Satire vs Fake News: You Can Tell by the Way They Say It

Dipto Das and Anthony J. Clark
Department of Computer Science
Missouri State University
Springfield, Missouri, USA
Email: dipto175@live.missouristate.edu

***Abstract*—In recent times, "fake news" has become an increasingly important concept. Primarily, because information is now able to more quickly and deeply propagate among users due to the pervasive nature of the Internet and digital media. For this reason, it has recently received a large amount of attention from computer science researchers. A large number of studies demonstrate methods for detecting misinformation in content shared on the Internet. On the other hand, satire and irony as a part of usual human communication have received less attention. Whereas, fake news means misinformation meant to deceive people, satire is misinformation meant to entertain or criticize. Thus, despite both satire and fake news being misinformation these two concepts have different objectives and impacts. Currently, few studies have focused on differentiating between satire and fake news. In this paper, we present the limitations of existing works for classifying satire and fake news; discuss the feasibility of using a subjective concept like storytelling as a way to classify satire and fake news; and present a supervised learning approach to classify satire and fake news.**

I. Introduction

Satire and fake news are both based on misinformation. The difference between them is their motivation. Though existing literature thoroughly investigates how to detect misinformation in digital contents, there has not been much research to identify motivation. We argue that the way misinformation is conveyed, i.e. the style of storytelling, is a good indicator of the motivation and effort of the person(s) behind that misinformation. We also show how this concept can be used to design a supervised learning model for distinguishing between satire and fake news.

Though fake news detection is a well studied field of computer science, to the best of our knowledge, Golbeck et. al. [5] is the only work in existing literature to address the problem of classifying satire and fake news. In their work, they present a dataset for fake news and satire. They showed applicability of naïve bayes algorithm to classify satire and fake news from the corresponding texts. However, we found that their approach is highly biased to the buzzwords of the period when the articles of the dataset were collected. For example, we found that the dataset contains terms like Obama, Trump, etc. and the naïve bayes model by [5] uses these terms to distinguish between satire and fake news. However, these terms are very specific to American politics during the time around the election of 2016. Thus, this approach looses universality with respect to time.

We argue that since the motivation and the targeted audience of satire and fake news are different, there will be difference in the storytelling approach while propagating these different types of articles. Fake news are shared with a view to deceiving people. This objective of deception often becomes successful when there is no reliable medium of verifying information and the targeted audience also do not have sufficient data and context information. On the other hand, the motivation behind satire is to criticize someone. The objective of satire fulfills when its targeted audience have access to enough context information to understand the basis, i.e. event behind it.

We used the dataset presented by Golbeck et al. [5]. First, we show how preprocessing the data can improve performance of their proposed model. Next, we identify the most influential factors behind their model and evaluate their correlation with the time period of the data collection and found high bias. We studied how storytelling approaches vary with the categories of articles – satire and fake news. Then, we used the variation of tones used in articles to differentiate satire and fake news. Since, the storytelling approach is largely independent of any particular time, we argue that our proposed approach is more widely applicable than the approach by Golbeck et al. [5].

The contribution of this paper is divided into two parts. First, we identify flaws of the existing approach and showed how performance of the existing model can be improved by using the text data from the articles. Second, we discuss how the approach of conveying message differs from satire to fake news, and propose a supervised learning approach to classify satire and fake news.

II. Background

Prior studies [16], [21] discuss the definition of "fake news". According to them, news satire, news parody, manipulation, fabrication, and large-scale hoaxes are different kinds of fake news. However, the problem with this definition is that it does not consider the motivation. In our work, we followed the definition by Golbeck et al. [5]. According to them, fake news is misinformation that is presented with the motivation to deceive the consumers. They excluded satire from the definition of fake news because of the different motivations. Golbeck et al. [5] did not provide a definition for satire, so, we followed the definition by Merriam-Webster Dictionary [13] that says satire is "a literary work holding up human vices and follies

978-1-7281-4127-5/19/$31.00 ©2019 IEEE
DOI 10.1109/TransAI46475.2019.00012

to ridicule or scorn; or trenchant wit, irony, or sarcasm used to expose and discredit vice or folly."

Since satire and fake news only differ in motivation, we have to first consider how human users actually recognize satire from fake news. Without access to information about the source of the article (e.g. website that publishes the article might be known for sharing satire) an important clue about the nature of the article can be the storytelling approach of the article. Narrative trajectory based on sentiment is an important indicator of the storytelling patterns of text articles [23], [4], [15], [17]. The main idea behind this is that though sentence-wise sentiment scores of an article corresponds to individual reader experience, if we filter/smooth the sentence-wise scores for a large amount of text, the variation can indicate narrative style/pattern of the articles of specific category [4]. Existing literature uses several different sentiment analysis approaches, including: Wordnet [17], [20], PCA [15], [1], and the IBM Tone Analyzer [8], [22], [7]. In our work, we used IBM Tone Analyzer because of its wide spectrum of considered sentiments.

III. Improvement on the Existing System and Drawbacks

Here, we use the dataset prepared by Golbeck et. al. [5]. They collected and annotated 203 satirical stories and 283 fake news stories. Their dataset was collected articles related to American politics after January 2016. They justified this decision to ensure minimal topic variation in the dataset. They also performed an empirical analysis on the themes of the articles in the dataset and found seven different categories: (1) hyperbolic position against a person or a group, (2) hyperbolic position in favor of a person or a group, (3) discredit a normally credible source, (4) sensationalist crime and violence, (5) racist messaging, (6) paranormal theories, and (7) conspiracy theories. They showed the applicability of multinomial naïve Bayes classifier in the classification context of satire and fake news. Their classifier achieved 79.1% accuracy with ROC area (a representation and interpretation of the area under a receiver operating characteristic (ROC) curve obtained by predictions by the model [6]) of 0.88. They concluded that this shows a high difference between the type of language in satire and fake news in their dataset.

At first, we used multinomial naïve Bayes classifier proposed by Golbeck et. al. [5] with some changes. Instead of using the text directly, we stemmed (reduced words to their root/base forms; e.g.: working $\rightarrow$ work) the words using Lovins Stemmer algorithm [11]. This reduced the probability of considering the same word differently due to different structures of the sentences. We discarded the stopwords (the words that do not have much significance in word based queries, e.g.: articles) defined in [12]. Including these steps improved the accuracy of the performance to an accuracy of 80.3% with a ROC area of 0.87.

In our study, we investigated how the model makes decision or distinguishes satire from fake news. We find out which words the classifier was using to differentiate between satire and fake news. We used Shannon information gain [19] based attributes evaluation on the word vectors of the article corpus for this purpose. The top 15 words contributing most to classification of satire and fake news are: Obama, report, Donald, good, people, Clinton, Trumps, years, Barack, jobs, States, dress, United, Hillary, and government. Words with the most information gains are shown as wordcloud in Figure 1.

Fig. 1: Wordcloud of the words with high information gain.

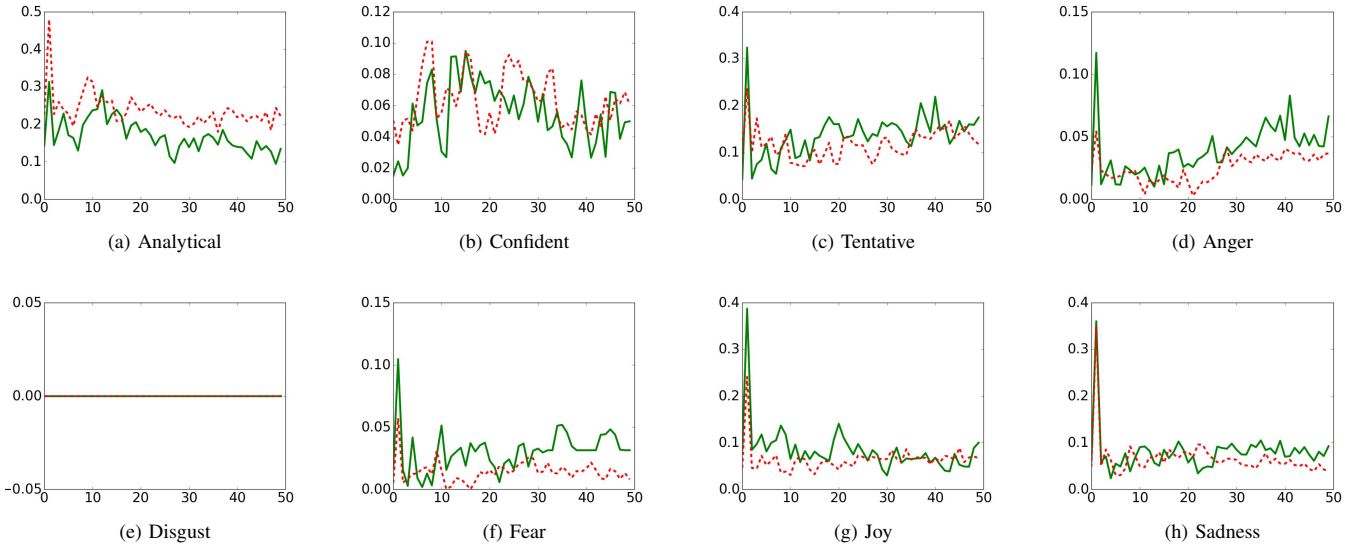
Here, we can see that the words that contribute most while using naïve Bayes classifier are mostly proper nouns or part of proper nouns (e.g. United, States) related to recent American politics. The other high information gain yielding words are also closely related to American politics. Since, the dataset was curated within the specific domain of American politics, it is expected to have many words regarding this as distinguishing terms. However, high information gain of the proper nouns show that the model is highly specific to the terms used in a specific period of time. This can be viewed as a drawback of both the existing naïve Bayes classifier [5] and our improved version.

IV. Tone as a Way to Differentiate

We hypothesize that the person or group who create fake news and satire use different approaches in their content creation or writing. Thus, the tone conveyed in a satire will be different from the tone conveyed in a fake news. Also, it is likely that the trajectory of this level of sentiments/tones will have different trajectories according to different categories of articles – satire and fake news.

We used the IBM Tone Analyzer to calculate different aspects of each article. It outputs scores (in a scale from 0.0 to 1.0) representing the tone conveyed by sentences. IBM Tone Analyzer calculates 13 kinds of tone that belong to 3 different classes.

a) Language Scores: IBM tone analyzer takes three aspects of language of an article as follows: analytical (the amount of technical substance and reasoning); confidence (the



(a) Analytical (b) Confident (c) Tentative (d) Anger

(e) Disgust (f) Fear (g) Joy (h) Sadness

Fig. 2: Comparison between narrative trajectories of satire (green solid line) and fake news (red dashed line) for different tones.

degree of expression of certainty); and tentative (the amount of words expressing uncertainty).

b) Emotion Scores: IBM tone analyzer calculates the probability of a sentence to express each of the following emotions: angry, joy, fear, disgust, and sadness.

c) Social Scores: IBM tone analyzer calculates the likelihood of a sentence to express five personality characteristics as follows: agreeableness, conscientiousness, emotion, extraversion, and openness.

For constructing narrative trajectories, we followed the algorithm presented by [22]. We calculated these scores for each article in both categories. Then, we used the scores of each sentence in an article to construct the narrative trajectory of that particular article. We considered the scores for a specific tone in an article as a signal S_{raw}. Next, we used a Hanning smoothing window with size = 3, to construct a smooth signal S_{smooth}. Then, we cropped the signal to remove the boundary effects introduced by filtering. Finally, the smoothed and cropped signal S_{crop} is interpolated to have a canonical length of 50 samples. We refer this final signal as the narrative trajectory.

We argue that a satire article would differ from a fake news article in the way of describing an event. For example, since the motivation behind creating a fake news is to make people believe something, the content creator needs to make it look like a real news, hence, be more analytic while writing. Likewise, if a fake news tries to disseminate a conspiracy theory, it will try to convey fear. Whereas a satire needs to be funny to the readers, a fake news obviously will not have such tone in it. We constructed narrative trajectories for all articles in both categories. Then, to verify the applicability of our argument, we calculated the resultant signal of summation of all the signals from the articles in each category.

As we can see, satire articles in the dataset often had different narrative trajectories with slightly different amplitudes than the fake news articles in the dataset. For example, analytical scores for satire articles were not as high as the ones for fake news (Figure 2a); satire articles' angry tone level was often higher than that of fake news (Figure 2d) which might indicate the exaggeration of emotion in satire posts and attempt of the fake news to look like unbiased like a real news. We also observed that social scores had almost no trajectory in their narrative approach, and thus there was not much difference in the signals generated for satire and fake news categories. We also did not observe much difference from the graphs for disgust emotion tone score trajectory and confidence language score trajectory.

V. Domain Independent Classification Based on Tone

According to the discussion in the previous section, we argue that we can use tone information to classify satire and fake news articles. If we use the tone scores to train the models instead of the text directly, it will make the models less dependent on the exact text data, and thus, less confined to any specific domain or time period.

We argue that the headlines of satire and fake news articles might have relevant sentiment information about the article. Therefore, we calculated the subjectivity and polarity of sentiment conveyed by the headline using TextBlob [10]. We extracted the tone data using IBM Tone Analyzer. We recorded the overall tone data conveyed by the article as document tone data. Then, we calculated sentence-wise tone data using IBM Tone Analyzer. Thus, we obtained features as following: (1) two features from headline: subjectivity and polarity; (2) thirteen tone data (three language tone, five emotion tone, five social tone) for document; (3) thirteen summation of tone data for all sentences in the document. We also added the number

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Satire	0.729	0.212	0.775	0.729	0.751	0.518	0.827	0.833
Fake	0.788	0.271	0.743	0.788	0.765	0.518	0.827	0.788
Weighted Avg.	0.758	0.242	0.759	0.758	0.758	0.518	0.827	0.811

TABLE I: Performance of classification task with tone data extracted from articles (article text independent approach)

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Fake	0.905	0.254	0.782	0.905	0.839	0.660	0.911	0.894
Satire	0.746	0.095	0.887	0.746	0.811	0.660	0.911	0.919
Weighted Avg.	0.826	0.174	0.834	0.826	0.825	0.660	0.911	0.907

TABLE II: Performance of classifier model with text, tone, and theme data combined

of sentences in the article as a feature to train our model. In total, we have 29 features for learning our model.

The dataset provided by Golbeck et. al. [5] has 203 satire articles (41.7%) and 283 fake news articles (58.3%). Hence, the dataset is slightly biased. Therefore, we decided to apply SMOTE (Synthetic Minority Over-sampling Technique) [2] on the minority class satire with 40% oversampling ratio. We used Random Forest classifier for this classification task between satire and fake news. We achieved 75.8% accuracy with ROC area 0.83. Detailed performance results are shown in Table I.

We achieved comparable performance without using text data unlike the existing work [5]. We hypothesize that if we use tone data along with text data, it will show increased performance in classifying satire and fake news. Like the existing work [5], we also added the theme information with these features. With all these features combined, we achieved 82.5% accuracy with ROC area 0.91. We show the detailed performance results using Random Forest classifier [9] for this classification in Table II. We used Scikit-learn [14] for training the model.

VI. Discussion

We used data processing steps like stopwords elimination and stemming that improved the performance of the system by a small margin. Whereas naïve Bayes text classifier is limited by the used terms in the articles in the dataset and thus the trained model is likely to be confined to be useful for only specific domain and time period, our proposed approach using tone data extracted from the text is less dependent on exact words of articles and thus is less likely to be confined to any specific domain or time period. We achieved comparable performance using this approach and we showed that combining tone information with text and theme data of the articles can improve the performance of the model by a considerable margin. However, we further investigated the contribution of the features of our model to classify satire and fake news articles using Shannon information gain [19]. Table III shows the top five features in our model with highest information gain. We can see that though word vectors generated from model are associated with our model, tone and theme based features have highest information gain, and thus can be good features for classifying satire and fake news.

Feature	Information Gain
Conspiracy (theme)	0.1035
Document Joy (tone)	0.0668
Document Analytical (tone)	0.0402
Sentences Analytical (tone)	0.0395
Sensationalist Crime/Violence (theme)	0.0390

TABLE III: Five features with topmost information gain values (type of the feature is inside parentheses)

VII. Future Works and Conclusions

The existing works on narrative style focus on English texts. Since novelty of our proposed model is being domain and period independent, we plan to study its applicability across different languages. Existing studies in satire detection suggest images to be useful [3], [18], hence images associated with the articles can be incorporated in a multimodal model for classification of satire and fake news. The model proposed by this paper uses tone information of the articles. Our model shows promising 75.8% accuracy without using text data directly and improved 82.5% accuracy while combined with text data.

References

[1] Christopher M Bishop. 2006. *Pattern recognition and Machine Learning*. Springer.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[3] Dipto Das and Anthony J Clark. 2018. Sarcasm Detection on Flickr Using a CNN. In *International Conference on Computing and Big Data (ICCBD)*.

[4] Jianbo Gao, Matthew L Jockers, John Laudun, and Timothy Tangherlini. 2016. A multiscale theory for the dynamical evolution of sentiment in novels. In *International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*. IEEE, 1–4.

[5] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B Everett, and others. 2018. Fake News vs Satire: A Dataset and Analysis. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 17–21.

[6] James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.

[7] IBM. 2017a. The science behind the service. https://console.bluemix.net/docs/services/tone-analyzer/science.html#the-science-behind-the-service. (2017). Online; accessed 29 September 2018.

[8] IBM. 2017b. Tone Analyzer, Understand emotions and communication style in text. https://www.ibm.com/watson/services/tone-analyzer/. (2017). Online; accessed 29 September 2018.

[9] Andy Liaw, Matthew Wiener, and others. 2002. Classification and regression by random forest. *R news* 2, 3 (2002), 18–22.
[10] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, and others. 2014. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing* (2014).
[11] Julie Beth Lovins. 1968. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics* 11, 1-2 (1968), 22–31.
[12] Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. (1996). http://www.cs.cmu.edu/ mccallum/bow.
[13] Merriam-Webster Dictionary. n.a. Satire Definition. https://www.merriam-webster.com/dictionary/satire. (n.a.). Online; accessed 25 September 2018.
[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.
[15] Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5, 1 (2016), 31.
[16] Victoria L Rubin, Yimin Chen, and Niall J Conroy. 2015. Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. American Society for Information Science, 83.
[17] Spyridon Samothrakis and Maria Fasli. 2015. Emotional sentence annotation helps predict fiction genre. *PloS one* 10, 11 (2015), e0141922.
[18] Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1136–1145.
[19] Claude E Shannon. 1948. A note on the concept of entropy. *Bell System Tech* 27, 3 (1948), 379–423.
[20] Carlo Strapparava, Alessandro Valitutti, and others. 2004. Wordnet affect: an affective extension of wordnet.. In *The International Conference on Language Resources and Evaluation*, Vol. 4. Citeseer, 1083–1086.
[21] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining “fake news" A typology of scholarly definitions. *Digital Journalism* 6, 2 (2018), 137–153.
[22] M Iftekhar Tanveer, Samiha Samrose, Raiyan Abdul Baten, and M Ehsan Hoque. 2018. Awe the Audience: How the Narrative Trajectories Affect Audience Perception in Public Speaking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 24.
[23] Kurt Vonnegut. 1999. *Palm Sunday: an autobiographical collage*. Dial Press.