Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues

Or Levi[1,*], **Pedram Hosseini**[2,*], **Mona Diab**[2,3] and **David A. Broniatowski**[2]

[1]AdVerifai
[2]The George Washington University, Washington DC, USA
[3]Amazon AWS AI, Seattle, USA
`or@adverifai.com, phosseini@gwu.edu`

Abstract

The blurry line between nefarious fake news and protected-speech satire has been a notorious struggle for social media platforms. Further to the efforts of reducing exposure to misinformation on social media, purveyors of fake news have begun to masquerade as satire sites to avoid being demoted. In this work, we address the challenge of automatically classifying fake news versus satire. Previous work have studied whether fake news and satire can be distinguished based on language differences. Contrary to fake news, satire stories are usually humorous and carry some political or social message. We hypothesize that these nuances could be identified using semantic and linguistic cues. Consequently, we train a machine learning method using semantic representation, with a state-of-the-art contextual language model, and with linguistic features based on textual coherence metrics. Empirical evaluation attests to the merits of our approach compared to the language-based baseline and sheds light on the nuances between fake news and satire. As avenues for future work, we consider studying additional linguistic features related to the humor aspect, and enriching the data with current news events, to help identify a political or social message.

1 Introduction

The efforts by social media platforms to reduce the exposure of users to misinformation have resulted, on several occasions, in flagging legitimate satire stories. To avoid penalizing publishers of satire, which is a protected form of speech, the platforms have begun to add more nuance to their flagging systems. Facebook, for instance, added an option to mark content items as "Satire", if "the content is posted by a page or domain that is a known satire

publication, or a reasonable person would understand the content to be irony or humor with a social message" (Facebook). This notion of humor and social message is also echoed in the definition of satire by Oxford dictionary as "the use of humour, irony, exaggeration, or ridicule to expose and criticize people's stupidity or vices, particularly in the context of contemporary politics and other topical issues".

The distinction between fake news and satire carries implications with regard to the exposure of content on social media platforms. While fake news stories are algorithmically suppressed in the news feed, the satire label does not decrease the reach of such posts. This also has an effect on the experience of users and publishers. For users, incorrectly classifying satire as fake news may deprive them from desirable entertainment content, while identifying a fake news story as legitimate satire may expose them to misinformation. For publishers, the distribution of a story has an impact on their ability to monetize content.

Moreover, in response to these efforts to demote misinformation, fake news purveyors have begun to masquerade as legitimate satire sites, for instance, carrying small badges at the footer of each page denoting the content as satire (Golbeck et al., 2018). The disclaimers are usually small such that the stories are still being spread as though they were real news (Funke, 2019).

This gives rise to the challenge of classifying fake news versus satire based on the content of a story. While previous work (Golbeck et al., 2018) have shown that satire and fake news can be distinguished with a word-based classification approach, our work is focused on the semantic and linguistic properties of the content. Inspired by the distinctive aspects of satire with regard to humor and social message, our hypothesis is that using semantic and linguistic cues can help to capture

---

*Authors contributed equally

these nuances.

Our main research questions are therefore, RQ1) are there semantic and linguistic differences between fake news and satire stories that can help to tell them apart?; and RQ2) can these semantic and linguistic differences contribute to the understanding of nuances between fake news and satire beyond differences in the language being used?

The rest of paper is organized as follows: in section 2, we briefly review studies on fake news and satire articles which are the most relevant to our work. In section 3, we present the methods we use to investigate semantic and linguistic differences between fake and satire articles. Next, we evaluate these methods and share insights on nuances between fake news and satire in section 4. Finally, we conclude the paper in section 5 and outline next steps and future work.

2 Related Work

Previous work addressed the challenge of identifying fake news (Conroy et al., 2015; Shu et al., 2017), or identifying satire (Burfoot and Baldwin, 2009; Reganti et al., 2016; Rubin et al., 2016), in isolation, compared to real news stories.

The most relevant work to ours is that of Golbeck et al. (Golbeck et al., 2018). They introduced a dataset of fake news and satirical articles, which we also employ in this work. The dataset includes the full text of 283 fake news stories and 203 satirical stories, posted between January 2016 and October 2017, with a main focus on American politics. These fake and satirical stories were verified manually such that each fake news article is paired with a rebutting article from a reliable source. This data carries two desirable properties. First, the labeling is based on the content and not the source, and stories spread across a diverse set of sources. Second, as also mentioned in (Golbeck et al., 2018), the fact that fake news and satire articles both focus on American politics minimizes the possibility that the topic of the articles will influence the classification.

In their work, Golbeck et al. studied whether there are differences in the language of fake news and satirical articles on the same topic that could be utilized with a word-based classification approach. A model using the Naive Bayes Multinomial algorithm is proposed in their paper which serves as the baseline in our experiments.

3 Method

In the following subsections, we investigate the semantic and linguistic differences of satire and fake news articles.[1]

3.1 Semantic Representation with BERT

To study the semantic nuances between fake news and satire, we use BERT (Devlin et al., 2018), which stands for Bidirectional Encoder Representations from Transformers, and represents a state-of-the-art contextual language model. BERT is a method for pre-training language representations, meaning that it is pre-trained on a large text corpus and then used for downstream NLP tasks. Word2Vec (Mikolov et al., 2013) showed that we can use vectors to properly represent words in a way that captures semantic or meaning-related relationships. While Word2Vec is a context-free model that generates a single word-embedding for each word in the vocabulary, BERT generates a representation of each word that is based on the other words in the sentence. It was built upon recent work in pre-training contextual representations, such as ELMo (Peters et al., 2018) and ULMFit (Howard and Ruder, 2018), and is deeply bidirectional, representing each word using both its left and right context. We use the pre-trained models of BERT and fine-tune it on the dataset of fake news and satire articles using Adam optimizer with 3 types of decay and 0.01 decay rate. Our BERT-based binary classifier is created by adding a single new layer in BERT's neural network architecture that will be trained to fine-tune BERT to our task of classifying fake news and satire articles.

3.2 Linguistic Analysis with Coh-Metrix

Inspired by previous work on satire detection, and specifically Rubin et al. (Rubin et al., 2016) who studied the humor and absurdity aspects of satire by comparing the final sentence of a story to the first one, and to the rest of the story - we hypothesize that metrics of text coherence will be useful to capture similar aspects of semantic relatedness between different sentences of a story.

Consequently, we use the set of text coherence metrics as implemented by Coh-Metrix (McNamara et al., 2010). Coh-Metrix is a tool for producing linguistic and discourse representations of

---

	PCA Component	Description	estimate	std.error	statistic	
Satire associated	**RC19**	First person singular pronoun incidence	1.80	0.41	4.38	***
	RC5	Sentence length, number of words	0.66	0.18	3.68	***
	RC15	Estimates of hypernymy for nouns	0.61	0.19	3.18	**
	RC49	Word Concreteness	0.54	0.17	3.18	**
	RC35	Ratio of casual particles to causal verbs	0.56	0.18	3.10	**
	RC91	Text Easability PC Referential cohesion	0.45	0.16	2.89	**
	RC20	Incidence score of gerunds	0.43	0.16	2.77	**
	RC32	Expanded temporal connectives incidence	0.44	0.16	2.75	**
	RC9	Third person singular pronoun incidence	0.44	0.16	2.67	**
	RC43	Word length, number of letters	0.45	0.20	2.27	*
	RC46	Verb phrase density	0.37	0.16	2.25	*
	RC97	Coh-Metrix L2 Readability	0.34	0.16	2.16	*
	RC61	Average word frequency for all words	0.50	0.24	2.13	*
	RC84	The average givenness of each sentence	0.37	0.18	2.11	*
	RC65	Text Easability PC Syntactic simplicity	0.38	0.18	2.08	*
	RC50	Lexical diversity	0.37	0.18	2.05	*
Fake news associated	**RC30**	Agentless passive voice density	-1.05	0.21	-4.96	***
	RC73	Average word frequency for content words	-0.72	0.20	-3.68	***
	RC59	Adverb incidence	-0.62	0.18	-3.43	***
	RC55	Number of sentences	-0.79	0.26	-3.09	**
	RC62	Causal and intentional connectives	-0.42	0.15	-2.72	**
	RC34	LSA overlap between verbs	-0.35	0.16	-2.22	*
	RC44	LSA overlap, adjacent sentences	-0.36	0.16	-2.16	*
	RC47	Sentence length, number of words	-0.36	0.18	-2.03	*
	RC89	LSA overlap, all sentences in paragraph	-0.34	0.17	-1.97	*
	(Intercept)		-0.54	0.19	-2.91	

Table 1: Significant components of our logistic regression model using the Coh-Metrix features. Variables are also separated by their association with either satire or fake news. **Bold**: the remaining features following the step-wise backward elimination. *Note*: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

a text. As a result of applying the Coh-Metrix to the input documents, we have 108 indices related to text statistics, such as the number of words and sentences; referential cohesion, which refers to overlap in content words between sentences; various text readability formulas; different types of connective words and more. To account for multicollinearity among the different features, we first run a Principal Component Analysis (PCA) on the set of Coh-Metrix indices. Note that we do not apply dimensionality reduction, such that the features still correspond to the Coh-Metrix indices and are thus explainable. Then, we use the PCA scores as independent variables in a logistic regression model with the fake and satire labels as our dependent variable. Significant features of the logistic regression model are shown in Table 1 with the respective significance levels. We also run a step-wise backward elimination regression.

Those components that are also significant in the step-wise model appear in bold.

4 Evaluation

In the following sub sections, we evaluate our classification model and share insights on the nuances between fake news and satire, while addressing our two research questions.

4.1 Classification Between Fake News and Satire

We evaluate the performance of our method based on the dataset of fake news and satire articles and using the F1 score with a ten-fold cross-validation as in the baseline work (Golbeck et al., 2018).

First, we consider the semantic representation with BERT. Our experiments included multiple pre-trained models of BERT with different sizes and cases sensitivity, among which the large un-

cased model, **bert_uncased_L-24_H-1024_A-16**, gave the best results. We use the recommended settings of hyper-parameters in BERT's Github repository and use the fake news and satire data to fine-tune the model. Furthermore, we tested separate models based on the headline and body text of a story, and in combination. Results are shown in Table 2. The models based on the headline and text body give a similar F1 score. However, while the headline model performs poorly on precision, perhaps due to the short text, the model based on the text body performs poorly on recall. The model based on the full text of headline and body gives the best performance.

To investigate the predictive power of the linguistic cues, we use those Coh-Metrix indices that were significant in both the logistic and step-wise backward elimination regression models, and train a classifier on fake news and satire articles. We tested a few classification models, including Naive Bayes, Support Vector Machine (SVM), logistic regression, and gradient boosting - among which the SVM classifier gave the best results.

Table 3 provides a summary of the results. We compare the results of our methods of the pre-trained BERT, using both the headline and text body, and the Coh-Mertix approach, to the language-based baseline with Multinomial Naive Bayes from (Golbeck et al., 2018)[2]. Both the semantic cues with BERT and the linguistic cues with Coh-Metrix significantly outperform the baseline on the F1 score. The two-tailed paired t-test with a 0.05 significance level was used for testing statistical significance of performance differences. The best result is given by the BERT model. Overall, these results provide an answer to research question RQ1 regarding the existence of semantic and linguistic difference between fake news and satire.

4.2 Insights on Linguistic Nuances

With regard to research question RQ2 on the understanding of semantic and linguistic nuances between fake news and satire - a key advantage of studying the coherence metrics is explainability. While the pre-trained model of BERT gives the

---

[2]We were not able to reproduce the same result as in the original paper, possibly due to the difference between Scikit-learn and Weka, tools that were used to train classifiers in our work and the original paper, respectively. To make our results comparable, we replicated the experiments of the original paper using Scikit-learn.

Model	P	R	F1
Headline only	0.46	0.89	0.61
Text body only	0.78	0.52	0.62
Headline + text body	**0.81**	**0.75**	**0.78**

Table 2: Results of classification between fake news and satire articles using BERT pre-trained models, based on the headline, body and full text. **Bold**: best performing model. *P*: Precision, and *R*: Recall

Method	P	R	F1
Baseline	0.70	0.64	0.67
Coh-Metrix	0.72	0.66	0.74*
Pre-trained BERT	**0.81**	**0.75**	**0.78***

Table 3: Summary of results of classification between fake news and satire articles using the baseline Multinomial Naive Bayes method, the linguistic cues of text coherence and semantic representation with a pre-trained BERT model. Statistically significant differences with the baseline are marked with '*'. **Bold**: best performing model. *P*: Precision, and *R*: Recall. For Coh-Metrix, we report the *mean* Precision, Recall, and F1 on the test set.

best result, it is not easily interpretable. The coherence metrics allow us to study the differences between fake news and satire in a straightforward manner.

Observing the significant features, in bold in Table 1, we see a combination of surface level related features, such as sentence length and average word frequency, as well as semantic features including LSA (Latent Semantic Analysis) overlaps between verbs and between adjacent sentences. Semantic features which are associated with the gist representation of content are particularly interesting to see among the predictors since based on Fuzzy-trace theory (Reyna, 2012), a well-known theory of decision making under risk, gist representation of content drives individual's decision to spread misinformation online. Also among the significant features, we observe the causal connectives, that are proven to be important in text comprehension, and two indices related to the text easability and readability, both suggesting that satire articles are more sophisticated, or less easy to read, than fake news articles.

5 Conclusion and Future Work

We addressed the challenge of identifying nuances between fake news and satire. Inspired by the humor and social message aspects of satire articles,

we tested two classification approaches based on a state-of-the-art contextual language model, and linguistic features of textual coherence. Evaluation of our methods pointed to the existence of semantic and linguistic differences between fake news and satire. In particular, both methods achieved a significantly better performance than the baseline language-based method. Lastly, we studied the feature importance of our linguistic-based method to help shed light on the nuances between fake news and satire. For instance, we observed that satire articles are more sophisticated, or less easy to read, than fake news articles.

Overall, our contributions, with the improved classification accuracy and towards the understanding of nuances between fake news and satire, carry great implications with regard to the delicate balance of fighting misinformation while protecting free speech.

For future work, we plan to study additional linguistic cues, and specifically humor related features, such as absurdity and incongruity, which were shown to be good indicators of satire in previous work. Another interesting line of research would be to investigate techniques of identifying whether a story carries a political or social message, for example, by comparing it with timely news information.

References

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Facebook. Fact-checking on facebook: What publishers should know. http://bit.ly/2ouEDwJ. [Online].

Daniel Funke. 2019. Facebooks fact-checking project sometimes forces hoaxers to change their content. http://bit.ly/2m89KNr. [Online].

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya

Chanduka, Paul Cheakalos, Jennine B Everett, et al. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, pages 17–21. ACM.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Danielle S McNamara, Max M Louwerse, Philip M McCarthy, and Arthur C Graesser. 2010. Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Aishwarya N Reganti, Tushar Maheshwari, Upendra Kumar, Amitava Das, and Rajiv Bajpai. 2016. Modeling satire in english text for automatic detection. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 970–977. IEEE.

Valerie F Reyna. 2012. A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making*.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

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