

The goal of these questions is to understand how these agents camouflage themselves among U.S. Twitter users in order to be more appealing to them. We use the markers of deceptive language to measure how deceptive trolls' tweets are compared to legitimate users. Since deception generally

entails messages and information knowingly transmitted to create a false conclusion (Buller et al., 1994), it stands to reason that trolls use deceptive language to mislead others into believing the information they share. In social media, people tend to be truth-biased on assessing messages they receive (Levine, Park, and McCornack, 1999). Because of that, the accuracy of human detection of deception remains little better than chance (Frank and Feeley, 2003). There is compelling evidence from prior deception research that a variety of language features, either spoken or written, can be valid indicators of deceit (Buller and Burgoon, 1996; Burgoon et al., 1996). One example of the psychological side effects of deception is the observation that people manage the discomfort caused by lying by distancing themselves from the deceptive message they created (DePaulo et al., 2003). Psychological distancing was found to manifest itself through a decrease in self-reference (e.g., "I," "me," "myself") and an increase in group reference (e.g., "they," "he"), which are strategies that indicate a lack of commitment toward the deceptive statement (DePaulo et al., 2003; Hancock et al., 2007). These pronouns become effective linguistic markers of deceptive language. Our analysis reveals that Russian troll accounts that discussed the 2016 U.S. election used deceptive language to influence public opinion and spread biased political information on social media. A classifier was built to identify these trolls based on different deceptive language cues were it resulted in a high accuracy (average F1 score of 82% and recall 88%).

In the remainder of the paper, we first synthesize previous work and background information and relate that to our work. After that, we describe the dataset used for analysis. We then discuss deceptive language markers. Finally, we discuss the findings, conclusions, and proposed directions for future research.

Related Work

We identify deception as misleading the audience via a piece of information. Deceptive information includes but is not limited to lies, fake news, and rumors disseminated to change peoples' cognition or beliefs (Rubin, 2017). Social media that focus primarily on content are highly susceptible to deception, since most communication is text-based and done asynchronously.

A growing body of research suggests that we can learn a great deal about people's underlying thoughts, emotions, and motives by counting and categorizing the words they use to communicate, where the communication can be verbal or written. Several studies on deception detection have demonstrated the effectiveness of linguistic cue identification, as the language of truth-tellers is known to differ from that of deceivers—see (Larcker and Zakolyukina, 2012). Prior work has examined deceptive language in several domains, including fake reviews (Ott et al., 2011; Feng, Banerjee, and Choi, 2012), online games (Zhou et al., 2004), online dating profiles (Toma and Hancock, 2012), interview dialogues (Levitan, Maredia, and Hirschberg, 2018), and opinions on controversial topics (Mihalcea and Strapparava, 2009). However, deception detection in social media has not been studied yet since the type of communication is different from interviews and emails. Even though there is no clear consensus on reliable predictors of deceptive language, prior work has identified several deceptive cues that can be identified in text, extracted and constructed conceptually, to represent several categories, such as complexity, specificity, and non-immediacy. Ott et al. (2011) compared approaches to automatically detecting deceptive opinion spam using a crowd-sourced dataset of fake hotel reviews. Other research has collected deceptive data by asking subjects to write or record deceptive and truthful opinions about controversial topics such as the death penalty or abortion, or about a person that they like or dislike (Mihalcea and Strapparava, 2009). Zhou et al. (2004) consider computer-mediated deception in role-playing games designed to be played over instant messaging and e-mail. Literature on linguistic analysis of deception suggests that changes in word quantity, pronouns, emotional terms, and distinction markers may reflect deception (Burgoon et al., 2003; DePaulo et al., 2003). Linguistic features such as n-grams and language complexity have been analyzed as cues to deception (Pérez-Rosas et al., 2017; Yancheva and Rudzicz, 2013). Moreover, expressing emotions, specifically negative ones, has been shown to be linked to deception (Zhou et al., 2004; Burgoon et al., 2003). Syntactic features such as part of speech tags have also been found to be useful for structured data (Ott et al., 2011; Feng, Banerjee, and Choi, 2012). Building on previous research on deception detection using language, new ways to analyze such data have emerged, such as developing software that can automate the detection of linguistic cues. One of the best known software platforms used for text-based deception detection is Linguistic Inquiry and Word Count (LIWC) (Pennebaker and King, 1999), which groups words into psychologically motivated categories. The main idea of LIWC coding is text classification according to truth conditions. LIWC has been extensively employed to study deception detection (Vrij, 2000; Hancock et al., 2007; Mihalcea and Strapparava, 2009). When deception detection is implemented with standard classification algorithms such as decision trees and logistic regression, it achieves an accuracy of 74% (Fuller, Biros, and Wilson, 2009). When using existing psycholinguistic lexicons as LIWC for detecting deceptive opinions, the accuracy of the classifier achieves an average accuracy rate of 70% (Mihalcea and Strapparava, 2009). By comparison, human judges only achieve a 50-63% success rate in identifying deception (Rubin and Conroy, 2011).

Data Collection

Trolls. To collect Twitter data on Russian trolls, we used a list of 2,752 Russian troll accounts compiled and released by the U.S. Congress.[1] After that, we collected all of the trolls' discussions. To collect the tweets, we used *Crimson Hexagon*, a social media analytic platform that provides paid datastream access. This tool allowed us to obtain tweets and retweets produced by trolls and subsequently deleted in

[1] https://www.recode.net/2017/11/2/16598312/russia-twitter-trump-twitter-deactivated-handle-list

2016. We were interested in understanding troll activity during the election year. We collected data starting from 2015.

Trolls were already active in 2015, posting over a million tweets, 44% of them in Russian, with 31% of the posts with an identifiable location coming from Russia. These accounts were actively demonizing Ukrainian President Petro Poroshenko and campaigning against Ukrainian nuclear power plants. Late in the year, the accounts started tweeting about U.S. elections, talking about debates between Republican and Democratic presidential candidates.

In 2016, the 1,148 trolls posted 1,226,185 tweets, of which 27% were written in Russian. Over 90% of the tweets had identifiable locations, with 65% from the U.S., 27% from Russia, and 2% from Belarus. Troll activity increased in the months leading to the elections, with spikes in activity related to external events. Interestingly, the biggest spike of activity was on October 6th. The tweets were mainly pro-Trump, although no specific topics are discernible. The next day, the *Access Hollywood* tape was released, which showed Trump using derogatory and sexist language. Table 1 presents descriptive statistics of the troll accounts.

Non-Trolls. To collect non-troll tweets, we use two strategies. First, we collect such tweets using a list of hashtags and keywords that relate to the 2016 U.S. Presidential election. This list is crafted to contain a roughly equal number of hashtags and keywords associated with each major Presidential candidate: we select 23 terms, including five terms referring to the Republican Party nominee Donald J. Trump (#donaldtrump, #trump2016, #neverhillary, #trumppence16, #trump), four terms for Democratic Party nominee Hillary Clinton (#hillaryclinton, #imwithher, #nevertrump, #hillary), and several terms related to debates. To make sure our query list was comprehensive, we add a few keywords for the two third-party candidates, including the Libertarian Party nominee Gary Johnson (one term), and Green Party nominee Jill Stein (two terms). Our second strategy is to collect tweets from the same users that do not include the same key terms mentioned above and making sure that we exclude any users who have re-tweeted a troll. Users who did not retweet a troll may help with shaping a better understanding of troll behaviours online. Our collection yielded a total of 12,361,285 tweets produced by 1,166,760 unique users. Table 1 shows descriptive statistics of non-troll accounts.

Table 1: Descriptive statistics of the dataset.

Trolls	*Number*
# unique Russian trolls in the data	1,148
# tweets	1,226,155
# retweets by trolls	688,019
# original tweets by trolls	538,136
# trolls who posted original tweets	1,032
Non-Trolls	*Number*
# unique non-trolls	1,166,760
# tweets by non-trolls	12,361,285
# retweets by non-trolls	9,868,403
# original tweets by non-trolls	2,492,882
# of non-trolls who posted original tweets	140,062

Deceptive Language

We conjecture that political trolls use deception to deliberately mislead others about their true intention. Deception has an emotional and cognitive cost to the deceiver, which can often emerge through the language used to deceive. Studies examined physiological responses of the deceiver utilizing behavioural coding with well-trained experts, or applying content-based criteria to written transcripts for deception detection (Zhou et al., 2004). After that, automated linguistic techniques were developed to analyze the linguistic properties of texts to examine the linguistic profiles of deceptive language (Newman et al., 2003; Zhou et al., 2004).

Deceptive (and truthful) language has been studied through different approaches (Shuy, 1997) based on theoretical assumptions of how deception should be reflected in language. Interpersonal Deception Theory (IDT) explains deception in interpersonal contexts (Buller and Burgoon, 1996). While not developed for online text, it provides a theoretical and evidentiary foundation for the cues in our study. Verbal immediacy theory (VI) was proposed to infer people's attitude or affect. The general construct of immediacy refers to verbal and nonverbal cues that create a psychological sense of closeness or distance (Zhou et al., 2004).

Criteria-based content analysis (CBCA) was developed to determine the credibility of child witness' testimonies in trials for sexual offenses and recently applied to assess testimonies by adults (Raskin and Esplin, 1991). It holds that a statement derived from memory of an actual experience is different in content and quality from a statement based on fantasy (Undeutsch, 1989; Steller and Koehnken, 1989). A similar theory, reality monitoring (RM), was designed to study memory. It holds that a truthful memory differs in quality from remembering a made up event (Johnson and Raye, 1981). Previous research has used this framework extensively to distinguish truth from lies (Bond and Lee, 2005).

Twitter messages lack facial expressions, gestures, and conventions of body posture and distance, so text itself is the only source for us to infer personal opinions and attitudes and verify message credibility. Moreover, previous work has identified deception as a characteristic that can be measured through verbal cues (Tsikerdekis and Zeadally, 2014). Lately, automated linguistic techniques in which computer programs are used to analyze the linguistic properties of text have been used to examine the linguistic profiles of deceptive language—see (Newman et al., 2003; Zhou et al., 2004; Bond and Lee, 2005).

Linguistic cue dictionaries are borrowed from different sources. The first is the *Multiple Perspective Question Answering* (MPQA) opinion corpus developed by University of Pittsburgh. This lexicon includes patterns to account for the various ways in which speakers argue. Lexicon entries are in the form of regular expression patterns. The second is Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010). LIWC is a text analysis program that computes features consisting of normalized word counts

for 93 semantic classes. LIWC dimensions have been used in many studies to predict outcomes including personality (Pennebaker and King, 1999), deception (Newman et al., 2003), and health (Pennebaker, Mayne, and Francis, 1997). LIWC produces the percentage of each variable type by dividing the frequency of the observed variable by the total number of words in the sample, with the exception of word count, words per sentence, and question marks, which are reported frequencies. All features computed for the users are normalized by the number of tweets each posted except for LIWC features since they are computed as percentages.

Building on this research, we identified 49 linguistic cues as potential markers of deceptive language. We used specialized lexicons designed to operationalize language-based measures. Below we justify our choice of each measure as a potential deception marker.

Uncertainty

Based on IDT theory, deceivers tend to use less structured and more evasive language. In contrast, truth-tellers tend to be more certain about their statements. Linguistic markers of certainty, such as "always" or "never," are strong indicators of truthfulness (Levitan, Maredia, and Hirschberg, 2018; Rubin, Liddy, and Kando, 2006). Prior research has shown that subjective language can help recognize certainty in textual information (Rubin, Liddy, and Kando, 2006). Deceivers express greater uncertainty by using more modifiers and model verbs in their text than truth tellers (Zhou et al., 2004; Buller and Burgoon, 1996). The increased use of hedges has been linked to more uncertainty (Rubin, Liddy, and Kando, 2006; Levitan, Maredia, and Hirschberg, 2018). Below we list linguistic cues that may increase uncertainty.

- **Modifier** is a word, phrase, or sentence element that limits or qualifies the sense of another word, phrase, or element in the same construction [2]. Inspired by previous research, we use a list of modifier words borrowed from MPQA. We count occurrences of each modal word in each user's list of tweets and follow the same technique for other lexicon-based measures.
- **Modality** is an expression of an individual's "subjective attitude" (Bybee, Perkins, and Pagliuca, 1994) and "psychological stance" (Mitra, Wright, and Gilbert, 2017) towards a proposition or claim. Words as "should" and "sure" denote assertion of a claim, while "possibly" and "may" express speculation. Modality can be identified as an auxiliary verb that is characteristically used with a verb of predication and expresses necessity or possibility. [3] We measure modality expressed in the text by using a list of necessity and possibility words borrowed from MPQA.
- **Subjectivity** is an aspect of language used to express opinions and evaluations (Banfield, 1982; Wiebe, 2000). Since being certain can be identified as being objective, we hypothesized that subjectivity can provide meaningful signals for deception detection and used OpinionFinder's subjectivity lexicon comprising 8,222 words (Wilson, Wiebe, and Hoffmann, 2005).
- **Quotations** serve as a reliable indicator for accuracy, where quoted content is correlated with being uncertain about its content (De Marneffe, Manning, and Potts, 2012). We hypothesize that trolls use more quoted content in their tweets. We compute this measure by counting the number of quotations present in a user's tweets.
- **Questions**. Based on IDT's interactivity principle, deceivers attempt to increase the interactivity of the communication in an effort to increase believability. Thus, in such interactions, deceivers are expected to ask more questions. Previous work has showed that deceivers use more questions during their discussions (Hancock et al., 2007). Hence, we include questions, measured as question marks in each user's tweets, as a potential indicator of deception.
- **Hedges** are words that express lack of commitment to the truth value of a claim, reveal skepticism, caution, or display an open mind about a proposition. Previous research has shown that deceptive speech contains more hedges (Tausczik and Pennebaker, 2010). We included hedges as potential deception markers in tweets. To measure hedges, we used a curated set of hedging cues from Ken (2005); Hyland (1998).

Non-immediacy

Following IV theory, being non-immediate is related to being deceptive. Deceivers tend to acquire more avoidance strategies. For example, "you and I worked" is equivalent to "we worked" in meaning; however, the former is more non-immediate than the latter. Moreover, IDT theory describes non-immediacy as a method of dissociation where deceivers may use language to distance themselves from the content of their messages. Non-immediacy can be measured through lack of self reference, group reference, and generalization.

- **Self reference**, measured through first person singular pronouns (i.e., "I", "me", or "my"), is one of the ways deceivers can express non-immediacy. Theoretical and empirical observations suggest that deceivers attempt to distance themselves from their deception and not take ownership of a statement by using fewer first-person singular pronouns (Newman et al., 2003; Zhou et al., 2004; Hancock et al., 2007; Toma and Hancock, 2012).
- **Group reference** is measured by using third-person pronouns (i.e., "they", "she"). Research suggests that liars are less likely to use third-person pronouns in their deceptive interactions than in truthful ones (Newman et al., 2003). In contrast, (Zhou et al., 2004) showed that deceptive senders used more group reference compared to truthful senders. This is a strategy to distance themselves from the deceptive message they created Ickes, Reidhead, and Patterson (1986). This feature is obtained from LIWC.
- **Generalization** refers to a person (or object) as a class that includes the person (or object). Hypothesizing that a non-immediate and more general narrative can be associated with higher deception, we employed MPQA's list of

[2] https://www.dictionary.com/browse/modifier

[3] www.webster.com

generalization words to incorporate features corresponding to these language markers.

- **Indefinite articles** Another way to be general is the usage of indefinite articles like "a", "the", and "an", which signal an upcoming noun (Tausczik and Pennebaker, 2010). Indefinite articles are more likely to refer to general concepts than definite articles since they suggest concreteness (Danescu-Niculescu-Mizil et al., 2012). To measure indefinite articles, we used LIWC's list of articles.

Specificity

Based on IDT, RM and CBCA theories, being specific in describing an event or a situation has been proven to relate to truthfulness. Previous research has shown that deceivers are less specific in their text (Burgoon et al., 2003). Being specific includes the usage of discourse markers, causation cues, emotional words, and sense terms.

- **Discourse Markers**. Liars may be particularly wary of using discourse markers that delimit what is in their story and what is not (Newman et al., 2003). Exclusion words, conjunctions, and negations are discourse markers that require a deceiver to be more specific and precise when communicating their messages.We hypothesize that trolls use fewer discourse markers compared to non-trolls. We employed LIWC's list of exclusion, negation, and conjunction words to incorporate features corresponding to these language markers.
- **Causation** is another linguistic marker similar to distinction markers, since it adds specificity and detail to a story and increases the possibility of self-contradiction. Causation words include "because", "effect", and "hence". Previous research has showed that deceivers use fewer causation terms when lying (Hancock et al., 2007). We hypothesize that trolls use fewer causation words. We used LIWC and MPQA's list of causation words.
- **Emotions**. One strategy to avoid being specific is to express more emotions. Previous works have found that deceivers tend to use more emotional language compared to truth tellers (Zhou et al., 2004; Burgoon et al., 2003). Fake content uses more positive words (Pérez-Rosas et al., 2017) and deceivers use negative emotion words (Newman et al., 2003). To measure the extent of emotions expressed in tweets, we used LIWC's comprehensive list of positive and negative emotion words.
- **Sense Terms** like "see", "touch", and "listen" are used to add more details and specifics to narrative. Previous research has suggested that providing such sensory details may be more difficult for a person who is fabricating an opinion or a memory (Johnson and Raye, 1998; Vrij, 2000). Other studies have confirmed that deceivers are more likely to use words that pertain to the senses when lying (Hancock et al., 2007). We employ LIWC's list of sense terms.
- **Use of numbers**. Mentions of numbers is commonly used as a marker of specificity (Li and Nenkova, 2015). Since deceivers tend to be less specific, we hypothesize that trolls use fewer numbers in their text.
- **Relativity** is a linguistic marker available in LIWC, which includes words related to motion, space, and time (i.e., "before"). Previous work identified that legitimate content expresses more relativity (Pérez-Rosas et al., 2017).

Information complexity

Another implication of IDT theory is that deceivers share information that is less complete by sharing content with reduced lexical diversity. Based on CBCA and RM theories, deceivers' language describing an imagined event may fail to reflect the rich diversity of an actual event, where higher sentence complexity results in lower perception of deception (Briscoe, Appling, and Hayes, 2014). Moreover, deceivers display less lexical and content diversity (Zhou et al., 2004). Information complexity is measured by average word length, sentence length, words that have more than six letters, and the amount of punctuation. We used the LIWC to produce the count of words per sentence, words with six letters, and the amount of punctuation. We calculated the average length of a user's set of tweets by summing all the tweets and normalizing by the total tweet count.

Information Quantity

Deceivers may be more hesitant and less forthcoming then truth-tellers and express their hesitancy by using fewer words and sentences. Previous research found deceivers' messages in text-based chats were briefer (Burgoon et al., 2003). We hypothesize that trolls use less information than non-trolls, where information quantity is measured by the number of words, verbs, adverbs, nouns, and prepositions. We use LIWC and NLTK to tokenize tweets and calculate these features.

Persuasion

Persuasion involves convincing a target to accept a message. We hypothesize that deceivers attempt to provide persuasive and credible statements to redirect the listener's attention from any false information.

- **URLs**. The sharing of URLs is a persuasive act that can contribute to a sophisticated and persuasive writing style. Previous research showed that persuasive arguments consistently use more links (Tan et al., 2016; Khazaei, Lu, and Mercer, 2017).
- **Function words** have little lexical or ambiguous meaning and express grammatical relationships among other words within a sentence, or specify the attitude or mood of the speaker. The use of function words in communication reveals deep aspects of the communicators such as his/her honesty and sense of self (Pennebaker, 2011). Previous research has shown that persuasive comments include fewer function words (Khazaei, Lu, and Mercer, 2017). We hypothesize that trolls use fewer function words. To calculate this feature, we used LIWC's list of function words.
- **Examples**. We recorded the normalized number of any mentions of the phrases "for example", "for instance", "e.g." and their synonyms in each tweet based on the notion that providing illustrations and further explanations

is another component of persuasive language, as has been shown in previous research (Tan et al., 2016).

- **Present Focus**. Linguistic cues that are used to talk about the present and the future such as "today", "is", and "now" are commonly used in non-persuasive comments (Xiao, 2018). We used LIWC to get a list of present tense words.
- **Reward**. Words such as "take", "prize", and "benefit" that reference rewards, incentives, and positive goals appear regularly in non-persuasive comments (Xiao, 2018). We hypothesize that troll tweets are less reward-focused then non-troll tweets. We used LIWC to identify the list of reward-focused words.
- **Number of Hashtags**. Previous research has shown that hashtags can serve as useful signals of rumors (Castillo, Mendoza, and Poblete, 2011). We include the hashtag count of tweets as a potential persuasive marker.

Morality

Moral foundation theory Haidt and Graham (2007) describes moral differences across cultures. This theory holds that there is a small number of basic moral values, and people differ in how they endorse these values. Moral foundations include care and harm, fairness and cheating, loyalty and betrayal, authority and subversion, and purity and degradation. We hypothesize that deceptive tweets contain fewer moral linguistic cues than non-deceptive tweets. We measure morality using the list of moral foundation words (Haidt and Graham, 2007).

Metadata

Metadata features obtained from Twitter API include the number of followers, the number of followees, total tweet count, user status count, and number of retweets. No previous work linked the predictability of such features with deception. However, we hypothesize that such features could be an indicator of deceptiveness. Previous research has showed that troll accounts usually have fewer followers and more followees (Badawy, Ferrara, and Lerman, 2018).

Results

What Topics Do Trolls Discuss?

To have a better understanding of the trolls and their activity, we studied the top hashtags, words, and mentions used in both troll and non-troll posts. Trolls use generic hashtags, such as #news, #politics and #sports, which allows their content to be more widely viewed. Thus, when a user searches for "#news" he is exposed to troll tweets. Another interesting insight is that trolls choose controversial topics that many Twitter users are discussing, such as the Black Lives Matter movement. This also makes them appear to be Americans who care about U.S. civil movements. While trolls mention Hillary Clinton with the "neverhillary" hashtag, non-trolls utilize the hashtag "imwithher" more frequently. Based on the top words used in both troll and non-troll tweets, we can get a sense of what topics these two user groups are discussing. We see that trolls discuss recent issues in American society, such as school shootings. In contrast, non-trolls discuss the leaked "Access Hollywood tape".

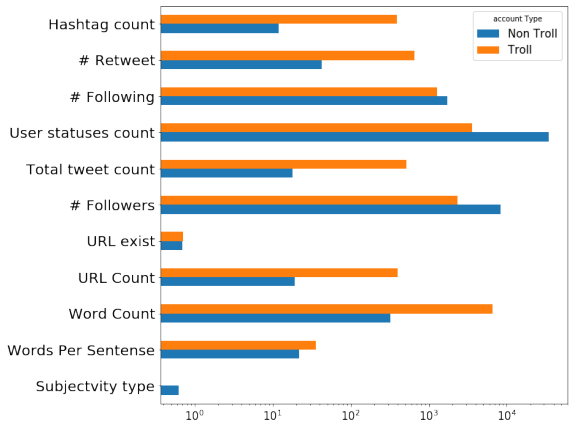


Figure 1: Log mean values for the difference between trolls and non-trolls in descriptive features

Do Trolls Use Deceptive Language?

To study whether trolls use deceptive language, we compare linguistic markers of deception in troll and non-troll tweets. For each linguistic dimension, we conduct a two-tailed t-test over the troll and non-troll datasets to verify the significance of differences for the mean between the two groups. Some linguistic dimensions are positively correlated deception; i.e., if a text contains more of that linguistic dimension, it is more likely to be deceptive. We show in Figures 2 and 3 the log mean values for deception markers.

For metadata features and descriptive features such as hashtag count, URL count, etc., we show the differences between their log mean values in figure 1 below. Figure 1 shows that trolls have significantly fewer followers and more tweets and retweets than non-trolls. This finding echoes findings from prior work (Badawy, Ferrara, and Lerman, 2018). Moreover, trolls use significantly more URLs and hashtags in their tweets, while non-trolls have more tweets and status counts. Figure 2 and figure 3 show the different linguistic measures in troll vs. non-troll tweets for features with positive and negative correlation with deception, respectively. Below we discuss the potential of linguistic measures described in the method section as markers of deception.

- **Uncertainty** Uncertainty was linked to deception. We hypothesized that trolls will use language that introduces uncertainty, such as modifiers, model verbs, etc. However, our results show that trolls use significantly fewer modifiers, model verbs, and hedges than non-trolls, which *contradicted* our hypothesis. On the other hand, other linguistic cues of uncertainty, such as the use of quotations and questions, was *significantly higher* in trolls compared to non-trolls. Moreover, trolls use less subjective language compared to non-trolls. Since subjectivity is used to express opinions and evaluations (Banfield, 1982; Wiebe, 2000), this implies that trolls are less certain, which leads to more deception.

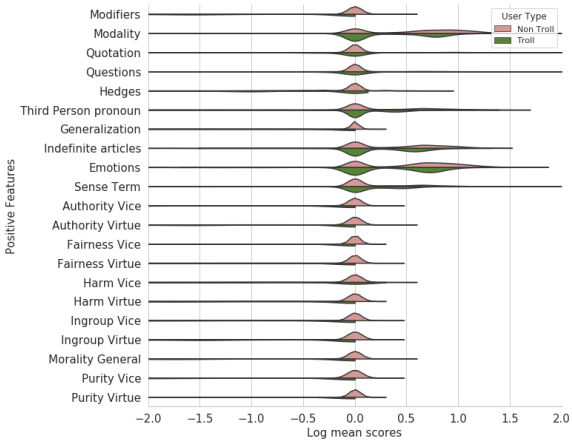


Figure 2: Log mean values for features with positive correlation with deception

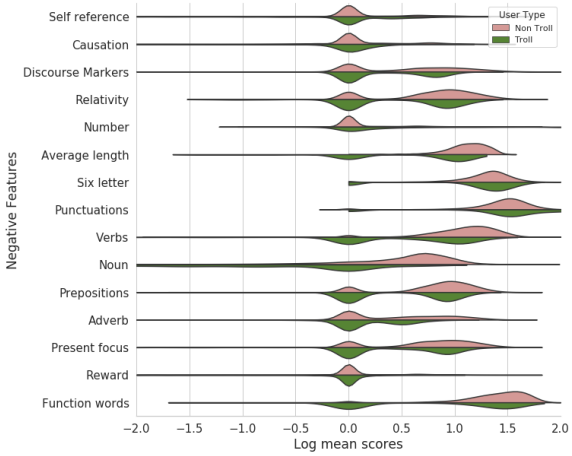


Figure 3: Log mean values for features with negative correlation with L470 deception

- **Non-immediacy** Deceivers tend to use linguistic cues that indicate avoidance, including self reference, group reference, and generalization. Our results show that trolls refer to themselves and others significantly less then non-trolls. This supports previous research that indicates that deceivers use less self and group reference to distance themselves from others (Newman et al., 2003; Zhou et al., 2004; Hancock et al., 2007; Toma and Hancock, 2012). On the contrary, trolls use significantly fewer general terms and indefinite articles compared to non-trolls, which *contradicts* our hypothesis that they use more general narrative to distance themselves from the deception.
- **Specificity** Research suggests that liars may be wary of using discourse markers, which can delimit what is in their story and what is not (Newman et al., 2003). Our results *matched* previous research and show that trolls use significantly fewer discourse markers. Similarly, the usage of causation words add specificity and details to a story and increase the possibility of self-contradiction. We found that trolls tend to use fewer causation words like "because" and fewer sense terms. Moreover, we found that trolls tend to write with significantly less emotion compared to non-trolls. This *contradicts* previous work that found that deceivers tend to express more emotional language (Zhou et al., 2004; Burgoon et al., 2003). Another indicator of specificity is relativity words; we show that trolls tend to use fewer relativity words, *confirming* previous work (Pérez-Rosas et al., 2017).
- **Information Complexity** We find that trolls have less complex, shorter tweets, compared to non-trolls and less complex words (with fewer than six letters). However, they use significantly more words per sentence and more punctuation compared to non-trolls.
- **Information Quantity** We hypothesized that trolls use fewer words and sentences to express their hesitancy. Trolls composed tweets with significantly fewer nouns, verbs, adverbs, and prepositions, which *confirms* research on deception (Burgoon et al., 2003). However, trolls used significantly more words in total compared to non-trolls. Even though trolls have higher word count compared to non-trolls, these words are not important parts of speech, such as nouns and verbs.
- **Persuasion** Trolls used highly persuasive linguistic cues. For example, the use of links in text has been shown to be part of persuasive arguments (Tan et al., 2016; Khazaei, Lu, and Mercer, 2017), and we have found that trolls used significantly more URLs in their tweets compared to non-trolls. Moreover, trolls use fewer function words, which was also confirmed by previous work (Khazaei, Lu, and Mercer, 2017). Furthermore, trolls use significantly fewer present-focused words compared to non-trolls, where the use of present tense has been confirmed to be part of non-persuasive comments (Xiao, 2018). When tweets are less reward-oriented, they are considered more persuasive (Xiao, 2018). In our data, trolls use significantly fewer reward-focused words compared to non-trolls. Trolls used significantly more hashtags, which confirmed our hypothesis that persuasive tweets contain more hashtags than non-persuasive ones. The results *confirm* our hypothesis that trolls use persuasive language as a way to deceive.
- **Morality** We found that trolls show significantly fewer moral values compared to non-trolls. This confirms the hypothesis that using fewer moral cues in the text might imply that the user is trying to be deceptive.

Can Trolls be Identified?

Identifying trolls is a considerable challenge given their small number. The resulting classification task is highly unbalanced, and a trivial algorithm marking every account as non-troll will have high accuracy, but low recall. To test our ability to detect trolls and to see which features are most important in distinguishing between trolls and non-trolls, we leverage two classifiers and multiple models. The first model serves as a baseline, with each model including progressively more variables. We use two off-the-shelf

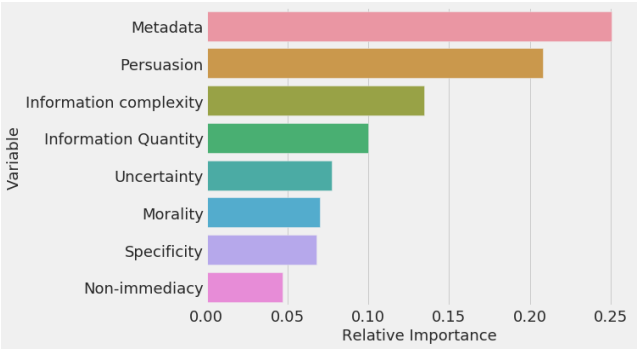


Figure 4: Relative importance of the feature categories using Gradient Boosting for the full model

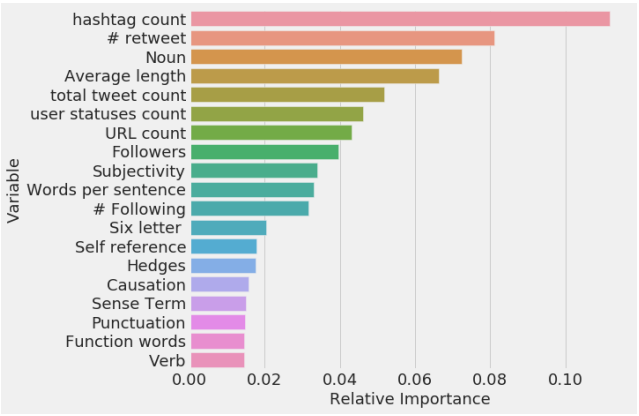


Figure 5: Relative importance of the features using Gradient Boosting for the full model (best performing fold) in predicting users who are trolls

machine learning algorithms: Random Forest (RF) and Gradient Boosting Classifier (GBM) and train classifiers using Stratified 10-fold cross-validation with the following preprocessing steps: *(i)* replace all categorical missing values with the most frequent value in the column, and *(ii)* replace missing values with the mean of the column. We tune the GBM classifier to have a learning rate of 0.1, 500 trees, and max depth of 3 for each tree. To deal with severe imbalance between the majority labels (non-trolls) and minority labels (trolls), we use the *Synthetic Minority Over-sampling Technique + Edited Nearest Neighbor Rule* (SMOTE-ENN) (Batista, Prati, and Monard, 2004) to over-sample from the minority label and undersample from the majority label, to keep a ratio of 1:5 trolls-to-non-trolls in every training fold. We tried different ratios and we did not see much difference in terms of performance between the ratio we chose and the ratios that are closer to 1:1. We decided to pick a ratio that keeps the synthetic data to a minimum while marinating a decent predicative performance.

For GBM, the better performing classifier, we obtain an average F1-score 0.82 and average recall of 0.88 for 10-fold cross validation. For RF, we obtain 0.8 for both the average F1-score and the average recall for the 10-fold.

GBM F1-scores across the 10-folds have a smaller variance than the RF scores. Thus, GBM does not only offer a better average F1-score, but the lower variance between folds shows that it is a more stable model to use.

Feature Importance

To better what features contribute to the accurate identification of trolls, we look at the feature importance plot of Gradient Boosting for the full model. The *Variable Importance by Category* plot (cf., Figure 4) provides a list of the categories of variables in descending order by a mean decrease in the Gini criterion. The top category variables contribute more to the model than the bottom ones and can discriminate better between trolls and non-trolls. In other words, features are ranked based on their predictive power according to the model.

Figure 5 shows the top 20 features in descending order of their importance to contributing to the prediction of trolls. According to the full model and the best GBM fold classifier, the number of hashtags, retweets, and tweets as well as the number of nouns and average length of users' tweets are the most predictive feature of whether users are trolls. Deception markers, including self reference and hedges, round out the top features.

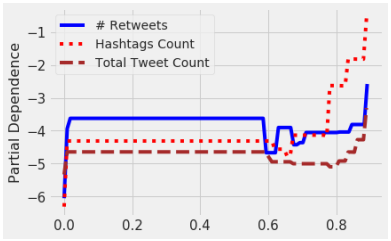
Using *Partial Dependence* plots, we show that the classification outcome has positive relationships with the following features: # of retweets and overall tweet counts, as well as the number of hashtags and urls. Figure 6a visualizes these relationships with the y-axis showing its magnitude and the x-axis the distribution of the feature under examination. Figure 6a suggests moving from left to right that the number of hashtags used increases the probability of being a troll, particularly toward the end of the distribution; higher number of total tweets and retweet counts are also associated with higher likelihood of being a troll, particularly toward the end, while being flat for most of the distribution.

On the other hand, we can see that the outcome has a negative relationship with the number of nouns used, word count, and average tweet length, as shown in Figure 6b. This means that having fewer nouns and posting shorter tweets with fewer words are characteristics associated with higher probability of being a troll.

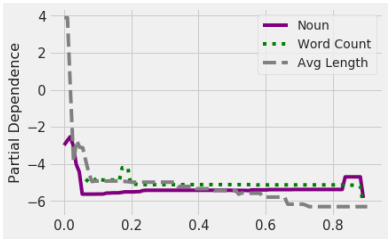
Conclusion

In this paper, we addressed the issue of understanding Russian troll activity on Twitter in 2016. Specifically, we identified linguistic markers of deception that could be good predictors when identifying such trolls. Based on these linguistic markers, we addressed the task of automatic identification of trolls. By developing a theory-driven model working on millions of tweets of Russian trolls and legitimate users, we unfold ways to identify trolls using social media text signals of deception.

Our results showed that Russian troll accounts that discussed the 2016 U.S. election used deceptive language to influence public opinion and spread biased political information on social media. The theory-driven linguistic analysis was able to capture features of the deceptive language. For



(a) Upward Trends



(b) Downward Trends

Figure 6: Partial Dependence plots for some of the top features considered in the full model (best preforming fold). These partial dependence plots are for the Gradient Boosting Classifier fitted to the balanced dataset. Each plot shows the dependence of the outcome variable (troll/non-troll) on the feature under consideration, marginalizing over the values of all other features (Note: x-axis values are CDF-normalized).

example, we found that troll accounts use significantly more persuasive language cues and less complex and specific language. We used these language cues to build a classifier that was able to identify trolls with high accuracy (average F1 score is 82% and recall is 88%). While metadata features were quite distinctive and predictive of trolls, several linguistic features were also predictive of troll accounts, particularly features related to information complexity and persuasion. We show that higher numbers of hashtags, tweets, and retweets are associated with higher likelihood of being a troll, as well as fewer usage of nouns and posting shorter tweets with fewer words. Our work has several limitations. First, not all trolls who are identified as trolls were active in 2016, so having a full picture of all trolls' activity might be harder to achieve. Second, we lack sufficient information on how the troll list was compiled in the first place. This might be an issue, since the methodology taken to identify these trolls could include certain biases that might affect our conclusions. Third, users who are identified as non-trolls might actually be bot accounts not identified in the list. Lastly, our model might be limited by missing potential confounding variables. Despite all of these limitations, the identification of such malicious actors who are mainly responsible of the spread of misinformation is extremely important. Although the data suggest important overall differences in deceptive linguistic patterns across trolls and non-trolls, not all linguistic variables changed as a function of deception. Further investigation into discourse markers that can identify trolls is needed. Moreover, this work can be extended by including higher level interaction terms, such as syntactic constructions and discourse relations.

Acknowledgements

The authors gratefully acknowledge support by the Air Force Office of Scientific Research (award #FA9550-17-1-0327). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFOSR or the U.S. Government.

References

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. *ASONAM* 258–265.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

Banfield, A. 1982. Unspeakable sentences.

Batista, G. E.; Prati, R. C.; and Monard, M. C. 2004. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD explorations newsletter* 6(1):20–29.

Bond, G. D., and Lee, A. Y. 2005. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology* 19(3):313–329.

Borenstein, S. 2016. Getting at the truth behind lying in politics.

Briscoe, E. J.; Appling, D. S.; and Hayes, H. 2014. Cues to deception in social media communications. In *HICSS*, 1435–1443.

Buller, D. B., and Burgoon, J. K. 1996. Interpersonal deception theory. *Communication theory* 6(3):203–242.

Buller, D. B.; Burgoon, J. K.; Daly, J.; and Wiemann, J. 1994. Deception: Strategic and nonstrategic communication. *Strategic interpersonal communication* 191–223.

Burgoon, J. K.; Buller, D. B.; Guerrero, L. K.; Afifi, W. A.; and Feldman, C. M. 1996. Interpersonal deception: Xii. information management dimensions underlying deceptive and truthful messages. *Communications Monographs* 63(1):50–69.

Burgoon, J. K.; Blair, J.; Qin, T.; and Nunamaker, J. F. 2003. Detecting deception through linguistic analysis. In *ISI*, 91–101.

Bybee, J. L.; Perkins, R. D.; and Pagliuca, W. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*, volume 196. University of Chicago Press Chicago.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *WWW*, 675–684. ACM.

Danescu-Niculescu-Mizil, C.; Cheng, J.; Kleinberg, J.; and Lee, L. 2012. You had me at hello: How phrasing affects memorability. In *ACL*, 892–901.

De Marneffe, M.-C.; Manning, C. D.; and Potts, C. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics* 38(2):301–333.

DePaulo, B. M.; Lindsay, J. J.; Malone, B. E.; Muhlenbruck, L.; Charlton, K.; and Cooper, H. 2003. Cues to deception. *Psychological bulletin* 129(1):74.

Feng, S.; Banerjee, R.; and Choi, Y. 2012. Syntactic stylometry for deception detection. In *ACL*, 171–175.

Frank, M. G., and Feeley, T. H. 2003. To catch a liar: Challenges for research in lie detection training. *Journal of Applied Communication Research* 31(1):58–75.

Fuller, C. M.; Biros, D. P.; and Wilson, R. L. 2009. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems* 46(3):695–703.

Gerber, T. P., and Zavisca, J. 2016. Does russian propaganda work? *The Washington Quarterly* 39(2):79–98.

Gottfried, J., and Shearer, E. 2016. *News Use Across Social Medial Platforms 2016*. Pew Research Center.

Haidt, J., and Graham, J. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20(1):98–116.

Hancock, J. T.; Curry, L. E.; Goorha, S.; and Woodworth, M. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes* 45(1).

Hyland, K. 1998. *Hedging in scientific research articles*, volume 54. John Benjamins Publishing.

Ickes, W.; Reidhead, S.; and Patterson, M. 1986. Machiavellianism and self-monitoring: As different as "me" and "you". *Social Cognition* 4(1):58–74.

Jin, F.; Wang, W.; Zhao, L.; Dougherty, E. R.; Cao, Y.; Lu, C.-T.; and Ramakrishnan, N. 2014. Misinformation propagation in the age of twitter. *IEEE Computer* 47(12):90–94.

Johnson, M. K., and Raye, C. L. 1981. Reality monitoring. *Psychological review* 88(1):67.

Johnson, M. K., and Raye, C. L. 1998. False memories and confabulation. *Trends in cognitive sciences* 2(4):137–145.

Ken, H. 2005. Metadiscourse: Exploring interaction in writing. *Continuum, London*.

Khazaei, T.; Lu, X.; and Mercer, R. 2017. Writing to persuade: Analysis and detection of persuasive discourse. *iConference*.

Larcker, D. F., and Zakolyukina, A. A. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50(2):495–540.

Levine, T. R.; Park, H. S.; and McCornack, S. A. 1999. Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communications Monographs* 66(2):125–144.

Levitan, S. I.; Maredia, A.; and Hirschberg, J. 2018. Linguistic cues to deception and perceived deception in interview dialogues. In *NAACL*, 1941–1950.

Li, J. J., and Nenkova, A. 2015. Fast and accurate prediction of sentence specificity. In *AAAI*, 2281–2287.

Mihalcea, R., and Strapparava, C. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *ACL-IJCNLP*, 309–312.

Mitra, T.; Wright, G. P.; and Gilbert, E. 2017. A parsimonious language model of social media credibility across disparate events. In *CSCW*, 126–145. ACM.

Newman, M. L.; Pennebaker, J. W.; Berry, D. S.; and Richards, J. M. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29(5).

Ott, M.; Choi, Y.; Cardie, C.; and Hancock, J. T. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *HLT*, 309–319.

Pennebaker, J. W., and King, L. A. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology* 77(6):1296.

Pennebaker, J. W.; Mayne, T. J.; and Francis, M. E. 1997. Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology* 72(4):863.

Pennebaker, J. W. 2011. The secret life of pronouns. *New Scientist* 211(2828):42–45.

Pennycook, G., and Rand, D. G. 2018. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking.

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2017. Automatic detection of fake news. *COLING* 3391–3341.

Raskin, D. C., and Esplin, P. W. 1991. Statement validity assessment: Interview procedures and content analysis of children's statements of sexual abuse. *Behavioral Assessment*.

Rubin, V. L., and Conroy, N. J. 2011. Challenges in automated deception detection in computer-mediated communication. *Proc American Soc for Inf Science and Technology* 48(1):1–4.

Rubin, V. L.; Liddy, E. D.; and Kando, N. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications*. Springer. 61–76.

Rubin, V. L. 2017. Deception detection and rumor debunking for social media. *The SAGE Handbook of Social Media Research Methods* 342–363.

Shuy, R. W. 1997. *The language of confession, interrogation, and deception*, volume 2. Sage publications.

Steller, M., and Koehnken, G. 1989. Criteria-based statement analysis.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *WWW*, 613–624.

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

Toma, C. L., and Hancock, J. T. 2012. What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication* 62(1):78–97.

Tsikerdekis, M., and Zeadally, S. 2014. Online deception in social media. *Communications of the ACM* 57(9):72–80.

Undeutsch, U. 1989. The development of statement reality analysis. In *Credibility assessment*. Springer. 101–119.

Vrij, A. 2000. Detecting lies and deceit: The psychology of lying and implications for professional practice.

Wiebe, J. 2000. Learning subjective adjectives from corpora. *AAAI/IAAI*.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP*.

Xiao, L. 2018. A message's persuasive features in wikipedia's article for deletion discussions. In *SMSociety*, 345–349.

Yancheva, M., and Rudzicz, F. 2013. Automatic detection of deception in child-produced speech using syntactic complexity features. In *ACL*, volume 1, 944–953.

Zhou, L.; Burgoon, J. K.; Nunamaker, J. F.; and Twitchell, D. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation* 13(1):81–106.