Text Classification of News Articles with Speed, Volume and Circuitousness

Aim and Objectives

In the era of information overload and rampant dissemination of news articles across digital platforms, distinguishing between reliable and unreliable sources has become increasingly challenging. Traditional classification approaches often rely solely on textual content analysis, overlooking crucial contextual factors that contribute to the credibility and nature of news articles. Narration in one such factor that plays a significant role in making unreliable and misinformation news sources believable. Here we propose quantifying narration of news articles by creating a high-dimensional latent representation of text. We use this semantic path to compute speed, volume and circuitousness and relate them to the bias or reliability of news articles. We also compute a a sentiment score of each news articles as an additional measure to detect bias in news sources. We use various state-of-the-art machine learning techniques and natural-language processing methods to represent text in a latent high dimensional space. Specifically, we use Google's Word2Vec model to represent text into a latent space We use these spacial representations to define features of the semantic path like speed, volume and circuitousness and develop a framework of relating these dimensions to various categories of news articles, including reliable, unreliable, hate news, fake news, clickbait, political, conspiracy theory, and satire.

Digital Object Identifier 10.1109/MCE.YYYY.Doi Number

Date of publication DD MM YYYY; date of current version DD MM YYYY

Literature Review

Tackling the dissemination of misinformation and fake news on social media platforms has become a critical area of research, with various methodologies and approaches being explored. In recent studies, Support Vector Machines (SVMs) have been utilized for supervised binary classification of fake news [1], [2]. These studies not only employed SVMs but also compared them with other classifiers, finding SVMs to outperform alternatives. Additionally, efforts have been made to enhance SVM classifier performance through user reputation and credibility analysis [3].

Apart from SVMs, Convolutional Neural Networks (CNNs) have emerged as prominent architectures for fake news classification. For instance, a CNN-based approach integrating text and image information was proposed by [4], which combined explicit and latent features for improved detection. Furthermore, CNN-based word-embedding models, such as those learned by syllable unit as suggested by [5], have shown promise in fake news detection.

In addition to CNNs, Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) have gained popularity in news classification tasks [6]. These architectures have been employed for both fake news classification and detection, contributing to the diverse range of approaches in this field.

Furthermore, machine learning systems have demonstrated effectiveness in clickbait detection

within news articles. Studies like [7] have proposed systems to detect the stance of headlines in relation to their corresponding article bodies, aiding in clickbait identification. Additionally, [8] introduced a hybrid categorization technique integrating various features and sentence structures for differentiating clickbait from non-clickbait articles.

However, despite the progress in fake news and clickbait detection, there remains a gap in addressing other harmful types of news articles, such as hate news, junk science, and extremely biased content. The lack of literature focusing on the detection or mitigation of these types of articles, coupled with insufficient available data, poses a significant challenge. Future research efforts could concentrate on bridging this gap to comprehensively address the spread of harmful misinformation in society.

With the increasing sophistication of machine learning models and the availability of vast amounts of data, there is potential for more nuanced approaches to detecting and mitigating various forms of harmful news content. Collaborative efforts between researchers, industry stakeholders, and policymakers will be crucial in developing effective strategies to combat misinformation and safeguard the integrity of information dissemination platforms.

Methodology

Dataset: We used the Fake News Corpus of News Articles [9] which is an open-source dataset comprising of millions of articles scraped from around 1001 domains.

Data Preprocessing: The news articles are first preproecessed and clean. First, all words are converted to lowercase and stops words are removed from the text content by using list of stop words from the nltk corpus. The words in clean text is then lemmatized. Next, we break the content into windows of sizes depending on the word count of text. As a thumb rule, we choose a window size of 50 for a word count between 100 and 250. Next, we use the Word2Vec model [10] that represents approximately 1 million words into a 300-dimensional latent space to look up the embedding vector, represented as, x_w, of each word in the windows. For each window, an average vector is computed.

Speed: It is the sum of distance between vectors of each window divided by length of text [11]. Let d_{i-j} denote the distance between i_{th} and j_{th} vector and L be the length of the text then:

$$\text{speed} = \frac{\sum d_{i-j}}{L} \tag{1}$$

Volume: To determine volume, we first begin by finding the minimum-volume ellipsoid that encloses all vectors in space associated with the text [11]. If the rank of the subspace spanned by the vectors is greater than 300, the dimensions are reduced using Principal Component Analysis and the volume of the convex hull is calculated as length of Eigen vectors of the vector space divided by square root of the Eigen vectors. However, if the dimensionality of the subspace is less than 300, then convex hull is flat and the problem is degenerate.

Circuitousness: It is the ratio of the distance travelled between all vectors to the distance travelled in the shortest path between the vectors [11]. We use the Minimal Spanning Tree to find the shortest path.

Furthermore, the sentiment score is assigned by computing the difference between the positive and negative words divided by the word count of the text content.

CONCLUSION

The conclusion of the research paper underscores the significance of narrative analysis in detecting various elements such as clickbait, fake news, hate news, and bias within news articles. Through the preprocessing and quantification of thousands of articles from the Fake News Corpus, the study revealed the importance of considering factors like speed, volume, and circuitousness in understanding the narrative structure of news content.

Moreover, the integration of sentiment analysis provided deeper insights into the tone and voice used in the articles. However, the conclusion highlights that these approaches can be further enhanced when combined with other Natural Language Processing (NLP) techniques. By leveraging a combination of methods, such as narrative analysis, sentiment analysis, and other NLP techniques, researchers can develop more comprehensive strategies for identifying and combating misinformation in news articles.

Looking ahead, future research will explore the training of deep learning models on a wider range of features, including speed, volume, circuitousness, and other latent path characteristics, for text classification purposes. By harnessing the power of deep learning and incorporating various textual features, researchers

aim to improve the accuracy and effectiveness of classification models for identifying different types of news content.

REFERENCES

1. Vasu Agarwal, H. Parveen Sultana, Srijan Malhotra, and Amitrajit Sarkar. Analysis of classifiers for fake news detection. *Procedia Computer Science*, 2019.
2. Hadeer Ahmed, Issa Traoré, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1, 2018.
3. Suchitra Deokate. Fake news detection using support vector machine learning algorithm. *International Journal for Research in Applied Science and Engineering Technology*, 7:438–444, 07 2019.
4. Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu. Ti-cnn: Convolutional neural networks for fake news detection. *ArXiv*, abs/1806.00749, 2018.
5. Pawan Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. Welfake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, PP:1–13, 04 2021.
6. Pritika Bahad, Preeti Saxena, and Raj Kamal. Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Comput. Sci.*, 165(C):74–82, jan 2019.
7. Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. *From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles*, pages 84–89. Association for Computational Linguistics, Copenhagen, Denmark, September 2017.
8. Abinash Pujahari and Dilip Singh Sisodia. Clickbait detection using multiple categorisation techniques. *Journal of Information Science*, 47(1):118–128, 2021.
9. OpenAI. https://github.com/several27/FakeNewsCorpus, 2022. Accessed: February 10, 2024.
10. Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.
11. Olivier Toubia, Jonah Berger, and Jehoshua Eliashberg. How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, 118(26):e2011695118, 2021.