Towards Automatic Fake News Classification

Souvick Ghosh
Rutgers University, USA. souvick.ghosh@rutgers.edu

Chirag Shah
Rutgers University, USA. chirags@rutgers.edu

ABSTRACT

The interaction of technology with humans has many adverse effects. The rapid growth and outreach of the social media and the Web have led to the dissemination of questionable and untrusted content among a wider audience, which has negatively influenced their lives and judgment. Many research studies have been conducted to tackle the detection and spreading of fake news, which is misinformation that looks genuine. While the first step of such tasks would be to classify claims associated based on their credibility, the next steps would involve identifying hidden patterns in style, syntax, and content of such news claims. We propose a generalized method based on Deep Neural Networks to detect if a given claim is fake or genuine. We have used a modular approach by combining techniques from information retrieval, natural language processing, and deep learning. Our classifier comprises two main submodules. The first submodule uses the claim to retrieve relevant articles from the knowledge base which can then be used to verify the truth of the claim. It also uses word-level features for prediction. The second submodule uses a deep neural network to learn the underlying style of fake content. Our experiments conducted on benchmark datasets show that for the given classification task we can obtain up to 82.4% accuracy by using a combination of two models; the first model was up to 72% accurate while the second model was around 81% accurate. Our detection model has the potential to automatically detect and prevent the spread of fake news, thus, limiting the caustic influence of technology in the human lives.

KEYWORDS

fake news, deep neural network, fake news detection, machine learning, classification, social media.

INTRODUCTION

With the increasing popularity and outreach of social media channels, fake news or misinformation could spread faster than ever before, reach a broader audience, and influence public opinion on a deeper level. Therefore, it has become increasingly important to address the concerns about fake news, using all system-level and human-level approaches. Creating public awareness on how to judge a news item for veracity is one such approach, as is developing algorithmic methods to act as a first step in combating this ever-increasing problem. As different disciplines of information science have been working towards mitigating this problem; it is essential to have a precise definition of the problem. While fake news means false stories or misinformation, it can be defined by other common terms like *satire, propaganda,* and *rumor*. In our current research, we have tried to address the problem of fake news by developing a classifier which can automatically detect fake news accurately. We have trained our classifier on several benchmark datasets which comprise short sentences containing fake and real news obtained from several credible and fake sources.

RELATED WORKS

Rubin, Chen, and Conroy (2015) identified three types of fake news in their work - serious fabrications, large-scale hoaxes, and humorous fake news. Conroy, Rubin, and Chen (2015) were one of the first researchers to use network analysis in fake news detection while Mukherjee and colleagues (2013) used words and the respective part-of-speech tags, together with bigrams to achieve a 68.1% accuracy on Yelp review classification. Shu and colleagues (2017) provided a detailed overview of the recent approaches towards fake news detection and similar problems. While the problem of fake news detection is relatively new, there have been several attempts to tackle it from an algorithmic (more specifically, machine learning) perspective. One such problem was proposed in the Fake News Challenge[1] (2017) where the participating teams were asked to detect the stance of the news claim. Researches have used modified versions of bidirectional LSTM/GRU architectures (Zeng, Zhou and Xu, 2017; Chopra, Jain, and Sholar, 2017), ensemble of classifiers (Thorne et al., 2017), vanilla CNNs, independent encoders, conditional encoder (Rakholia and Bhargava, 2016), multipass conditional encoders, attentive readers with or without weighted cross entropy function (Miller and Oswalt, 2017) and bidirectional LSTMs.

DATASET

The Fake News Challenge Dataset (Rubin, Chen, and Conroy, 2015) contains around 13000 short headlines and 2587 full articles. Each instance contains a headline (which is mostly short), a reference to one of the articles, and the stance of the article towards the claim. The stance could be *agree*, *disagree*, *discuss*, or *unrelated*. Though the challenge approached the fake news

[1] Fake news challenge stage 1 (fnc-i): Stance detection, 2017. URL http://www.fakenewschallenge.org/. https://www.theonion.com/

through stance detection, which is unique and interesting, yet it requires a classification based on a pair of claim and article. University of Washington Fake News Dataset contains a total of 49000 instances, each comprising a paragraph of news article collected from credible and fake news sources (e.g., The Onion[1]). Each claim has one of four possible labels: *hoax*, *propaganda*, *satire*, or *true news*. The length of each sentence ranges from 500 to 600 words. Although the dataset was developed for a similar problem, we made slight modifications to make it more generalizable. For example, we removed all sentences which were labeled as satire as we theorize that satire is more of a linguistic phenomenon (intended for humor) than fake news (Rashkin et al., 2017).

METHODOLOGY

The proposed model consists of several smaller submodules, each responsible for categorizing the instances based on a set of features. Finally, we combine the results through a voting process, which is based on a weighted average where the weights are also learned by the deep neural model. In this research, we have focused on two main submodules: the veracity detection submodule (based on information retrieval models and knowledge base) and the style based submodule. The main module can be extended by adding other submodules such as author metadata (background information, posting history, etc.) or cognitive authority of the source. The first submodule is responsible for checking the veracity of each claim given that we have already constructed a knowledge base. First, the most relevant documents are retrieved from the knowledge base. Secondly, given those documents, the stance of the claim towards the documents is inferred. For retrieval, we used TF-IDF method as a baseline and more advanced algorithms for comparison and improved performance. We also implemented and tested BM25, Vector Space and Language Models. After the *k* related articles are retrieved, in the second step of the algorithm, each article is classified into three labels '*Fake*', '*Suspicious*' or '*Legit*.' For the classification, any deep learning architecture can be used. The input features of the classifier are two one-hot bags-of-word vectors of size 5000, one corresponding to the news statement and the other to the article. Both vectors are fitted on the vocabulary of 5000 most frequently used words in the knowledge base. Additionally, it takes the cosine similarity between these two vectors as an additional input, hence, extending the final size of the input vector to 10001. The hidden layer of the model has 100 Rectified Linear Units (ReLU), and the final layer is a SoftMax layer with three output classes as mentioned before. The second submodule of our model is responsible for gaining valuable insights into how the writing style of fake news differs from real news. The syntax, semantics, and style of the written text can provide significant information about the intention of the authors. It has been widely observed that the language and tone of fake news presentation are more aggressive in general, and it involves a choice of words depicting strong emotions and biases (Rashkin et al., 2017). Our model uses a deep, bidirectional LSTM architecture.

RESULTS AND DISCUSSION

The veracity-based submodule retrieved the most relevant documents relative to the claim and classified the claim into three possible mutually exclusive categories: *fake*, *suspicious* and *real*. The accuracy of prediction was 67.1% for ternary classification and 72.12% for binary classification. The style-based submodule, when evaluated separately on the UW test dataset, predicts with an accuracy of 81.83% (the best performing architecture). Finally, by combining both the submodules using a weighted average, we were able to slightly increase the accuracy to 82.4%. Our contribution is not limited to constructing an accurate model, but it advances the literature on the fake news by evaluating how different retrieval techniques can be incorporated to deep neural architecture to create a more robust and flexible model. By modularizing the architecture, we allow for further enhancements and modules, such as the cognitive authority of source, mining of social media and public opinion and so on.

CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a universal model to verify the authenticity of news claims. By using different features like the authenticity of the source, perceived cognitive authority, style, and content-based factors, and natural language features, it is possible to accurately predict fake news. Our experiments conducted on benchmark datasets show that for the given classification task we can obtain up to 82.4% accuracy by using a combination of two models; the first model was up to 72% accurate while the second model was around 81% accurate. Our detection model has the potential to automatically detect and prevent the spread of fake news, thus, limiting the caustic influence of technology in the human lives.

FORMAT OF PRESENTATION

The presentation will be in the form of either a poster or a three to four minutes video which will highlight the details of the study and the key findings.

[1] https://www.theonion.com/

REFERENCES

Chopra, S., Jain, S., & Sholar, J. M. (2017). Towards Automatic Identification of Fake News: Headline-Article Stance Detec- tion with LSTM Attention Models.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4.

Miller, K., & Oswalt, A. (2017). Fake News Headline Classification using Neural Networks with Attention.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. S. (2013, July). What yelp fake review filter might be doing?. In *IC- WSM*.

Rakholia, N., & Bhargava, S. (2016). 'Is it true?–Deep Learning for Stance Detection in News'.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Pro- cessing* (pp. 2931-2937).

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22-36.

Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X., & Vlachos, A. (2017). Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*(pp. 80-83).

Zeng, Q., Zhou, Q., & Xu, S. (2017). Neural Stance Detectors for Fake News Challenge.

81st Annual Meeting of the Association for Information Science & Technology | Vancouver, Canada | 10 – 14 November 2018