

Attention-based Deep Learning Model for Accurate Detection of Misleading News

Lakshit Jain¹, Mayank Tomar², Mohammad Abdullah Wasi³, Priyanka Jain⁴

¹⁻⁴Department of Electronics and Communication Engineering Delhi Technological University Delhi, India
Email: lakshitjain_2k19ec099@dtu.ac.in, mayanktomar_2k19ec108@dtu.ac.in,
mohammadabdullahwasi_2k19ec109@dtu.ac.in, priyankajain@dtu.ac.in

Abstract— With the inception of big data, there is a lot of information that goes unfiltered, contributing to fake news. This false information can lead to catastrophic results in the long run. Hence, to curb this spread, we aim to provide constructive techniques that can help detect this kind of misleading news. In this paper, we have introduced a novelty model, the GRU-Attention Network, and shown the comparison of the novelty model with existing machine learning algorithms like Naïve Bayes, SVM, BERT, and DeBERTa. The analysis is based on the dataset, which was extracted from diverse sources and contains over 20,000 news samples. The proposed work results in an accuracy ranging from 85.7% to 99.9%, which is very good for classifying most of the information.

Index Terms— Naïve Bayes, SVM, BERT, DeBERTa, GRU-Attention Network.

I. INTRODUCTION

The phenomenon of faking the news has become a widespread issue in this world, affecting people from all walks of life. Not only is it spread deliberately by those with nefarious intentions, such as gaining political power or spreading propaganda against certain groups, but it is also often unknowingly shared by individuals who are not aware of its false nature. The proliferation of social media has made it easier than ever for false information to reach a large audience. With advancements of internet and the widespread use of social media, an overwhelming amount of information is readily available to the public, but determining its credibility can be a challenging task. "Fake news" is a term used to describe information that is false and intended to harm the reputation of an organisation, individual, or group, or to gain financially or politically. The spread of fake news not only undermines trust in media and institutions but also has the potential to create societal divisions, sow confusion, and lead to dangerous consequences. It is crucial that individuals take the time to fact-check information and question its source before sharing it with others. Additionally, media and social media platforms have a responsibility to take steps to combat the spread of fake news by verifying information and providing resources for fact-checking. It is a collective effort by everyone to counter the spread of false information and ensure that accurate information is being disseminated. Therefore, it is important for people to stay vigilant and educate themselves about the potential dangers of fake news and how to identify and avoid it.

The spread of false information, commonly referred to as "fake news," has become a major concern in today's age of instant global communication. The misinformation can lead to confusion and undue stress among the public and has the potential to harm democracy and individual freedoms. It is often used for personal or political gain and can result in violence and mistrust in government and media institutions. This practise is referred to as 'infodemic,'

which, just like the pandemic that we all went through in the recent years, spreads false information quickly. It is crucial for individuals to be discerning when consuming information and for authorities to take action to combat the spread of fake news.[1].

In this research, a new and innovative approach is proposed: the GRU-Attention Network for detecting fake news using state-of-the-art deep neural network models that incorporate attention mechanisms such as BERT and DeBERTa. The effectiveness of these attention-based models is assessed by evaluating their performance in terms of accuracy when applied to both training and validation datasets. This paper is structured in a logical and coherent manner, with the first section providing a comprehensive introduction to the subject matter. The second section delves into the existing algorithms that have been used for the detection of fake news, while the third section describes the methodology employed in this research. The fourth section presents a detailed comparison of the results obtained from the proposed algorithm with other existing algorithms. Finally, the fifth section concludes the paper by summarising the key findings and discussing potential future work in this field.

II. LITERATURE REVIEW

A. Fake News on social media

Social media encompasses wikis, microblogging, social bookmarking, forums, social websites, and other similar platforms[2][3]. The ease of sharing and the large reach of social media make it a prime channel for the dissemination of fake news. Fake news's effects on social media can be significant, as it can affect common people's opinions and lead to the spread of misinformation. Furthermore, it can also lead to political interference, manipulation of public opinion, and even manipulation of stock markets[4].

B. Natural Language Processing

NLP is an area of AI that focuses on the interaction between computers and human language. It uses various techniques to analyze, understand, and generate human language. It has a variety of applications, including speech-to-text and text-to-speech systems, analysis of emotions, machine translation, and information retrieval. Recently, then advancement of deep learning techniques has greatly improved the performance of NLP, resulting in more accurate and natural language understanding.

The creation of tools for separating actions from speech is a crucial topic of NLP research. This requires the usage of numerous pipelines that speak various languages, including Chunking, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Emotion Analysis and Detection, and Semantic Role Labelling. Together, these methods enable us to recognise and properly extract the intended meaning from spoken language.

Recent advancements in NLP have also led to the development of chatbots and virtual assistants that are able to understand and respond to natural language input. This has led to the widespread use of NLP in industries such as customer service and e-commerce, where it is used to improve customer engagement and automate repetitive tasks. Overall, NLP is a highly dynamic and rapidly evolving field that has the potential to revolutionize the way we interact with technology. With the continued advancements in machine learning techniques, it is likely that NLP will continue to play a significant role in shaping the future of human-computer interaction.

C. Naïve Bayes

The Naive Bayes Classifier algorithm is a statistical method that is based on Bayes' theorem and is commonly used in machine learning problems such as text classification and spam filtering. The algorithm makes the assumption that the predictors in the dataset are independent of each other, which is often referred to as the "naive" assumption [5]. Despite this assumption, the Naive Bayes Classifier algorithm has been shown to be highly effective in many real-world applications.

One such application is in the detection of fake news on social media sites. The Naive Bayes Classifier algorithm can be used to classify news articles as either real or fake based on a variety of features such as the text of the article, the source of the article, and the number of shares and likes the article receives. By training the algorithm on a large dataset of known real and fake news articles, it can be used to accurately identify and flag potentially fake news articles on social media platforms.

Another application of Naive Bayes Classifier is in the field of natural language processing where it can be used to identify text as spam or ham. The algorithm can be trained on a dataset of known spam and ham messages, and then used to classify new incoming messages as either spam or ham. It can also be used in sentiment analysis to classify text as positive, negative, or neutral.

Overall, the Naive Bayes Classifier algorithm is a powerful tool that can be used in a wide range of machine learning problems. Its simplicity and effectiveness make it a popular choice for many practitioners and researchers.

D. Support Vector Machine

A well-liked machine learning method for classification and regression applications is the Support Vector Machine (SVM). It is a technique for determining the best border (hyperplane) that categorizes the data points while maximizing the margin—the distance between the nearest data points in each class and the boundary[6]. The SVM algorithm works by plotting the data points in an dimensional space, where n is the number of features. Each data point is represented by a coordinate that corresponds to the values of each feature. The algorithm then tries to find the most suitable hyperplane that separates the data points into different classes. The data points closest to the boundary are known as support vectors, and they have a crucial role in determining the location of the hyperplane. In addition to classification, SVM can also be used for regression problems, in which case the goal is to find a function that best maps the input data to the output variable[7].

Overall, the Support Vector Machine algorithm is a powerful tool that can be used for a wide range of machine learning tasks, and its ability to work well with non-linearly separable data make it a popular choice among practitioners and researchers.

III. METHODOLOGY

This paper presents the use of NLP techniques for identifying misleading news stories, commonly known as fake news. This section provides an overview of the architecture of the model. It includes information about the dataset used for training and evaluation.

The dataset is collected from Kaggle, followed by pre-processing, feature extraction, training of the model, and finally evaluation. Figure 1 describes the system methodology.



Figure 1: Methodology Process

A. Datas et

The project's dataset came from <https://www.kaggle.com/c/fake-news/train.csv>, from where it was downloaded. a complete training set having the following characteristics:

- Id: unique id.
- Title: the title of a news article.
- Text: Content in the article.
- Label: a label that marks the article as fake.

B. Data Pre-Processing

The data on social media is extremely amorphous; the bulk of it is casual announcements with typos, slang, and poor grammar. For better inference, it is essential to clean the data before it can be used for extrapolative modelling. So, fundamental pre-processing was completed on the training data. Data Pre-processing comprises the following steps:

- Conversion to Lower case: To avoid using the same words more than once, the text was first converted to lower case. For instance, "Quick" and "quick" are both full of dissimilar terms when determining the word count.
- Removal of Punctuations: Removal of Punctuation mark will help the model to treat all the text equally. For eg. the word Hi and Hi! will be treated equally after the process of punctuation removal. Punctuation typically don't convey much meaning on its own.
- Stop-words removal: The majority of frequently occurring words that are insignificant are known as stop-words. These include a, the, of, on, at, etc. as examples. They frequently categorise a text based on its structure rather than its context. It would lead to a deficient presentation if viewed as a feature. Stop words are therefore not included in the training data as it helps the model focus more on important information and not on low-level information like stop-words.

- **Lemmatization:** It breaks down words into their underlying word forms. It does morphological analysis to identify the root word with the aid of a lexicon. Lemmatization was used in this study to raise the quality of the frequency-based vector values.
- **Handling missing or duplicate values:** When missing or duplicate values are found, the next step is to choose whether to drop them or impute them. While imputing values may inject bias into the dataset, dropping missing values could result in the loss of crucial data. Duplicate values should be eliminated to prevent the model from being trained on the same data more than once. You can achieve this by eliminating duplicates or by preserving only the first instance of a duplicate value.

C. Feature Extraction

Feature Extraction is the process of selecting and extracting relevant information from text data that can be used as input to a machine learning model. The goal of feature extraction is to extract the most informative and relevant features from the text that can help the model make accurate predictions. Feature Extraction steps are:

- **Count Vector:** It is a feature extraction technique to represent the frequency of words in a text. It is a simple method of converting text data into numerical feature vectors. This is done by counting the number of occurrences of each word in the text and then representing each word as a dimension in the feature vector. The count of the word in the text is the value of that dimension. The resulting vector is a sparse vector.
- **Term Frequency-Inverse Document Frequency(TFIDF):** It is a feature extraction approach used to depict a word's significance within a document. To calculate the term frequency of a word in a corpus of documents, the term frequency of a word in a document is multiplied by the word's inverse document frequency. As a result, terms that are used frequently in one text but not frequently across the corpus are given more weight.
- **Word Embeddings:** The representation of words as high-dimensional vectors that accurately convey their meaning is a feature extraction technique. Word embeddings are designed to map vocabulary words to a dense vector space, where each dimension of the vector corresponds to a distinct semantic or syntactic aspect of the word. Words with comparable meanings are clustered together in the resulting vector space, capturing the similarity between words.

D. Algorithms

- **Bidirectional Encoder Representations from Transformers (BERT):** It is designed to understand the context of a sentence by analyzing the relationship between words in a sentence, rather than just their individual meanings. BERT has been trained on a large corpus of text and has the ability to understand the meaning of words in context. This allows it to perform well on a wide range of natural language processing tasks such as text classification, question answering, and language translation [8]. In addition to its pre-training and fine-tuning capabilities, BERT's architecture is designed to handle long sequences of text, making it well-suited for handling news articles. It also has the ability to handle out-of-vocabulary words, which is important for handling new and emerging words or phrases in the context of fake news detection. Overall, BERT's pre-training, fine-tuning capabilities, and ability to handle long sequences of text and out-of-vocabulary words make it a powerful tool for fake news detection and a wide range of other NLP tasks.
- **Deeply Bidirectional Encoder Representations from Transformer-based Pre-training (DeBERTa):** DeBERTa's architecture is similar to BERT, but with some modifications such as a longer sequence length, more layers, and more attention heads. One of the features of DeBERTa is the use of dynamic masking. This technique adaptively masks different tokens for each input, which results in a more robust language model. This allows DeBERTa to better handle text data with a high degree of variability, such as social media posts or informal text. DeBERTa also uses a technique called "span based masked language model" which helps to better understand the relationship between words in a sentence, which in turn helps it to perform better on tasks such as coreference resolution, where understanding the relationships between different words is crucial.
- **GRU-Attention Network:** Gated recurrent unit (GRU) is a type of recurrent neural network (RNN) architecture that was designed to address the vanishing gradient problem of traditional RNNs. GRU uses gates to control the flow of information through the network and can selectively remember or forget past information based on the current input. This allows GRU to capture long-term dependencies in the input sequence and has made it a popular choice for applications such as natural language processing, speech recognition, and music generation. The GRU architecture consists of a set of units that maintain an internal state. These units are similar to the cells in a long short-term memory (LSTM) network but with fewer parameters. Each unit has two gates: an update gate and a reset gate. The update gate controls how much of the previous state should be retained and how much of the new input should be used to update the state. The reset gate determines how

much of the previous state should be forgotten. By selectively updating and forgetting past information, the GRU can capture long-term dependencies in the input sequence while avoiding the vanishing gradient problem.

E. Evaluation

The dataset after pre-processing is split into a training and validation dataset in the ratio of 80:20, respectively. Measuring the performance of machine learning algorithms is an important task to ensure that they are working correctly and achieving the desired results. One of the most commonly used metrics for evaluating the performance of classification algorithms is classification accuracy.

Classification accuracy is the capacity of a model or system to correctly identify the true class of an observation. It is described as the percentage of examples in a dataset that were correctly categorized. This indicates that it is determined by dividing the total number of occurrences in the dataset by the number of cases that were correctly categorized.

A model or system that can accurately identify the true class of an observation with a high frequency is said to be functioning well. This is indicated by a higher classification accuracy. Generally speaking, the model is thought to function better the higher the classification accuracy. Although accuracy is often the greatest indicator for assessing a model's performance, additional metrics like precision, recall, and F1-score should also be taken into account.

IV. RESULT

Five classifiers have been implemented, and their performance on how well they were able to classify the given news article is compared. For the comparison, we have used the accuracy metric.

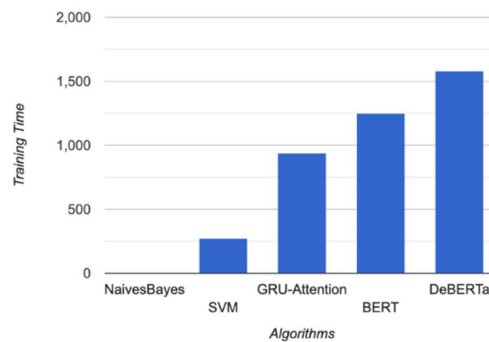
The following are the four algorithms that are used to detect fake news:

- Naive Bayes
- Support Vector Machine
- GRU-Attention Network
- BERT
- DeBERTa

TABLE I: COMPARISON OF ACCURACY SCORE

Algorithms	Accuracy Score
Naive Bayes	91.34%
Support Vector Machine	89.57%
GRU-Attention Network	92.43%
BERT	99.5%
DeBERTa	99.9%

TABLE II: TRAINING TIME FOR ALL ALGORITHMS



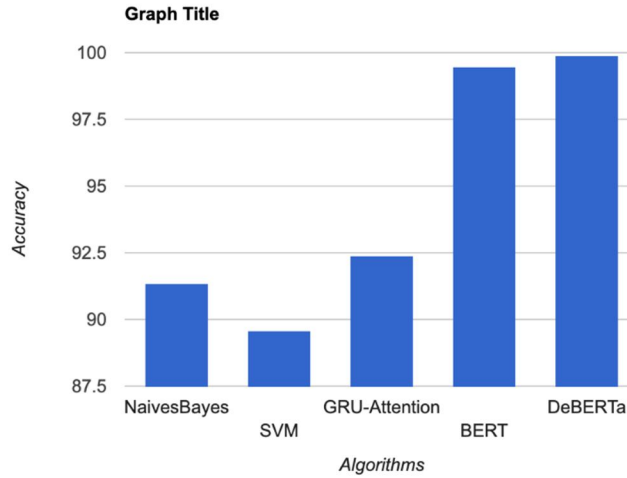


Figure 3: Accuracy results of all Algorithms

Figure 2 expresses the training time for all the algorithms. We can see that Naive Bayes requires the least amount of training time (about 3.7s), followed by SVM (276s). Naive Bayes takes 1/100th of the processing time taken by SVM because Naive Bayes assumes that features are independent of each other, thus it does not have to learn all the possible correlations between the features, whereas SVM looks for correlation between the features. The DeBERTa model takes the most training time of 1581s.

Figure 3 expresses the accuracy score for all the algorithms. We can see that DeBERTa model has the highest accuracy of 99.9%, followed by BERT model with 99.5%. SVM has the least accuracy of 89.57%.

V. RESULT

The objective of this research is to address the challenging task of classifying news articles, as the input data is in text format and news can have a wide range of features that must be taken into account. This paper presents a novelty model, the GRU-Attention Network, for identifying fake news on social media. The first stage focuses on defining the key concepts and characteristics of fake news, while the second stage reviews the comparison of the novelty model with existing methods for detecting fake news using various supervised learning algorithms. We propose a model with an accuracy of 92.43%. There are several potential hypotheses that can be made as to why a model may perform well on one dataset but not as well on another. The model may be overfitting to the training dataset and not generalising effectively to new data, which is one scenario that could apply. This indicates that the model has discovered patterns in the training data that only apply to that dataset and not to others. Another potential reason is that the features used in the model may not be relevant or informative for the new dataset. The model may be using features that are specific to the training dataset and are not present in the new dataset, causing the model to perform poorly.

A combination of factors such as overfitting, irrelevant features, covariate shift, and different styles of writing may contribute to a model performing well on one dataset but not as well on another. These variables should be taken into account when assessing a model's performance and when developing new models.

With huge computational power required for the attention-based models to be fine tuned, which will be costly and does not provide much improvement over traditional machine learning algorithms like Naive Bayes and SVM, which require much less computational power compared to attention-based models, and execution times for these models are also much less than attention-based models,

In the future, we plan to continue this research by experimenting with multi-model-based approaches and We plan to develop a hybrid approach for detecting fake news by incorporating information from the content, context, and structure of the news articles. This approach will make use of graph-based techniques to analyse the relationships between different entities in the news. We'll also look into how varying kernel sizes for parallel channel-based deep neural networks might help with performance. This approach will be able to better detect patterns and features that are indicative of fake news, as well as use visual information-based analysis for real-time detection in videos and images. Furthermore, we aim to incorporate domain experts' knowledge and fact-checking techniques for more accurate classification.

REFERENCES

- [1] Ramya, S. P., and R. Eswari. "Attention-based deep learning models for detection of fake news in social networks," IJCINI vol.15, no.4, pp.1-25,2021.
- [2] Mu-Yen Chen, Yi-Wei Lai, and Jiunn-Woei Lian., "Using deep learning models to detect fake news about covid-19". ACM Transactions on Internet Technology, vol. 0,May 2022.
- [3] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation identification," in Proc. 26th Int. Joint Conf. Artif. Intell., Aug. 2017, pp. 3901–3907.
- [4] Galli, A., Masciari, E., Moscato, V. et al. "A comprehensive Benchmark for fake news detection," J Intell Inf Syst, vol. 59, pp.237–261,2022.
- [5] Zolnierrek, A., Rubacha, B. "The empirical study of the naive bayes classifier in the case of markov chain recognition task,"Advances in Soft Computing, Springer, vol 30, pp. 329-335,2005
- [6] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998.
- [7] Zhang, Y,"Support vector machine classification algorithm and its application,"Information Computing and Applications, Springer, vol 308,pp.179-188, 2012
- [8] Devlin, J., Chang, M., Lee, K., & Toutanova, K., "Bert: pre-training of deep bidirectional transformers for language understanding," North American Chapter of the Association for Computational Linguistics: Human Language Technologies,vol. 1,pp. 4171–4186,2019