Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang and Nathan Hodas
Data Sciences and Analytics Group, National Security Directorate
Pacific Northwest National Laboratory
902 Battelle Blvd, Richland, WA 99354
firstname.lastname@pnnl.gov

Abstract

Pew research polls report 62 percent of U.S. adults get news on social media (Gottfried and Shearer, 2016). In a December poll, 64 percent of U.S. adults said that "made-up news" has caused a "great deal of confusion" about the facts of current events (Barthel et al., 2016). Fabricated stories in social media, ranging from deliberate propaganda to hoaxes and satire, contributes to this confusion in addition to having serious effects on global stability.

In this work we build predictive models to classify 130 thousand news posts as suspicious or verified, and predict four subtypes of suspicious news – satire, hoaxes, clickbait and propaganda. We show that neural network models trained on tweet content and social network interactions outperform lexical models. Unlike previous work on deception detection, we find that adding syntax and grammar features to our models does not improve performance. Incorporating linguistic features improves classification results, however, social interaction features are most informative for finer-grained separation between four types of suspicious news posts.

1 Introduction

Popular social media platforms such as Twitter and Facebook have proven to be effective channels for disseminating falsified information, unverified claims, and fabricated attention-grabbing stories due to their wide reach and the speed at which this information can be shared. Recently, there has been an increased number of disturbing incidents of fabricated stories proliferated through social media having a serious impact on real-world events (Perrott, 2016; Connolly et al., 2016)

False news stories distributed in social media vary depending on the intent behind falsification. Unlike verified news, suspicious news tends to build narratives rather than report facts. On one extreme is *disinformation* which communicates false facts to deliberately deceive readers or promote a biased agenda. These include posts generated and retweeted from propaganda and so-called clickbait ("eye-catching" headlines) accounts. The intent behind **propaganda** and **clickbait** varies from opinion manipulation and attention redirection to monetization and traffic attraction. **Hoaxes** are another type of *disinformation* that aims to deliberately deceive the reader (Tambuscio et al., 2015; Kumar et al., 2016). On the other extreme is **satire**, e.g., `@TheOnion`, where the writer's primary purpose is not to mislead the reader, but rather entertain or criticize (Conroy et al., 2015). However, satirical news and hoaxes may also be harmful, especially when they are shared out of context (Rubin et al., 2015).

Our novel contributions in this paper are twofold. We first investigate several features and neural network architectures for automatically classifying verified and suspicious news posts, and four sub-types of suspicious news. We find that incorporating linguistic and network features via a "late fusion" technique boosts performance. We then investigate differences between verified and suspicious news tweets by conducting a statistical analysis of linguistic features in both types of account. We show significant differences in use of biased, subjective language and moral foundations behind suspicious and trustworthy news posts.

Our analysis and experiments rely on a large Twitter corpus[1] collected during a two-week pe-

[1]Data available at: http://www.cs.jhu.edu/∼svitlana/

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 647–653
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2102

TYPE	NEWS	POSTS	RTPA	EXAMPLES
Propaganda	99	56,721	572	ActivistPost
Satire	9	3,156	351	ClickHole
Hoax	8	4,549	569	TheDcGazette
Clickbait	18	1,366	76	chroniclesu
Verified	166	65,792	396	USATODAY

Table 1: Twitter dataset statistics: news accounts, posts and retweets per account (RTPA).

riod around terrorist attacks in Brussels in 2016. Our method of collection ensures that our models learn from verified and suspicious news within a predefined timeframe, and further ensures homogeneity of deceptive texts in length and writing manner (Rubin et al., 2015).

Several tools have been recently developed to verify and reestablish trusted sources of information online e.g., Google fact checking (Gindras, 2016) and Facebook repost verification (Mosseri, 2016). These projects, among others, teach news literacy[2] and contribute to fact-checking online.[3] We believe our models and novel findings on linguistic differences between suspicious and verified news will contribute to these fact-checking systems, as well as help readers to judge the accuracy of information they consume in social media.

2 Data

Suspicious News We relied on several public resources that annotate suspicious Twitter accounts or their corresponding websites as propaganda, hoax, clickbait and satire. They include propaganda accounts identified by PropOrNot,[4] satire, clickbait and hoax accounts.[5] In total we collected 174 suspicious news accounts.[6] In addition, we manually confirmed that accounts and their corresponding webpages labeled by PropOrNot have one or more signs of propaganda listed below: (a) tries to persuade; (b) influences the specific emotions, attitudes, opinions, and actions; (c) target audiences for political, ideological, and religious purposes; and (d) contains selectively-omitting and one-sided messages.

Figure 1 presents a communication network between verified and suspicious news accounts. We observe that verified accounts are connected to

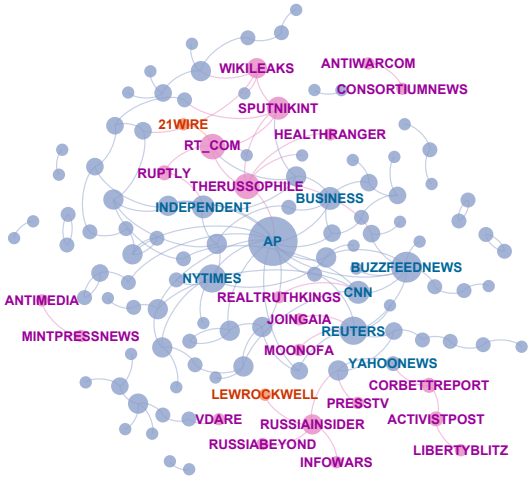


Figure 1: Communication network (@mention) among verified (blue), propaganda (pink), and clickbait (orange) accounts (no shared edges with Hoax and Satire accounts).

(via RTs and mentions) some suspicious news accounts – clickbaits and propaganda.

Verified News We manually constructed a list of 252 "trusted" news accounts that tweet in English and checked whether they are verified on Twitter. We release the final verified list of trusted and suspicious news accounts used in our analysis.[7]

Tweet Corpus We query the Twitter firehose from Mar 15 to Mar 29 2016 – one week before and after Brussels bombing on Mar 22 2016 for 174 suspicious and 252 verified news accounts. We collected retweets generated by any user that mentions one of these accounts and assign the corresponding label propagated from suspicious or trusted news.[8] We de-duplicated, lowercased, and tokenized these posts and applied standard NLP preprocessing. We extracted part-of-speech tags and dependency parses for 130 thousand tweets using SyntaxNet (Petrov, 2016).

3 Models

We propose linguistically-infused neural network models to classify social media posts retweeted from news accounts into verified and suspicious categories – propaganda, hoax, satire and clickbait. Our models incorporate tweet text, social graph, linguistic markers of bias and subjectivity, and moral foundation features. We experiment with several baseline models, and develop neural network architectures presented in Figure 2 in the

[2]News Literacy: http://www.thenewsliteracyproject.org/

[3]Fact checking: http://reporterslab.org/fact-checking/ Hoaxy: http://hoaxy.iuni.iu.edu/

[4]Propaganda: http://www.propornot.com/p/the-list.html

[5]http://www.fakenewswatch.com/

[6]To ensure the quality of suspicious account labels we manually verified them.

[7]The lists of verified and suspicious news: http://www.cs.jhu.edu/~svitlana/TwitterList

[8]Suspicious news annotations should be done on a tweet rather than an account level. However, these annotations are extremely costly and time consuming.

Keras framework.[9] We rely on state-of-the-art layers effectively used in text classification – Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) (Johnson and Zhang, 2014; Zhang and Wallace, 2015). The content sub-network consists of an embedding layer and either (a) one LSTM layer or (b) two 1-dimensional convolution layers followed by a max-pooling layer.

We initialize our embedding layer with pre-trained GloVe embeddings (Pennington et al., 2014). The social graph sub-network is a simple feed-forward network that takes one-hot vectors of user interactions, e.g. @mentions, as input. We are careful to exclude source @mentions from these vectors, as these were used to derive labels for our networks and would likely lead to overfitting. In addition to content and network signals, we incorporate other linguistic cues into our networks. For this we rely on the "late fusion" approach that has been shown to be effective in vision tasks (Karpathy et al., 2014; Park et al., 2016). "Fusion" allows for a network to learn a combined representation of multiple input streams. This fusion can be done early (in the feature extraction layers) or later (in the later extraction layers, or in classification layers). In our case, we use fusion as a technique for training networks to learn how to combine data representations from different modalities (network and text features) to boost performance. We train our models for 10 epochs using the ADAM optimization algorithm, and evaluate them using 10 fold cross-validation (Kingma and Ba, 2014).

Baselines We compare our neural network architectures to several baselines. Word- and document-level embeddings have been shown to be effective as input to simpler classifiers. We experiment with several feature inputs for testing baseline classifiers: (a) TFIDF features, (b) Doc2Vec vectors and (c) Doc2Vec or TFIDF features concatenated with linguistic or network features. In the case of Doc2Vec features, we induce 200-dimensional vectors for each tweet using the gensim library,[10] training for 15 epochs.

Bias cues Inspired by earlier work on identifying *biased language* on Wikipedia (Recasens et al., 2013) we extract *hedges* (expressions of tentativeness and possibility) (Hyland, 2005), *assertive verbs* (the level of certainty in the complement clause) (Hooper, 1975), *factive verbs* (presuppose the truth of their complement clause) (Kiparsky and Kiparsky, 1968), implicative verbs (imply the truth or untruth of their complement) (Karttunen, 1971) and report verbs (Recasens et al., 2013) from preprocessed tweets.

[9]Keras: https://keras.io/

[10]https://radimrehurek.com/gensim/models/doc2vec.html

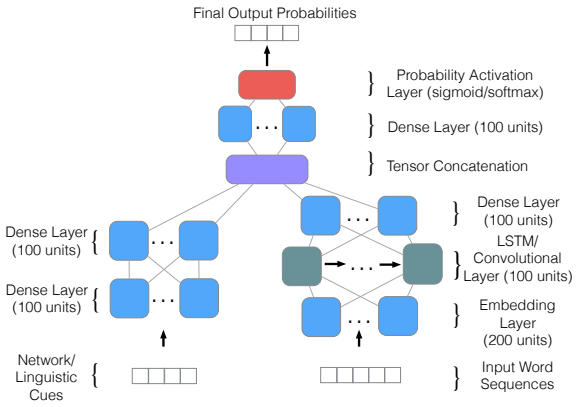


Figure 2: Neural network architecture for news classification fused with social network and linguistic cues.

Subjectivity cues We rely on external publicly available subjectivity, and positive and negative opinion lexicons to extract strongly and weakly subjective words (Riloff and Wiebe, 2003), positive and negative opinion words (Liu et al., 2005).

Psycholinguistic cues In addition to biased and subjective language cues, we extract Linguistic Inquiry Word Count (LIWC) features (Pennebaker et al., 2001) to capture additional signals of persuasive and biased language in tweets. LIWC features have been successfully used for deception detection before (Hancock et al., 2007; Vrij et al., 2007; Mihalcea and Strapparava, 2009). For example, persuasive language cues in LIWC include statistics and factual data, rhetorical questions, imperative commands, personal pronouns, and emotional language. Additional biased language cues captured by LIWC are quotations, markers of certainty, inclusions and conjunctions. Extra subjective language cues in LIWC cover positive and negative emotion and anxiety words.

Moral foundation cues According to Haidt and Graham (2007); Graham et al. (2009), there is a small number of basic widely supported moral values, and people differ in the way 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 suspicious news could appeal to specific moral foundations of their read-

ers in a way that is distinct from verified news accounts. Thus, they could help in predicting verified vs. suspicious news, as well as different suspicious news types.

4 Results

4.1 Classification

Table 2 presents classification results for Task 1 (binary) – suspicious vs. verified news posts and Task 2 (multi-class) – four types of suspicious tweets e.g., propaganda, hoaxes, satire and clickbait. We report performance for different model and feature combinations.

We find that our neural network models (both CNNs and RNNs) significantly outperform logistic regression baselines learned from all feature combinations.[11] The accuracy improvement for the binary task is 0.2 and F1-macro boost for the multi-class task is 0.07. We also observe that all models learned from network and tweet text signals outperform models trained exclusively on tweets. We report 0.05 accuracy improvement for Task 1, and 0.02 F1 boost for Task 2. Adding linguistic cues to basic tweet representations significantly improves results across all models. Finally, by combining basic content with network and linguistic features via late fusion, our neural network models achieve best results in binary experiments. Interestingly, models perform best in the multi-class case when trained on tweet embeddings and fused network features alone. We report 0.95 accuracy when inferring suspicious vs. verified news posts, and 0.7 F1-macro when classifying types of suspicious news.

Syntax and grammar features have been predictive of deception in the product review domain (Feng et al., 2012; Pérez-Rosas and Mihalcea, 2015). However, unlike earlier work we find that fusing these features into our models significantly decreases performance – by 0.02 accuracy for the binary task and 0.02 F1 for multi-class. This may be explained by the domain differences between reviews and tweets which are shorter, more noisy and difficult to parse.

4.2 Linguistic Analysis

We measure statistically significant differences in linguistic markers of bias, subjectivity and moral foundations across different types of suspicious news, and contrast them with verified news using resources described in Section 3. These novel findings presented in Table 3 provide deeper understanding of model performance in Table 2.

	Binary			Multi-class	
Features	**A**	**ROC**	**AP**	**F1**	**F1 macro**
Baseline 1: Logistic Regression (Doc2Vec)					
Tweets	0.65	0.70	0.68	0.82	0.40
+ network	0.72	0.80	0.82	0.88	0.57
+ cues	0.69	0.74	0.73	0.83	0.46
ALL	**0.75**	**0.84**	**0.84**	**0.88**	**0.59**
Baseline 2: Logistic Regression (TFIDF)					
Tweets	0.72	0.81	0.81	0.84	0.48
+ network	0.78	0.87	0.88	0.88	0.59
+ cues	0.75	0.85	0.85	0.86	0.49
ALL	**0.79**	**0.88**	**0.89**	**0.89**	**0.59**
Recurrent Neural Network					
Tweets	0.78	0.87	0.88	0.90	0.63
+ network	0.83	0.91	0.92	**0.92**	**0.71**
+ cues	0.93	0.98	0.99	0.90	0.63
+ syntax	0.93	0.96	0.96	0.90	0.64
ALL	**0.95**	**0.99**	**0.99**	0.91	0.66
Convolutional Neural Network					
Tweets	0.76	0.85	0.87	0.91	0.63
+ network	0.81	0.9	0.91	**0.92**	**0.70**
+ cues	0.93	0.98	0.98	0.90	0.61
ALL	**0.95**	**0.98**	**0.99**	0.91	0.64

Table 2: Classification results: predicting suspicion and verified posts reported as A – accuracy, AP – average precision, ROC – the area under the receiver operator characteristics curve, and inferring types of suspicious news reported using F1 micro and F1 macro scores.

Verified news tweets contain significantly less bias markers, hedges and subjective terms and less harm/care, loyalty/betrayal and authority moral cues compared to suspicious news tweets. Satirical news are the most different from propaganda and hoaxes; and propaganda, hoax and clickbait news are the most similar based on moral, bias and subjectivity cues.

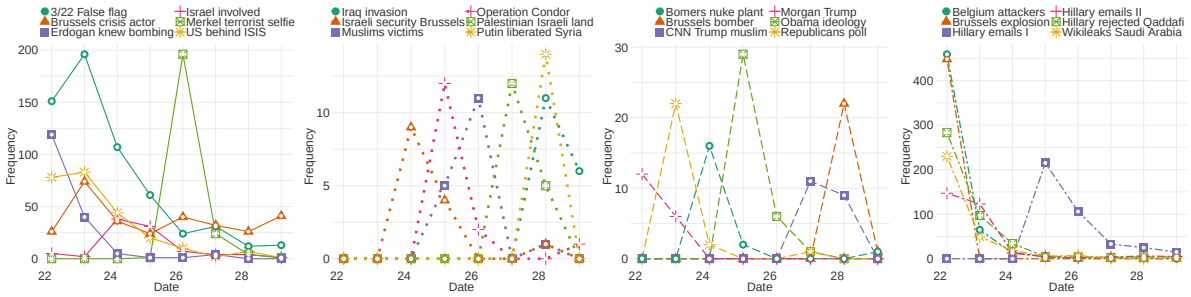
Propaganda news target morals more than satire and hoaxes, but less than clickbait. Satirical news contains more loyalty and less betrayal morals compared to propaganda, hoaxes and clickbait news. Propaganda news target authority more than satire and hoaxes, and fairness more than satire.

Hoaxes and propaganda news contain significantly less bias markers (e.g, hedging, implicative and factive verbs) compared to satire. However, propaganda and clickbait news contain significantly more factive verbs and bias language markers compared to hoaxes. Satirical news use significantly more subjective terms compared to other news, while clickbait news use more subjective cues than propaganda and hoaxes.

[11] We experimented with other baseline models, such as Random Forest, but found negligible difference between these results and those obtained via logistic regression.

CUES	V ↔ F	P ↔ S	P ↔ H	P ↔ C	S ↔ H	S ↔ C	H ↔ C
MORAL FOUNDATION CUES							
Harm	2.1↓↓↓ 2.8	2.8↑ 2.1	–	–	2.0↓↓ 2.6	–	–
Care	5.8↓↓↓ 9.0	9.4↑↑↑ 6.3	9.4↑↑↑ 5.1	9.4↓ 11.3	–	6.3↓↓↓ 11.3	5.1↓↓↓ 11.3
Fairness	–	0.8↑ 0.4	–	–	–	0.4↓ 1.0	–
Cheating	0.3↑ 0.2	–	–	–	–	–	–
Loyalty	2.1↓↓↓ 2.5	2.2↓↓↓ 7.6	–	–	7.6↑↑↑ 2.0	7.6↑↑↑ 2.3	–
Betrayal	1.7↓↓↓ 3.1	3.4↑↑↑ 0.2	3.4↑↑↑ 2.2	–	0.2↓↓↓ 2.2	0.2↓↓↓ 3.0	–
Authority	2.4↓↓↓ 2.9	3.0↑ 2.1	3.0↑ 2.3	–	–	–	–
BIASED LANGUAGE CUES							
Assertive	12.6↓↓↓ 13.8						
Bias	142.6↓↓↓ 164.4	–	165.5↑↑↑ 148.8	–	165↑↑↑ 148.8	–	148.8↓↓167.1
Factive	4.9↓↓↓ 5.5	5.5↓ 6.3	5.5↑ 4.7	5.5↓ 6.8	6.3↑ 4.7	–	4.7↓↓ 6.8
Hedges	14.2↓↓↓ 15.7	15.6↓↓↓ 20.0	–	–	20↑↑↑ 15.8	20↑↑↑ 13.4	–
Implicative	7.6↓↓↓ 8.9	8.6↓↓↓ 15.2	–	–	15.2↑↑↑ 8.8	15.2↑↑↑ 8.3	–
Report	30↓↓↓ 34.5	34.3↓ 36.0	–	–	–	–	–
SUBJECTIVE LANGUAGE CUES							
Subjective	28.8↓↓↓ 32.8	32.6↓↓↓ 39.5	–	–	39.5↑↑↑ 30.9	39.5↑↑↑ 32.5	–
Strong Subj	23.5↓↓↓ 25.3	24.8↓↓↓ 31.5	24.8↓↓↓ 26.3	24.8↓↓↓ 27.5	31.5↑↑↑ 26.3	–	–
Weak Subj	24.8↓↓↓ 30.8	31.2↓↓↓ 32.8	31.2↑↑↑ 24.1	–	32.8↑↑↑ 24.1	32.8↑↑ 30.7	24.1↓↓↓ 30.7

Table 3: Linguistic analysis of moral foundations, bias and subjective language shown as the **percentage of tweets with one or more cues** across verified (V) and suspicious (F) news – propaganda (P), hoaxes (H), satire (S) and clickbait (C). We report only statistically significant differences: p-value $\leq 0.05\uparrow$, $\leq 0.01\uparrow\uparrow$, $\leq 0.001\uparrow\uparrow\uparrow$ estimated using the Mann-Whitney U test. Subjective lexicon is from (Liu et al., 2005), weekly and strongly subjective terms are from (Riloff and Wiebe, 2003).



(a) Rumors (b) Clickbaits (c) Hoaxes (d) Propaganda

Figure 3: The most popular retweets over time across suspicious news types in contrast to rumors.

4.3 Suspicious News Retweet Patterns

In addition to contrasting linguistic realizations behind different types of suspicious news on Twitter, we are interested in qualitatively evaluating differences in retweet patterns across suspicious news types (Lumezanu and Klein, 2012; Mendoza et al., 2010; Kumar et al., 2016). Figure 3 presents top retweeted tweets over time across three types of suspicious news – hoaxes, propaganda, and clickbaits and contrasts them to well-studied retweeting behaviors of rumors. We observe that users retweeting propaganda, clickbaits and hoaxes send high volumes of tweets over short periods of time. Rumors[12] are less spiky but active over significantly longer periods of time compared to other suspicious news. We also notice that rumors and propaganda contain the majority of topics related to Brussels bombing, but clickbaits and hoaxes promote very divergent set of topics.

[12]We identified rumors relevant to Brussels bombing: http://www.cs.jhu.edu/∼svitlana/BrusselsRumorList

5 Summary

We built linguistically-infused neural network models that jointly learn from tweet content and social network interactions to classify suspicious and verified news tweets and infer specific types of suspicious news. Future work may focus on utilizing more sophisticated discourse and pragmatics features, and inferring degrees of credibility. We hope our findings on bias and subjectivity in suspicious news will help readers to better judge about credibility of news in social media.

6 Acknowledgments

The research described in this paper was conducted under the High-Performance Analytics Program and 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.

References

Michael Barthel, Amy Mitchell, and Jesse Holcomb. 2016. Many americans believe fake news is sowing confusion. http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/. Accessed: 2017-01-31.

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.

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52(1):1–4.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of ACL*. pages 171–175.

Richard Gindras. 2016. Labeling fact-check articles in google news. https://blog.google/topics/journalism-news/labeling-fact-check-articles-google-news/. Accessed: 2016-12-12.

Jeffrey Gottfried and Elisa Shearer. 2016. News use across social media platforms 2016. http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016. Accessed: 2017-01-30.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* 96(5):1029.

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

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

Joan B. Hooper. 1975. On assertive predicates. In J. Kimball, editor, *Syntax and Semantics*. volume 4, pages 91–124.

Ken Hyland. 2005. *Metadiscourse*. Wiley Online Library.

Rie Johnson and Tong Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058* .

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of CVPR*. pages 1725–1732.

Lauri Karttunen. 1971. Implicative verbs. *Language* pages 340–358.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Paul Kiparsky and Carol Kiparsky. 1968. *Fact*. Indiana University.

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of WWW*. pages 591–602.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of WWW*. pages 342–351.

Feamster Lumezanu and H Klein. 2012. Measuring the tweeting behavior of propagandists. In *Proceedings of ICWSM*.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: can we trust what we rt? In *Proceedings of the first workshop on social media analytics*. ACM, pages 71–79.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP*. pages 309–312.

Adam Mosseri. 2016. News feed fyi: Addressing hoaxes and fake news. http://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/. Accessed: 2017-01-30.

Eunbyung Park, Xufeng Han, Tamara L Berg, and Alexander C Berg. 2016. Combining multiple sources of knowledge in deep cnns for action recognition. In *Proceedings of AWACV*. IEEE, pages 1–8.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71:2001.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. pages 1532–1543.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. *Proceedings of EMNLP* pages 1120–1125.

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.

Slav Petrov. 2016. Announcing syntaxnet: The world's most accurate parser goes open source. *Google Research Blog, May* 12:2016.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of ACL*. pages 1650–1659.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP*. pages 105–112.

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.

Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. 2015. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of WWW*. pages 977–982.

Aldert Vrij, Samantha Mann, Susanne Kristen, and Ronald P Fisher. 2007. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior* 31(5):499–518.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820* .