Getting Real on Fake News

Connor Masini and Thanaphong Phongpreecha

I. Problem Description

Recently, the circulation and distribution of news articles of dubious credibility has become very widespread. With the ease of sharing news articles through social media platforms like Twitter and Facebook, these "Fake News" articles can quickly spread and manipulate hundreds of users into thinking they are true. The purpose of this project is to create a tool for the analysis of the credibility of Twitter messages. Such a tool would allow users to check the validity of a news article they find hard to believe. To check the validity of the article, we will examine factors including things like subscriber count, subscribed count, verified user account information, number of re-tweets, and of course the information found in the article itself. The combination of all these factors should allow us to create an effective tool that determines the legitimacy of a news article.

II. Related Works

Labeling of dataset of news for its veracity can primarily be achieved via three means of analysis: expert, crowd sourcing, and computational oriented. While experts opinions are reliable, the process is time-consuming and difficult to scale, resulting in attempts to employ reinforcement learning for an automated labeling process. [1] The research in this area is closely related to rumor detection and and information credibility evaluation. These methods are typically based on building predictive models to classify whether a news article is fake. At a high-level, models developed in this line of work typically rely on:

- **Linguistic**: Analysis of the article headlines, contents, or users' stance using natural language processing (NLP) is one of the primary methods of analyzing credibility. The underlying assumption for this methodology is that since fake news pieces are intentionally created for financial or political gain rather than to report objective claims, they often contain opinionated and inflammatory language, crafted as clickbait or designed to incite confusion. [2] Using methods to capture linguistic features such as N-grams, bag-of-words, Latent Dirichlet Allocation (LDA) models, or word embedding with models such as recurrent neural networks (RNN) or supported vector machines (SVM), a number of studies have constructed models with varying success for different datasets. [1], [3]–[6] These approaches, however, are limited by the fact that the linguistic characteristics of fake news are not yet fully understood. Furthermore, the distinguishing characteristics of unreliable news vary across different topics, and media platforms. [7]
- **Network and Content Propagation:** It can be very challenging, if not impossible, to identify useful features from textural content alone as intentional spreaders of fake news may manipulate the content to make it appear like real news. To address this problem several studies instead focus on exploring the topology of information diffusion, *i.e.* graph mining. By employing RNN to capture features such as news propagation frequency combined with embedded user techniques, studies were able to achieve high F_1 and accuracy scores. [6], [8], [9]
- **Combinatorial:** Only a few recent studies have started to combine these features in the hope of achieving better accuracy within a shorter period of time (since we want to inhibit the propagation of false news). For example, a model integrating user's features with the post's textural features and propagation resulted in a high accuracy of 90% successful classification by using Long-Short Term Memory (LSTM), a specific type of RNN. [7], [10]. In a separate study, that used just the user and textural features with a logistic regression classifier, a lower classification success rate was achieved. [11] This emphasizes the importance of selecting features as well as the efficacy of the RNN.

III. Objectives

Previous research has provided useful insights to tackle this issue, however, the models used in many studies also come with significant setbacks for use in real-world applications. For example, using content propagation would require the news to spread significantly before it can be detected and dealt with accordingly. Additionally, there is still a significant accuracy gap that can be improved.

In this work, we aim to classify fake news using a combination of linguistic and network features. Moreover, we aim to explore different models and their structures to improve the accuracy. For example, exploring the use of end-to-end memory network compared to either LSTM or GRU would allow us to create a more effective tool. End-to-end memory networks were shown to be able to capture long-term structure within sequences and are an effective way to classify documents. [12], [13] We would also explore the use of algorithms to capture textural features prior to feeding to the model, for example, instead of just bag-of-word or N-grams, we will be investigating word-embedding techniques such as word2vec that can better preserve meaning and relations between words. [14]

We will be using data obtained through the Twitter API with labeled training data provided by a previous study.[1] [9]

1. http://alt.qcri.org/∼wgao/data/rumdect.zip

IV. Milestones

Project Schedule

Week 1	2/19-2/25	• Literature Review • Dataset Acquisition • Set-up Development Environment
Weeks 2-4	2/26-3/18	• Design Models • Examine/Preprocess Features • Begin Training of Models
Weeks 5-7	3/19-4/8	• Analyze Deep Features • SOMETHING Architecture • Model Training
Weeks 8-9	4/9-4/25	• Finalize Results • Prepare Final Report • Project Submission

References

[1] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

[2] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM, 2015.

[3] David M Markowitz and Jeffrey T Hancock. Linguistic traces of a scientific fraud: The case of diederik stapel. *PloS one*, 9(8):e105937, 2014.

[4] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 647–653, 2017.

[5] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

[6] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1751–1754. ACM, 2015.

[7] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806. ACM, 2017.

[8] Liang Wu and Huan Liu. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. 2018.

[9] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, pages 3818–3824, 2016.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Arkaitz Zubiaga. Learning class-specific word representations for early detection of hoaxes in social media. *arXiv preprint arXiv:1801.07311*, 2018.

[12] Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, page 2237, 2017.

[13] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[14] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.