Detecting fake news using hybrid neural networks

Eliza Vardanyan
Department of Computer Science
Earlham College
Richmond, Indiana
evarda17@earlham.edu

.

1 ABSTRACT

Detection of misinformation has become of great relevance and importance in the past few years. A significant amount of work has been done in the field of fake news detection using natural text processing tools. Combined with many other filtering algorithms. However, these studies lacked a thorough linguistic analysis of the text to assist in detecting fake news. In order to improve the fake news detection accuracy rate, my project addresses the potential in using a thorough linguistic analysis for fake news detection. This study will examine the impact of linguistic analysis and user features in the prediction accuracy of misinformation. I will use a Hybrid approach to complete lexical analysis and integrate user feature's weight in the process of fake news detection.

2 INTRODUCTION

Detection of misinformation has become of great relevance and importance in the past few years. With the ongoing COVID-19 pandemic, many have witnessed the harmful and dangerous impact of misinformation, especially in the healthcare field. With the rapid spread of misinformation in politics, health care, and in many more areas, rose the issue of separating fake news from the real one. According to Paskin, fake news refers to "particular news articles that originate either on mainstream media (online or offline) or social media and have no factual basis but are presented as facts and not satire" [10]. A significant amount of work has been done in the field of fake news detection using language-based Approach[5]. However, these approaches failed to investigate the role that user profile contributing to the text can play in detecting fake news. In this research paper, I will be focusing on a hybrid approach to detecting fake news. This approach will follow a hybrid model, which will extract textual features using LSTM (Long Short-Term Memory) and then apply the user features' weight to produce vectors. This hybrid model is based on Ruchansky et al.[11]. My project is different from the previous work done in this field with its sentiment analysis when using LSTM. I will be using NLTK (Natural Language Toolkit). In one of the three components of the Hybrid model proposed in Ruchansky et al.[11], sentiment analysis will occur. I believe that sentiment analysis can contribute to achieving higher accuracy.

The paper is structured as follows: Section 2 discusses approaches used in other works in this field. Section 3 focuses on overall experimental design of the research, as well as describes the necessary tools used in the process. Section 4 provides information regarding future work and a timeline.

CS388 Methods for Research and Dissemination, , Earlham College

3 RELATED WORK

This section addresses previous work in the field of fake news detection. It provides an overview of the models used and touches upon some of the most common tools used in those models. Fake news detection has heavily depended on Neural Networks as a way of text processing due to its accuracy. This is due to the high accuracy and precision of classification that Neural Networks demonstrated in the past. In this project, I will be combining Recurrent Neural Networks (RNN) in a hybrid model. Other than using RNN, this hybrid model will also integrate three main characteristics of the news: text, user response, and news source. Kwon et al. introduce three main features associated with rumor classification: 1. Temporal Features; 2. Structural Features; 3. Linguistic Features [2]. The focus of this paper will be the use of Linguistic Features for sentiment analysis. This section will review related work done in fake news detection from a Linguistic analysis point of view.

Zhang et al.[14] introduced FAKEDETECTOR, which focuses on deep diffusive network models. This approach is unique with its four component analyses of the dataset. First, it completes an article credibility analysis that includes textual content, then it does creator credibility analysis, goes over creator-article publishing historical records, and computes subject credibility analysis. The paper reviews the word cloud of 'true' and 'false' statements in the first component analysis. It uses words that differentiate true and false news/statements as signals for future distinguishing purposes. These are part of the Explicit Feature Extraction steps in this paper. The next step in this analysis that again distinguishes this work from related work in the field is the Latent Feature Extraction. This step goes further, exploring hidden signals related to the publishers, statements, creators, authors. This is done with the use of deep recurrent neural network models. The RNN model results are then used as an input for Deep Diffusive Unit and Gated Diffusive Unit models. These models draw correlations between creators, authors, publishers, articles, and the content's subject.

Zhou et al. [15] is unique in its approach to the problem with its observance of fact-tampering attack method of detecting fake news. The paper highlights the importance of linguistic characteristics analysis in fake news detection. The proposed idea mainly focuses on linguistic features without doing fact-checking of statements.

To visually portray this approach, the authors propose a knowledge graph known as the Straw Man Solution Approach. This idea of knowledge graphing via Crowdsourcing is widespread among well-known companies such as Google and Reddit. The paper is significant in fake news detection since it reviews all the most common and effective methods in the field and highlights the benefits and the defects of these methods. This paper shows that, in the real-world, linguistic characteristics, if observed alone, are not as effective. However, when combined with other fake news detection filters, they can significantly improve accuracy. Another related paper in this field, Yardi et al, introduces three features of identifying spam on the Twitter platform. Those features are URL analysis, username pattern analysis, and keyword detection in spams [12]. Similar work completed by O'Donovan et al. mentions that "the most useful indicators" of news credibility are URLs, retweets, and tweets' length [2] [8].

4 IMPLEMENTATION

My approach consists of two main components: automatic identification of features from the given input of news statements and classification. The first component of our approach is built on a hybrid learning model. This model utilizes Long Short-Term Memory (LSTM) based on RNNs. My proposed model views fake news detection through the lens of three main components: text analysis, source, and user response. This section will describe the elements of the hybrid model and the three main components in more detail. The design of this model can be seen in Figure 1.

4.1 Dataset

The dataset we will use is the result of data collection in Jing Ma et al. The dataset is constructed using Twitter (www.twitter.com) and Sina Weibo (www.weibo.com) [6]. It consists of 2,313 rumors and 2,351 non-rumors. Rumor, according to social psychology literature, is defines as "a story or a statement whose truth value is unverified or deliberately false" [3]. With this definition in mind, we will refer to fake news as rumors and authentic news as non-rumors.

4.2 Data pre-processing

These documents will be pre-processed to fit the input format of the RNN model. We will complete document vectorization to represent the text numerically and feed it to RNN. Pre-processing of data also consists of normalization and tokenization. Our pre-processing model will convert input data into numbers, and these predictor vectors will be of fixed length.

4.2.1 Tokenization. Tokenization is converting the data into tokens of strings while stop words are "a set of commonly used words in any language" [1].

4.2.2 Normalization. Normalization of the text consists of removing stop words from the text, stemming, and synonym mapping [4]. Stemming maps "different forms of nouns into a single semantically similar word" [4]. There are 153 stop word terms in English that are removed when we complete data pre-processing.

4.3 RNNs

Recurrent Neural Networks "are a class of neural networks that allow previous outputs to be used as inputs while having hidden states" [1] When comparing neural Networks, I decided that RNN would be the best model to represent articles since, as shown in Table 1, RNN works with textual data while CNN is more accurate with image [9].

Table 1: Types of Neural Networks

	RNN	CNN
Data	Sequence data (Time series, text, audio)	Image data
Recurrent connections	Yes	No
Parameter sharing	Yes	Yes
Spatial relationship	No	Yes
Vanishing and Exploding Gradient	Yes	Yes

RNN has some limitations and problems associated with it. One of those is the problem of vanishing gradient. This has been resolved with the introduction of back-propagation, which leads us to LSTM.

4.3.1 The learning architectures: LSTM. The main focus of my project will be LSTM. LSTM networks are a modified version of RNNs [7]. Like mentioned earlier, RNNs have a problem with vanishing gradients. With LSTM networks, this problem is resolved. LSTM allows an easier way of "remembering past data in memory" [7]. The training of the LSTM model is based on back-propagation. LSTM consists of three gates: input gate, forget gate, and output gate. Each of these gates uses the Sigmoid function, which will be used in my approach.

4.4 Text analysis

This part of the project focuses on mining "particular linguistic cues" [11] where specific patterns of conjunctions and words can be associated with misinformation. We will be using recurrent neural networks, following the model proposed by Ma et al. [6] that will be using linguistic features to analyze text. When analyzing article a_j, we capture temporal patterns present in data. We represent feature vector with y_t, and it will have the following form [11]:

$$y_t = (\eta, \delta t, y_u, y_\tau),$$

where η represents number of engagements and δt captures the time that has been between user engagements. in the equation above y_u captures user feature vectors while y_τ is representation of the "text characteristic of an engagement" [11].

4.5 User response and Source

The response feature mostly focuses on the response that the article has received. This will be studied with the help of a social graph [11]. Related work in this field also approaches this feature with the "hand-crafted social network dependent behavior approach, e.g., the number of Facebook likes" [11]. However, this approach requires a lot of labor since it's hand-crafted. Source characteristics of fake news focus on studying "the source of an epidemic analyzing the social graph." Similar work has been focusing on identifying anomalies. An example of such work includes Yu et al. [13] uses a Bayes model to detect anomalies. User response is represented

via weight matrix W_a, which is fixed for all y_t. When user response weight is applied to article feature vectors, the resulting output is a vector:

$$v_j = tanh(W_r h_T + b_r),$$

where h_T is the last hidden state in the fully connected layer or RNN.

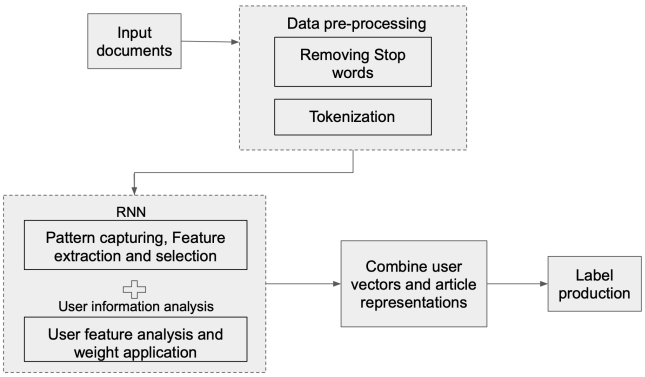

Figure 1: Basic project framework

5 TIMELINE

Week 1 (5 Oct)
- Start implementation of LSTM networks
- Implement sentiment analysis with the use of NLTK
- Implement article representation and user response vectorization
Week 2 (12 Oct)
- Report the findings from LSTM networks and vectorization in the first draft of the paper
Week 3 (19 Oct)
- Implement user response vector's weight application to RNN model
Week 4 (26 Oct)
- Start producing labels for the articles
Week 5 (2 Nov)
- Finish label production
- Report the results from label production in the second draft of the paper
Week 6 (9 Nov)
- Compare my results to those of Ruchansky et al.
- Improve any text analysis aspect if possible
Week 7 (16 Nov)
- Finalize and polish everything
- If there is time, experiment with a different dataset same model
Week 8 (23 Nov, finals)
- Finalize and complete the final paper

6 ACKNOWLEDGEMENT

I would like to thank Dr. David Barbella for the support and feedback that he provided through the preparation of my project, developing it into a polished research proposal.

REFERENCES

[1] Shervine Amidi Afshine Amidi. 2020. What are stop words? Retrieved September 18, 2020 from https://kavita-ganesan.com/what-are-stop-words/#.X2mXZy2z1QI
[2] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2018. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th international conference on social media and society*. 226–230.
[3] Gordon W Allport and Leo Postman. 1947. The psychology of rumor. (1947).
[4] Muhammad Bux Alvi, Naeem A Mahoto, Majdah Alvi, Mukhtiar A Unar, and M Akram Shaikh. 2018. Hybrid classification model for twitter data-a recursive preprocessing approach. In *2018 5th International Multi-Topic ICT Conference (IMTIC)*. IEEE, 1–6.
[5] Matthee M. de Beer, D. 2020. Approaches to Identify Fake News: A Systematic Literature Review. Retrieved September 6, 2020 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7250114/
[6] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. (2016).
[7] Aditi Mittal. 2019. Understanding RNN and LSTM. Retrieved September 21, 2020 from https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e
[8] John ODonovan, Byungkyu Kang, Greg Meyer, Tobias Höllerer, and Sibel Adalii. 2012. Credibility in context: An analysis of feature distributions in twitter. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 293–301.
[9] Aravind Pai. 2020. CNN vs. RNN vs. ANN - Analyzing 3 types of Neural Networks in Deep Learning. Retrieved September 18, 2020 from https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/
[10] Danny Paskin. 2018. Real or fake news: who knows? *The Journal of Social Media in Society* 7, 2 (2018), 252–273.
[11] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 797–806.
[12] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. 2010. Detecting spam in a twitter network. *First Monday* (2010).
[13] Rose Yu, Xinran He, and Yan Liu. 2015. Glad: group anomaly detection in social media analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10, 2 (2015), 1–22.
[14] Jiawei Zhang, Bowen Dong, and S Yu Philip. 2020. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1826–1829.
[15] Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019. Fake news detection via NLP is vulnerable to adversarial attacks. *arXiv preprint arXiv:1901.09657* (2019).