Engineering Multifaceted Features for Robust Fake News Classification

Uthkarsh Reddy Junuthula
uthkarshreddyjunuthula@my.unt.edu

Jashia Mitayeegiri
JashiaMitayeegiri@my.unt.edu

Mohammed Abdul Azeem
abdulazeemmohammed@my.unt.edu

Mohammed Sabith Abul Faiz
mohammedsabithabulmohammedsabithabul@my.unt.edu

1 Goals and Objectives:

1.1 Motivation:

In this age of social media, we are consuming news at a much higher rate than ever. In our day to day lives, we come across a wide range of news articles, however, the challenge is to analyse which stories are true and which are fabrications. The alarming rise in the broadcasting of fake news poses a grave threat to society at large, influencing public opinion and even shaping political landscapes. The urgency to tackle this issue is further amplified by data revealing a considerable increase in the spread of misinformation in recent years.

1.2 Significance:

Fake news has been prevalent for ages, but we are currently new to online fake news which is virtual and has got no bounds to it as the user is not

physically present and no one can be found as the main source of it. It has got high influences on the people and has shifted the locus of interest related to the contemporary issue. The dataset we have chosen will help to detect the rumours and fake news used in the articles. This mainly helps us to gain insight into the fake news and the latent pattern involved in it which promotes us to implement a safer internet space. Analytics will enable us to recognize the vulnerable communities and the influence of fake news on people and also to protect these communities by providing them with advocacy support. The better detection of algorithms will permit us to capture the facts and the fakes used and authorised to filter content and limit such kind of content generated. We can enforce better community guidelines and effective regulations which comply with various legal policies.

2 Objectives

The objectives of this project are multi-layered:

- **Data Collection and Preprocessing:** To collect a comprehensive data-set of news articles and preprocess it for analysis, including handling missing values, filtering the data-set, and normalizing the text.
- **Feature Engineering:** To extract and engineer relevant features from the news data that can be used to train the machine learning model. This includes analyzing text length, sentiment, stylistic elements, and readability scores.
- **Model Development and Training:** To develop a machine learning model that uses these features to classify news articles as real or fake.
- **Evaluation and Optimization:** To evaluate the model's performance and optimize it for higher accuracy and efficiency.

2.1 Features:

For this project, we have extracted a variety of features from both the text and the title of the news articles to distinguish between real and fake news. These features include and analyze both the content and style of the articles. We have analysed the Textual content through preprocessing methods such as tokenization and lemmatization and performed the sentiment on both

titles and texts too. Stylistic features include factors such as counts of exclamation marks, question marks, use of all caps, and repeated letters, which often help in differentiating sensational content from factual content. Use of proper vocabulary is measured through the number of unique words used, and the readability scores gives us the insights into the complexity of the text. Additionally, a combined analysis of titles and texts offers a more complete understanding of the articles.

3 Related Work (Background):

In recent times, the issue of fake news and misinformation has escalated significantly, posing a major challenge. The growth of social media and digital news platforms has accelerated the swift spread of false information to vast audiences. A 2019 Pew Research survey discovered that 68% of American adults feel that their trust in governmental institutions is impacted by fake news (Mitchell et al., 2019)[1]. Furthermore, a separate study indicated that more than 90% of individuals have at some point been deceived by misleading news headlines (Silverman, 2016)[2]. Based on these studies and many more, the focus has been recently shifted majorly towards creating automated techniques for identifying fake news. Current research primarily employs machine learning for categorizing articles as either authentic or fraudulent. An early study by Rubin et al. (2016) [3] focused on analysing fraudulent news stories from the 2016 U.S. presidential election. In this study, they extracted various textual features, both syntactic and semantic, which were utilized in a support vector machine framework, attaining over 90% accuracy in identifying fraudulent stories. Potthast et al. (2018) [4] examined stylometric elements like the complexity of writing, punctuation use, and readability, noting that fraudulent news often employs simpler language and more dramatic tones. Rashkin et al. (2017) [5] examined the verifiability of claims in articles, developing a model to assess the accuracy of individual statements in news stories. Recently, deep learning approaches, including RNNs and transformers, have become prevalent in fake news detection. For example, Wang et al. (2018) [6] applied LSTMs to a substantial dataset from Weibo, achieving a 95% accuracy rate. In a more recent study, Zhou et al. (2020) [7] implemented RoBERTa models, generating sentence-level embeddings from news content. These neural network frameworks have demonstrated effectiveness in learning the semantic nuances of text. Despite these advancements, many studies

face challenges such as limited dataset sizes, oversimplified models, and a lack of model transparency. Future research needs to focus on developing larger, more diverse datasets, models that are both interpretable and sophisticated, and frameworks adaptable to various news sources and topics (Shu et al., 2020) [8]. As the methods of misinformation continue to evolve, the development of dynamic, low-effort detection systems becomes increasingly vital.

4 Dataset:

The dataset contains 3 columns namely Title, Text and Label. The articles provided in this dataset range from political commentary to news reports with a mix of factual and fake news, which have been labelled accordingly.

Here's a brief overview of the contents of the dataset:

- Title: This column contains the titles of articles or texts.
- Text: This column has the main article.
- Label: This column is populated with integer values, where 0 denotes fake news and 1 denotes factual news.

5 Detail design of Features Extraction:

A. Text Preprocessing:

- **Tokenization:** In this phase, the title and text columns will be split into individual words also known as tokens, making them easier to analyse and process for the next steps.
- **Stop word Removal:** Common words like 'the', 'is', and 'in', known as stopwords, often don't contribute to the meaning of a text for analysis purposes. Removing them will help focus the analysis on more significant words, which are more likely to contribute to the classification of the news as real or fake.
- **Lemmatization:** This step involves converting words to their base or root form. For example, 'falling' becomes 'fall'. It helps in reducing

the complexity of the text data and merging different forms of a word into a single representation.

B. Sentiment Analysis:

- **Polarity Scores:** Sentiment analysis is conducted on both titles and texts to capture the emotional tone of the articles. Polarity scores range from -1 (very negative) to +1 (very positive). Our hypothesis regarding this feature is that fake news could possibly exhibit distinct sentiment patterns (e.g., overly negative, or positive tones) compared to genuine articles.

C. Stylistic Elements:

- **Exclamation and Question Marks:** Counts of these punctuation marks are considered, as an overuse of these can indicate a focus towards sensationalism or a particular emotional appeal, which is common in misleading news.

- **All-Caps Words:** Again, just like exclamation and question marks, frequent use of all-caps can also be a stylistic indicator of sensationalism, which could be more predominant in fake news.

- **Repeated Letters:** Unnecessary repetition of letters in words (e.g., "Nooooo") can be a stylistic feature associated with informal or exaggerated language.

D. Vocabulary Richness:

- **Unique Word Counts:** This refers to the count of distinct words in titles and texts. A richer vocabulary might indicate a more experienced or diverse language use, possibly correlating with the authenticity of the content.

E. Readability Scores:

- **Text Complexity Assessment:** Using metrics like the Flesch Reading Ease score from the "teststate" library, the complexity of the text is assessed. This score is based on sentence length and word syllable

count. It can provide us insights into the target audience and the style of the text.

F. Vectorization:

- **TF-IDF Representation:** Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate the importance of a word in a document, relative to a corpus. This technique will be applied to transform the title and text columns into numerical representations, emphasizing words that are unique to each document.

Implementation

Using the TF-IDF vectors along with text-specific features

The TF-IDF vectors along with the text-specific features will enhance the text classification as these specific features will provide us with content that the TF-IDF vectors might not be able to capture, for example, the use of punctuations might not be considered rightly with the TF-IDF vectors but by adding them the performance will be increased.

Using Word2vec vector along with text-specific features

The word2vec is a widely used word embedding technique where the vectors are generated based on the context and neighborhood of a word. We will have the texts to be tokenized and the word2vec model is built with the corpus, later the model generates the vectors of each word. In our case, we have the word vectors and take the average of it for the sentence vectors.

Using the TF-IDF weighted Word2vec vectors

The method will enable us to capture the advantages by both the word2vec and the TF-IDF. We can make the vectors much more specific as we add the TF-IDF weights of the words as the word in a document is assigned weight which reflects the importance of the word in the document. This captures the semantic relationship between the words encoded in the word2vec vectors

and the importance of words in the documents and the Corpus encoded in the TF-IDF.

Using the TF-IDF weighted Word2vec vectors along with text-specific features

This method gives a comprehensive approach to the classification as we include all the above features making a features fusion which will make the dataset include all the important features.

Results

From the below results, we could see the integration of TF-IDF weights, word2vec vectors, and the text-specific features has made the best model with a low number of iterations having an accuracy of 94.71% and furthermore the total trainable parameters are just 5,889. The loss and accuracy of the best model are as shown. From the graphs, we can see that the model is not overfitting or underfitting.

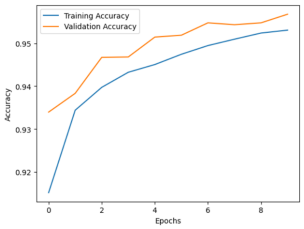

Figure 1: Accuracy

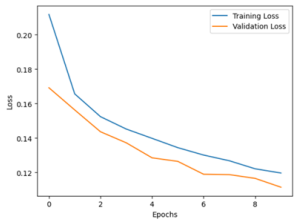

Figure 2: Loss

Models	TF-IDF vectors + text characteristics	Word2vec vectors + text characteristics	TF-IDF weighted Word2vec vectors	TF-IDF weighted Word2vec vectors + text characteristics
Logistic Regression	92%	92%	88%	91%
Linear Discriminant analysis	89%	90%	87%	90%
Artificial neural networks	93.32%	93.95%	89.47%	94.17%

Figure 3: Comparision

In the machine learning the logistic regression has been getting accuracy of 92% with just the TF-Idf vectors an characters this model is simple but gives the 2nd best accuracy, the scores and the ROC AU are as follows:

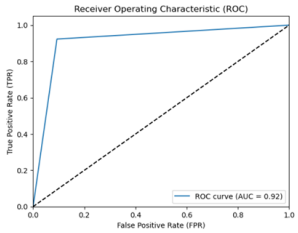

Figure 4: ROC

	precision	recall	f1-score	support
0	0.91	0.92	0.91	6929
1	0.92	0.91	0.92	7379
accuracy			0.92	14308
macro avg	0.92	0.92	0.92	14308
weighted avg	0.92	0.92	0.92	14308

Figure 5: FPR

Implementation Status Report

Work Completed:

Uthkarsh Reddy Junuthula **Work (25% of contribution)**
Contribution: Feature Engineering
Responsibility: Researching and extracting features that would be useful for building a robust classification system.

Jashia Mitayeegiri **Work (25% of contribution)**
Contribution: Model Building

Responsibility: Based on the features extracted, have built the models and analysed the results.

Mohammed Abdul Azeem **Work (25% of contribution)**
Responsibility: Data Preprocessing and Model Implementation
Contribution: Explored the data, Data Visualization on the data labels, the labels of the data seem to be balanced. Visualized the word cloud. Performed text cleaning which involves removal of HTML tags, special characters, punctuations, exclamation marks, question marks. Removed numbers, performed Stemming and Lemmatization. Applied Vectorization on the text data.

ammed Sabith Abdul Faiz **Work (25% of contribution)**
Contribution: Report preparation and PPT preparation
Responsibility: Based on the discussions and codes created the project increment 1 report and the PPT.

Work to be completed:

Utkarsh Reddy Junuthula **Work (25% of contribution)**
Responsibility: Using merger model architecture
Description: We would like to have a merger model architecture which will take two inputs simultaneously and process through different sets of layers, and at the end, the processing is merged to get the results out.
Issues/Concerns: We have used the text characteristics extracted and also the vectors which might require different hidden layers to process the input.

Jashia Mitayeegiri **Work (25% of contribution)**
Responsibility: Using transfer learning to impact the model with prior knowledge
Description: For enhancing the model further, we will be using the transfer learning approach by leveraging the advantage of the pre-trained models. The model which has prior knowledge along with the features we extracted will be able to achieve higher accuracy than the 1st iteration.
Issues/Concerns: The models built are simple but have good ac-

curacy but the use of transfer learning will let the model have prior knowledge.

Mohammed Abdul Azeem **Work (25% of contribution)**
Responsibility: Use of Recurrent Neural Networks
Description: We have only used neural networks which might not capture the context of the sentence, but we would like to see how the Recurrent Neural Networks like GRU or LSTM would work for classification as the RNNs have been very effective in terms of context.
Issues/Concerns: Many of the state-of-the-art models in natural language processing are made up of RNNs, specifically the encoder-decoder architecture. The use of RNNs will provide us with better results and enable a comparison.

ammed Sabith Abdul Faiz **Work (25% of contribution)**
Responsibility: Analysing the results of various models
Description: The results from various other models are compared and the reasons why a specific model has been performing is analysed. The results are also compared with the state of the art models.
Issues/Concerns: The comparison with the state-of-the-art models is imperative as it enables us to understand why the model is performing well or if there are any tweaks that will make the model more efficient.

6 Data-set and GitHub

Link to the data-set: `https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification`.

The project can be found at `https://github.com/Uthkarshh/Feature_Engineering_Project_Group_8`.

References

[1] Mitchell, A., Gottfried, J., Barthel, M., & Sumida, N. (2019). Distinguishing Between Factual and Opinion Statements in the News. *Pew Research Center*.

[2] Silverman, C. (2016). This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook. *BuzzFeed News*.

[3] 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.

[4] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231-240.

[5] 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. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931-2937.

[6] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 849-857.

[7] Zhou, X., Jain, A., Phoha, V. V., & Zafarani, R. (2020). Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2), 1-25.

[8] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3), 171-188.