Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking

Hannah Rashkin[†] **Eunsol Choi**[†] **Jin Yea Jang**[‡]
Svitlana Volkova[‡] **Yejin Choi**[†]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
{hrashkin,eunsol,yejin}@cs.washington.edu
[‡]Data Sciences and Analytics, Pacific Northwest National Laboratory
{jinyea.jang,svitlana.volkova}@pnnl.gov

Abstract

We present an analytic study on the language of news media in the context of political fact-checking and fake news detection. We compare the language of real news with that of satire, hoaxes, and propaganda to find linguistic characteristics of untrustworthy text. To probe the feasibility of automatic political fact-checking, we also present a case study based on PolitiFact.com using their factuality judgments on a 6-point scale. Experiments show that while media fact-checking remains to be an open research question, stylistic cues can help determine the truthfulness of text.

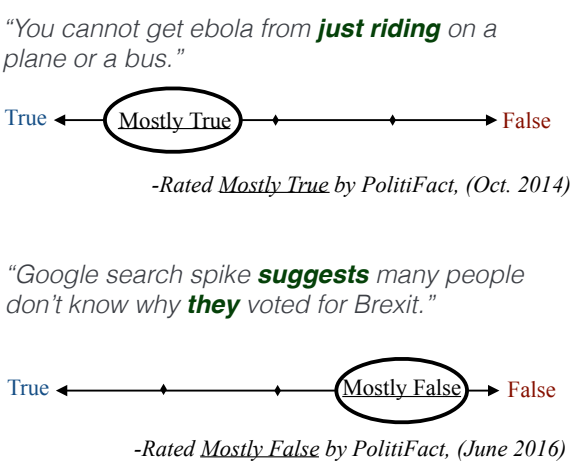

Figure 1: Example statements rated by PolitiFact as *mostly true* and *mostly false*. Misleading phrasing - bolded in green - was one reason for the in-between ratings.

1 Introduction

Words in news media and political discourse have a considerable power in shaping people's beliefs and opinions. As a result, their truthfulness is often compromised to maximize impact. Recently, fake news has captured worldwide interest, and the number of organized efforts dedicated solely to fact-checking has almost tripled since 2014.[1] Organizations, such as PolitiFact.com, actively investigate and rate the veracity of comments made by public figures, journalists, and organizations.

Figure 1 shows example quotes rated for truthfulness by PolitiFact. Per their analysis, one component of the two statements' ratings is the misleading phrasing (bolded in green in the figure). For instance, in the first example, the statement is true as stated, though only because the speaker hedged their meaning with the quantifier *just*. In the second example, two correlated events – Brexit and Google search trends – are presented ambiguously as if they were directly linked.

Importantly, like above examples, most fact-checked statements on PolitiFact are rated as neither entirely true nor entirely false. Analysis indicates that falsehoods often arise from subtle differences in phrasing rather than outright fabrication (Rubin et al., 2015). Compared to most prior work on deception literature that focused on binary categorization of truth and deception, political fact-checking poses a new challenge as it involves a graded notion of truthfulness.

While political fact-checking generally focuses on examining the accuracy of a single quoted statement by a public figure, the reliability of general news stories is also a concern (Connolly et al., 2016; Perrott, 2016). Figure 2 illustrates news types categorized along two dimensions: the *intent* of the authors (desire to deceive) and the *content* of the articles (true, mixed, false).

[1] https://www.poynter.org/2017/there-are-now-114-fact-checking-initiatives-in-47-countries/450477/

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2931–2937
Copenhagen, Denmark, September 7–11, 2017.

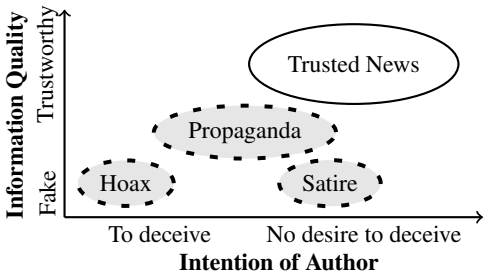

Figure 2: Types of news articles categorized based on their intent and information quality.

In this paper, we present an analytic study characterizing the language of political quotes and news media written with varying intents and degrees of truth. We also investigate graded deception detection, determining the truthfulness on a 6-point scale using the political fact-checking database available at PolitiFact.[2]

2 Fake News Analysis

News Corpus with Varying Reliability To analyze linguistic patterns across different types of articles, we sampled standard trusted news articles from the English Gigaword corpus and crawled articles from seven different unreliable news sites of differing types. Table 1 displays sources identified under each type according to US News & World Report.[3] These news types include:

- *Satire*: mimics real news but still cues the reader that it is not meant to be taken seriously
- *Hoax*: convinces readers of the validity of a paranoia-fueled story
- *Propaganda*: misleads readers so that they believe a particular political/social agenda

Unlike hoaxes and propaganda, satire is intended to be notably different from real news so that audiences will recognize the humorous intent. Hoaxes and satire are more likely to invent stories, while propaganda frequently combines truths, falsehoods, and ambiguities to confound readers.

To characterize differences between news types, we applied various lexical resources to trusted and fake news articles. We draw lexical resources from prior works in communication theory and stylistic analysis in computational linguistics. We tokenize the text with NLTK (Bird et al., 2009) and compute per-document count for each lexicon, and report averages per article of each type.

First among these lexicons is the Linguistic Inquiry and Word Count (LIWC), a lexicon widely used in social science studies (Pennebaker et al., 2015). In addition, we estimate the use of strongly and weakly subjective words with a sentiment lexicon (Wilson et al., 2005). Subjective words can be used to dramatize or sensationalize a news story. We also use lexicons for hedging from (Hyland, 2015) because hedging can indicate vague, obscuring language. Lastly, we introduce intensifying lexicons that we crawled from Wiktionary based on a hypothesis that fake news articles try to enliven stories to attract readers. We compiled five lists from Wiktionary of words that imply a degree a dramatization (comparatives, superlatives, action adverbs, manner adverbs, and modal adverbs) and measured their presence.

News Type	Source	# of Doc	# Tokens per Doc.
Trusted	Gigaword News	13,995	541
Satire	The Onion	14,170	350
	The Borowitz Report	627	250
	Clickhole	188	303
Hoax	American News	6,914	204
	DC Gazette	5,133	582
Propaganda	The Natural News	15,580	857
	Activist Report	17,869	1,169

Table 1: News articles used for analysis in Section 2.

Discussion Table 2 summarizes the ratio of averages between unreliable news and truthful news for a handful of the measured features. Ratios greater than one denote features more prominent in fake news, and ratios less than one denote features more prominent in truthful news. The ratios between unreliable/reliable news reported are statistically significant ($p < 0.01$) with Welsch t-test after Bonferroni correction.

Our results show that first-person and second-person pronouns are used more in less reliable or deceptive news types. This contrasts studies in other domains (Newman et al., 2003), which found fewer self-references in people telling lies about their personal opinions. Unlike that domain, news writers are trying to appear indifferent. Editors at trustworthy sources are possibly more

[2] All resources created for this paper including corpus of news articles from unreliable sources, collection of Politifact ratings, and compiled Wiktionary lexicons have been made publicly available at homes.cs.washington.edu/~hrashkin/factcheck.html

[3] www.usnews.com/news/national-news/articles/2016-11-14/avoid-these-fake-news-sites-at-all-costs

LEXICON MARKERS	RATIO	EXAMPLE SOURCE	EXAMPLE TEXT	MAX
Swear (LIWC)	7.00	Borowitz Report	... Ms. Rand, who has been **damned** to eternal torment ...	S
2nd pers (You)	6.73	DC Gazette	**You** would instinctively justify and rationalize **your** ...	P
Modal Adverb	2.63	American News	... investigation of Hillary Clinton was **inevitably** linked ...	S
Action Adverb	2.18	Activist News	... if one **foolishly** assumes the US State Department ...	S
1st pers singular (I)	2.06	Activist Post	**I** think its against the law of the land to finance riots ...	S
Manner Adverb	1.87	Natural News	... consequences of **deliberately** engineering extinction.	S
Sexual (LIWC)	1.80	The Onion	... added that his daughter better not be **pregnant**.	S
See (LIWC)	1.52	Clickhole	New Yorkers ... can bask in the **beautiful image** ...	H
Negation(LIWC)	1.51	American News	There is **nothing** that outrages liberals more than ...	H
Strong subjective	1.51	Clickhole	He has one of the most **brilliant** minds in basketball.	H
Hedge (Hyland, 2015)	1.19	DC Gazette	As the Communist Party USA website **claims**...	H
Superlatives	1.17	Activist News	Fresh water is the single **most** important natural resource	P
Weak subjective	1.13	American News	... he made that **very clear** in his response to her.	P
Number (LIWC)	0.43	Xinhua News	... **7 million** foreign tourists coming to the country in **2010**	S
Hear (LIWC)	0.50	AFP	The prime minister also **spoke** about the commission ...	S
Money (LIWC)	0.57	NYTimes	He has proposed to lift the state **sales tax** on groceries	P
Assertive	0.84	NYTimes	Hofstra has **guaranteed** scholarships to the current players.	P
Comparitives	0.86	Assoc. Press	... from fossil fuels to **greener** sources of energy	P

Table 2: Linguistic features and their relationship with fake news. The ratio refers to how frequently it appears in fake news articles compared to the trusted ones. We list linguistic phenomena more pronounced in the fake news first, and then those that appear less in the fake news. Examples illustrate sample texts from news articles containing the lexicon words. All reported ratios are statistically significant. The last column (MAX) lists compares which type of fake news most prominently used words from that lexicon (P = propaganda, S = satire, H = hoax)

rigorous about removing language that seems too personal, which is one reason why this result differs from other lie detection domains. This finding instead corroborates previous work in written domains found by Ott et al. (2011) and Rayson et al. (2001), who found that such pronouns were indicative of imaginative writing. Perhaps imaginative storytelling domains is a closer match to detecting unreliable news than lie detection on opinions.

Our results also show that words that can be used to exaggerate – subjectives, superlatives, and modal adverbs – are all used more by fake news. Words used to offer concrete figures – comparatives, money, and numbers – appear more in truthful news. This also builds on previous findings by Ott et al. (2011) on the difference between superlative/comparative usage.

Trusted sources are more likely to use assertive words and less likely to use hedging words, indicating that they are less vague about describing events, as well. This relates to psychology theories (Buller and Burgoon, 1996) that deceivers show more "uncertainty and vagueness" and "indirect forms of expression". Similarly, the trusted sources use the *hear* category words more often, possibly indicating that they are citing primary sources more often.

The last column in Table 2 shows the fake news type that uses the corresponding lexicon most prominently. We found that one distinctive feature of satire compared to other types of untrusted news is its prominent use of adverbs. Hoax stories tend to use fewer superlatives and comparatives. In contrast, compared to other types of fake news, propaganda uses relatively more assertive verbs and superlatives.

News Reliability Prediction We study the feasibility of predicting the reliability of the news article into four categories: trusted, satire, hoax, or propaganda. We split our collected articles into balanced training (20k total articles from the Onion, American News, The Activist, and the Gigaword news excluding 'APW', 'WPB' sources) and test sets (3k articles from the remaining sources). Because articles in the training and test set come from different sources, the models must classify articles without relying on author-specific cues. We also use 20% of the training articles as an in-domain development set. We trained a Max-Entropy classifier with L2 regularization on n-gram tf-idf feature vectors (up to trigrams).[4]

The model achieves F1 scores of 65% on the out-of-domain test set (Table 3). This is a promising result as it is much higher than random, but still leaves room for improvement compared to the

[4]N-gram tfidf vectors have acted as competitive means of cross-domain text-classification. Zhang et al. (2015) found that for data sets smaller than a million examples, this was the best model, outperforming neural models.

Data	Sources	Random	MaxEnt
Dev	in-domain	0.26	0.91
Test	out-of-domain	0.26	0.65

Table 3: F1 scores of 4-way classification of news reliability.

performance on the development set consisting of articles from in-domain sources.

We examined the 50 highest weighted n-gram features in the MaxEnt classifier for each class. The highest weighted n-grams for trusted news were often specific places (e.g., "washington") or times ("on monday"). Many of the highest weighted from satire were vaguely facetious hearsay ("reportedly", "confirmed"). For hoax articles, heavily weighted features included divisive topics ("liberals", "trump") and dramatic cues ("breaking"). Heavily weighted features for propaganda tend towards abstract generalities ("truth", "freedom") as well as specific issues ("vaccines", "syria"). Interestingly, "youtube" and "video" are highly weighted for the propaganda and hoax classes respectively; indicating that they often rely on video clips as sources.

3 Predicting Truthfulness

Politifact Data Related to the issue of identifying the truthfulness of a news article is the fact-checking of individual statements made by public figures. Misleading statements can also have a variety of intents and levels of reliability depending on whom is making the statement.

PolitiFact[5] is a site led by Tampa Bay Times journalists who actively fact-check suspicious statements. One unique quality of PolitiFact is that each quote is evaluated on a 6-point scale of truthfulness ranging from "True" (factual) to "Pants-on-Fire False" (absurdly false). This scale allows for distinction between categories like mostly true (the facts are correct but presented in an incomplete manner) or mostly false (the facts are not correct but are connected to a small kernel of truth).

We collected labelled statements from PolitiFact and its spin-off sites (PunditFact, etc.) (10,483 statements in total). We analyze a subset of 4,366 statements that are direct quotes by the original speaker. The distributions of ratings on the PolitiFact scale for this subset are shown in Table 4. Most statements are labeled as neither completely true nor false.

[5]www.politifact.com/

	More True			More False		
	True	Mostly True	Half True	Mostly False	False	Pants-on-fire
6-class	20%	21%	21%	14%	17%	7%
2-class	62%			38%		

Table 4: PolitiFact label distribution. PolitiFact uses a 6-point scale ranging from: True, Mostly True, Half-true, Mostly False, False, and Pants-on-fire False.

We formulate a fine-grained truthfulness prediction task with Politifact data. We split quotes into training/development/test set of {2575, 712, 1074} statements, respectively, so that all of each speaker's quotes are in a single set. Given a statement, the model returns a rating for how reliable the statement is (Politifact ratings are used as gold labels). We ran the experiment in two settings, one considering all 6 classes and the other considering only 2 (treating the top three truthful ratings as true and the lower three as false).

Model We trained an LSTM model (Hochreiter and Schmidhuber, 1997) that takes the sequence of words as the input and predicts the Politifact rating. We also compared this model with Maximum Entropy (MaxEnt) and Naive Bayes models, frequently used for text categorization.

For input to the MaxEnt and Naive Bayes models, we tried two variants: one with the word tf-idf vectors as input, and one with the LIWC measurements concatenated to the tf-idf vectors. For the LSTM model, we used word sequences as input and also a version where LSTM output is concatenated with LIWC feature vectors before undergoing the activation layer. The LSTM word embeddings are initialized with 100-dim embeddings from GLOVE (Pennington et al., 2014) and fine-tuned during training. The LSTM was implemented with Theano and Keras with 300-dim hidden state and a batch size of 64. Training was done with ADAM to minimize categorical cross-entropy loss over 10 epochs.

Classifier Results Table 5 summarizes the performance on the development set. We report macro averaged F1 score in all tables. The LSTM outperforms the other models when only using text as input; however the other two models improve substantially with adding LIWC features, particu-

	2-CLASS		6-CLASS	
	text	+ LIWC	text	+ LIWC
Majority Baseline	.39	-	.06	-
Naive Bayes	.44	.58	.16	.21
MaxEnt	.55	.58	.20	.21
LSTM	.58	.57	.21	.22

Table 5: Model performance on the Politifact validation set.

MODEL	FEATURE	2-CLASS	6-CLASS
Majority Baseline		.39	.06
Naive Bayes	text + LIWC	.56	.17
MaxEnt	text + LIWC	.55	.22
LSTM	text + LIWC	.52	.19
LSTM	text	.56	.20

Table 6: Model performance on the Politifact test set.

larly in the case of the multinomial naive Bayes model. In contrast, the LIWC features do not improve the neural model much, indicating that some of this lexical information is perhaps redundant to what the model was already learning from text.

We report results on the test set in Table 6. We again find that LIWC features improves MaxEnt and NB models to perform similarly to the LSTM model. As in the dev. set results, the LIWC features do not improve the LSTM's performance, and even seem to hurt the performance slightly.

4 Related Work

Deception Detection Psycholinguistic work in interpersonal deception theory (Buller and Burgoon, 1996) has postulated that certain speech patterns can be signs of a speaker trying to purposefully obscure the truth. Hedge words and other vague qualifiers (Choi et al., 2012; Recasens et al., 2013), for example, may add indirectness to a statement that obscures its meaning.

Linguistic aspects deception detection has been well-studied in a variety of NLP applications (Ott et al., 2011; Mihalcea and Strapparava, 2009; Jindal and Liu, 2008; Girlea et al., 2016; Zhou et al., 2004). In these applications, people purposefully tell lies to receive an extrinsic payoff. In our study, we compare varying types of unreliable news source, created with differing intents and levels of veracity.

Fact-Checking and Fake News There is research in political science exploring how effective fact-checking is at improving people's awareness (Lord et al., 1979; Thorson, 2016; Nyhan and Reifler, 2015). Prior computational works (Vlachos and Riedel, 2014; Ciampaglia et al., 2015) have proposed fact-checking through entailment from knowledge bases. Our work takes a more linguistic approach, performing lexical analysis over varying types of falsehood.

Biyani et al. (2016) examine the unique linguistic styles found in clickbait articles, and Kumar et al. (2016) also characterize hoax documents on Wikipedia. The differentiation between these fake news types is also proposed in previous work (Rubin et al., 2015). Our paper extends this work by offering a quantitative study of linguistic differences found in articles of different types of fake news, and build predictive models for graded deception across multiple domains – PolitiFact and news articles. More recent work (Wang, 2017) has also investigated PolitiFact data though they investigated meta-data features for prediction whereas our investigation is focused on linguistic analysis through stylistic lexicons.

5 Conclusion

We examine truthfulness and its contributing linguistic attributes across multiple domains e.g., online news sources and public statements. We perform multiple prediction tasks on fact-checked statements of varying levels of truth (graded deception) as well as a deeper linguistic comparison of differing types of fake news e.g., propaganda, satire and hoaxes. We have shown that fact-checking is indeed a challenging task but that various lexical features can contribute to our understanding of the differences between more reliable and less reliable digital news sources.

6 Acknowledgements

We would like to thank anonymous reviewers for providing insightful feedback. The research described in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy, the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256082, in part by NSF grants IIS-1408287, IIS-1714566, and gifts by Google and Facebook.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. 2016. "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, pages 94–100.

David B. Buller and Judee K. Burgoon. 1996. Interpersonal deception theory. *Communication Theory* 6(3):203–242. https://doi.org/10.1111/j.1468-2885.1996.tb00127.x.

Eunsol Choi, Chenhao Tan, Lillian Lee, Cristian Danescu-Niculescu-Mizil, and Jennifer Spindel. 2012. Hedge detection as a lens on framing in the gmo debates: A position paper. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*. Association for Computational Linguistics, pages 70–79.

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):e0128193. https://doi.org/10.1371/journal.pone.0128193.

Kate Connolly, Angelique Chrisafis, Poppy McPherson, Stephanie Kirchgaessner, Benjamin Haas, Dominic Phillips, Elle Hunt, and Michael Safi. 2016. Fake news: An insidious trend that's fast becoming a global problem. https://www.theguardian.com/media/2016/dec/02/fake-news-facebook-us-election-around-the-world. Accessed: 2017-01-30.

Codruta Girlea, Roxana Girju, and Eyal Amir. 2016. Psycholinguistic features for deceptive role detection in werewolf. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 417–422.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Ken Hyland. 2015. Metadiscourse. In *The International Encyclopedia of Language and Social Interaction*, John Wiley & Sons, Inc., pages 1–11.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, pages 219–230.

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, pages 591–602.

Charles G Lord, Lee Ross, and Mark R Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37(11):2098–2109.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, pages 309–312.

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

Brendan Nyhan and Jason Reifler. 2015. The effect of fact-checking on elites: A field experiment on US state legislators. *American Journal of Political Science* 59(3):628–640.

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*. Association for Computational Linguistics, pages 309–319.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd Boyd, and Martha E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker-Conglomerates, Austin, TX. www.liwc.net.

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)*. Association for Computational Linguistics, pages 1532–1543.

Kathryn Perrott. 2016. 'Fake news' on social media influenced US election voters, experts say. http://www.abc.net.au/news/2016-11-14/fake-news-would-have-influenced-us-election-experts-say/8024660. Accessed: 2017-01-30.

Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2001. Grammatical word class variation within the british national corpus sampler. *Language and Computers* 36(1):295–306.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1650–1659.

Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology* 52(1):1–4.

Emily Thorson. 2016. Belief echoes: The persistent effects of corrected misinformation. *Political Communication* 33(3):460–480.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics, pages 18–22.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the Association for Computational Linguistics Short Papers*. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 347–354.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*. pages 649–657.

Lina Zhou, Judee K. Burgoon, Jay F. Nunamaker, and Doug Twitchell. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation* 13(1):81–106.