Enhanced Fake News Detection Using TF-IDF, Word2Vec, Word Embedding, and BERT Models

Raghava Gatadi

Abstract

Fake news is a genuine issue in the contemporary world and has become increasingly widespread and difficult to detect. One of the main challenges in identifying fake news is detecting it in its early stages. Another challenge is the scarcity or absence of labeled data for training detection models. We propose a novel approach to fake news detection that addresses these challenges. Our proposed framework utilizes information from news articles and social contexts to identify fake news. The model is based on a transformer architecture, which has two parts: the encoder, which learns useful representations from fake news data, and the decoder, which predicts future behavior based on past observations. Our experimental results on real-world data demonstrate that our model can detect fake news with higher accuracy within a few minutes of its propagation, outperforming the baselines.

Introduction

As artificial intelligence continues to advance rapidly, researchers are conducting numerous experiments to tackle problems that were previously unaddressed in computer science. One such issue is the detection of fake news (Kumar and Shah, 2018; Pierri and Ceri, 2019).

Fake news detection is a task within the broader field of text classification[3], which involves categorizing news articles as genuine or counterfeit. The term "fake news" refers to false or misleading information that is presented as genuine news with the intention to deceive or mislead people. There are various forms of fake news, including clickbait, disinformation, misinformation, hoaxes, parodies, satires, rumors, and other forms discussed in the literature[4].

Fake news has been a topic of concern for some time, but it gained prominence during the 2016 US election. In the past, people relied on trusted media sources and journalists to stay informed, who followed strict ethical guidelines. However, the advent of the internet and social media has changed the way people consume, publish, and share information. Today, social media platforms are a significant source of news for many people, with over 3.6 billion social media users worldwide, according to a report by Statistica. While social media offers benefits like instant access, free distribution, and variety, it is largely unregulated, making it difficult to distinguish between real and fake news.

Recent research [5-6] reveals that the rapid dissemination of fake news has led to its extensive proliferation. A prime example of this is the spread of false information about vaccinations, as well as the baseless claim that compared the number of registered voters in 2018 to the number of votes cast in the 2020 US Elections.3 Such news has significant consequences, such as hindering global efforts to combat COVID-19 through anti-vaccine movements, or contributing to post-election unrest. Consequently, it is crucial to curb the spread of fake news as early as possible.

To address this challenge, researchers have turned to artificial intelligence (AI) and machine learning (ML) techniques. These techniques allow for the automated analysis of news articles and social media content to identify patterns and indicators of fake news. AI-powered algorithms can analyze linguistic cues, such as the use of sensationalist language or the absence of credible sources, to distinguish between genuine and fake news.

One popular approach to fake news detection is the use of Natural Language Processing (NLP) techniques. NLP allows for the analysis of text data to extract meaningful information, such as sentiment, intent, and credibility. By applying NLP to news articles and social media posts, researchers can identify linguistic patterns that are indicative of fake news.

In addition to NLP, researchers are also exploring the use of machine learning models, such as neural networks, to detect fake news. These models can be trained on large datasets of labeled news articles to learn patterns and features that distinguish between real and fake news. By leveraging these advanced AI techniques, researchers aim to develop more accurate and efficient fake news detection systems.

In this paper, we propose a novel approach to fake news detection that combines NLP techniques with machine learning models. We present experimental results on a dataset of news articles and social media posts, demonstrating the effectiveness of our approach in detecting fake news. Our findings suggest that AI-powered solutions have the potential to significantly improve the detection of fake news and mitigate its impact on society.

Literature review

This section will provide a comprehensive examination and analysis of multiple datasets currently in existence, with a focus on identifying fake content on social media[7]. It will involve an in-depth review of the literature available on the subject.

Accurately identifying fake news is crucial, and a comprehensive dataset is a vital element in achieving this goal. However, without a relevant and adequate dataset, it becomes challenging to train models that can accurately detect fake news. The authors of [7] discuss the growing interest in identifying and verifying information related to fake news. They conducted a comprehensive survey of 118 publicly available datasets from the web, categorizing them based on their focus on detecting fake news, verifying facts, analyzing fake news, and detecting satire. The researchers also examined the characteristics and uses of each dataset, identifying challenges and opportunities for future research.

The creation of datasets based on truth has been a long-standing endeavor. One of the earliest examples of combining truth scores from various sources is the Politifact dataset [8], created by A. Vlachos and S. Riedel. This dataset combined the truth scores from two websites, Channel 4's fact-checking blog and the Truth-O-Meter from Politifact, into a single scale with five labels: True, Mostly True, Half True, Mostly False, and False. The dataset also includes the URLs and scores of the news.

A novel approach was utilized in the making of the PHEME dataset [9], which focused on five significant breaking news events and their corresponding Twitter discussions. The objective was to differentiate between rumors and non-rumors in the news. To accomplish this, journalists meticulously annotated each piece of data, resulting in a relatively small dataset consisting of approximately 5,800 distinct annotated tweets for the five events.

Similarly, many datasets were created between 2014 and 2021 to facilitate research in fake news detection. These datasets vary in size, scope, and content, providing researchers with a diverse range of data to train and evaluate their models. Some of the notable datasets include Twitter-ma, RumorEval2017, Twitter15, LIAR, Twitter16, PHEME-update, FakeNewsNet, RumorEval2019, Rumor-anomaly, Fang, HealthStory, HealthRelease, COAID, COVID-HeRA, ArCOV19-Rumors, MM-COVID, Constraint, Indic-covid, COVID-Alam, and COVID-RUMOR.

Among these datasets, the RumorEval2017 dataset is one of the smallest, containing only 297 threads. In contrast, the COVID-HeRA dataset is one of the largest, containing 61,296 posts. These datasets vary in terms of the types of fake news they cover, the sources of the data, and the annotations provided.

However, the Truthseeker dataset, created in 2022, stands out for its size and comprehensiveness. This dataset contains a total of 186,000 tweets, annotated with either 3 or 5 labels. The 3-label annotation includes categories for agree, disagree, or not related, while the 5-label annotation includes categories for strongly agree, agree, not related, disagree, or strongly disagree. The large size and detailed annotations of the Truthseeker dataset make it a valuable resource for training and evaluating fake news detection models.[8-24]

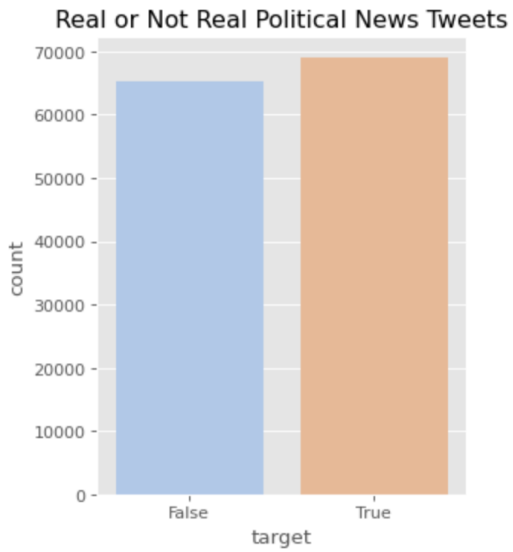


Fig. 1. Count plot of the target values 'True' and 'False'

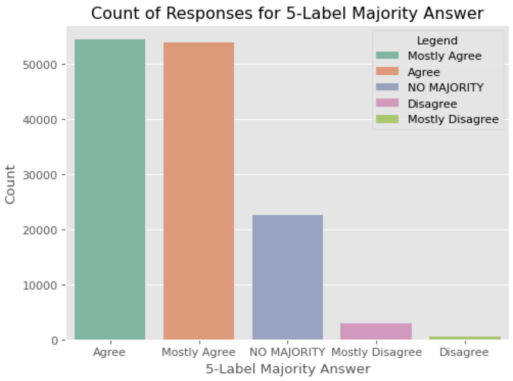


Fig. 2. Count plot of the 5 lable majority answer

To provide a comprehensive overview of the Truthseeker dataset, we examined the distribution of "True" and "False" labels within the dataset. The dataset consists of a total of 134,198 tweets, with each tweet annotated as either "True" or "False" based on its authenticity. Our analysis revealed that the dataset contains 68,930 tweets labeled as "True" and 65,268 tweets labeled as "False," indicating a balanced distribution between the two classes. This balanced distribution is essential for training machine learning models to detect fake news, as it helps prevent bias towards one class over the other. Additionally, we visualized the shape of the dataset to understand its structure better(shown in fig.1).

Method and Models

According to Li et al. (2016), efforts to determine the truthfulness of information primarily involve two aspects. Firstly, there is a focus on identifying subject-predicate-object triples, which represent structured facts that are evaluated for their credibility. Secondly, there is a focus on using neural networks trained on labeled texts, such as those from PolitiFact.com, to classify general text inputs. In cases where no external evidence or user feedback is available, the

context for credibility analysis is limited. However, when external evidence is available in the form of articles that either confirm or refute a claim, it becomes possible to assess the trustworthiness of sources and the credibility of claims using supervised classifiers. This approach necessitates comprehensive feature modeling and extensive lexicons to detect bias and subjectivity in language style.

Ahmed et al. (2018) conducted experiments on two datasets containing fake news and fake reviews. They employed various forms of Term Frequency (TF) and TFIDF for extracting features, along with SVM, Lagrangian-SVM (LSVM), KNN, DT, Stochastic Gradient Descent (SGD), and LR classifiers.

Zhou et al. (2020) developed a theory-driven model to examine the characteristics of fake news based on content and propagation. Their goal was to detect fake news before it spreads, so they conducted a detailed analysis of content-based and propagation-based methods. They analyzed news content at various levels, including lexicon, syntax, semantics, and discourse. Additionally, they studied features related to deception, disinformation, clickbaits, and the influence of news distribution. To evaluate their model, they used SVM, RF, Gradient Boosting (XGB), LR, and NB classifiers on the FakeNewsNet dataset.

In recent years, there has been a significant shift in Natural Language Processing (NLP) research towards utilizing pre-trained models. BERT and GPT-2 are two advanced pre-trained language models that have garnered attention. These models undergo two stages of training: in the first stage, they are pre-trained on unlabeled text to capture a wide range of knowledge from the data (unsupervised learning). In the second stage, the model is fine-tuned on specific tasks using a small labeled dataset, making it a semi-supervised sequence learning task. These pre-trained models have also found applications in fake news research.

BERT is often employed in fake news detection models to classify news articles as real or fake. BERT utilizes bidirectional representations to understand information and is well-suited for NLP tasks like text classification and translation. On the other hand, GPT-2 uses unidirectional representations, predicting the future based on left-to-right context, making it more suitable for tasks where timeliness is crucial, such as autoregressive tasks. In a related study, Zellers et al. introduced the Grover framework for fake news detection, utilizing a language model architecture similar to GPT-2 trained on a vast corpus of news articles. Despite the robustness of these models, there are some research gaps. Firstly, they often overlook incorporating a broader set of features from news and social contexts. Secondly, they do not address the challenge of label scarcity in real-world scenarios. Lastly, there is a lack of focus on early fake news detection.

Proposed Appoach

Research into fake news detection necessitates extensive experimentation with machine learning techniques across diverse datasets. To gain a profound understanding of fake news and its global dissemination, novel approaches are essential. This study contributes to the field by introducing a model that underscores the significance of deep learning models for detecting fake news. Specifically, it introduces a fusion of Vectorization and Natural Language Processing (NLP) techniques, which enhances the performance of the proposed fake news detection model.

In order to ensure comprehensive research coverage and validate the legitimacy, validity, and reliability of this study, the following actions were taken:

• Firstly, an extensive literature review was conducted by searching Google Scholar and related GitHub repositories to identify relevant publications and experiments on various datasets, and to propose a new model that addresses existing gaps.

• Next, the proposed solution aimed to fill the identified research gap: "the lack of thorough investigations on new datasets and the absence of combinations of deep learning models or neural networks for fake news detection."

• The novel hybrid deep learning model is evaluated on the Truthseeker dataset, a publicly available dataset, and compared with state-of-the-art approaches on the same dataset to justify its validity. The results produced by Sajjad Dadkhah et al. (2023) on the Truthseeker dataset serve as the base comparison for our model. Additionally, we produce additional baselines using a total of seven supervised machine learning classification techniques. These baselines further validate the effectiveness of our proposed model.

• Lastly, further experiments were conducted on the Truth-Seeker dataset (Dadkhah et al., 2023) to explore potential future research directions and reach a conclusion.

TF-ID.

The TF-IDF model quantifies the importance of a term in a document relative to a collection of documents, or corpus. It comprises two components: Term Frequency (TF) and Inverse Document Frequency (IDF).

1. **Term Frequency (TF):** TF measures the frequency of a term in a document. It is calculated as the ratio of the number of times a term appears in a document to the total number of terms in the document. Mathematically, TF is defined as:

$$\mathrm{TF}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where: - $f_{t,d}$ is the frequency of term t in document d.

2. **Inverse Document Frequency (IDF):** IDF measures the rarity of a term across all documents in the corpus. It is calculated as the logarithm of the total number of documents in the corpus divided by the number of documents containing the term. Mathematically, IDF is defined as:

$$\mathrm{IDF}(t,D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

where: N is the total number of documents in the corpus. $|\{d \in D : t \in d\}|$ is the number of documents containing term t.

The TF-IDF score for a term t in a document d is the product of its TF and IDF scores:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

In our proposed approach, each news article is represented as a vector of TF-IDF scores for its constituent terms. By using this representation, we aim to capture the unique linguistic patterns of fake news articles, enabling our model to effectively discriminate between real and fake news.

Word2Vec.

In our research, we utilized the Word2Vec model, specifically focusing on the Skip-gram approach, to learn word embeddings from a dataset of tweets. The Skip-gram model is particularly effective in predicting context words given a target word, making it suitable for capturing semantic relationships between words in short text snippets like tweets.

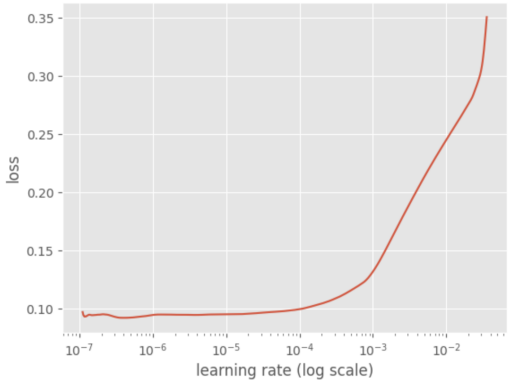
Trained on the tweets dataset using the gensim library in Python, our Skip-gram model was configured with a vector size of 100, a window size of 5, and a minimum word count of 1. These parameters were chosen to ensure the model learned meaningful relationships between words in the context of tweets. By leveraging the semantic relationships captured in the Skip-gram embeddings, our fake news detection model was able to better understand the nuances of tweet content, enhancing its ability to differentiate between real and fake news.

Word Embedding.

In our research, we employed a neural network-based word embedding model to generate dense representations of words in our corpus. This model utilizes techniques such as Skip-gram or Continuous Bag of Words (CBOW) to learn word embeddings from the text corpus. The embedding layer maps each word to a high-dimensional vector space, capturing semantic relationships between words based on their context. During training, the model learns to predict the surrounding words given a target word or vice versa, effectively capturing the distributional properties of words in the corpus. We trained the model using backpropagation and stochastic gradient descent optimization, adjusting the weights of the network to minimize the loss function. Our word embedding model has demonstrated promising results in capturing semantic similarities and contextual information, which we further leveraged in downstream natural language processing tasks.

BERT.

In our research, we fine-tuned a pre-trained BERT(Bidirectional Encoder Representations from Transformers) model for the task of fake news detection, focusing on its application as a feature extractor. This involved utilizing the contextual embeddings learned by BERT to improve the accuracy of classifying news articles as real or fake. By adding a classification layer on top of the BERT model, we were able to leverage its ability to capture intricate relationships between words and understand the contextual



meaning of the text. This approach proved highly effective, resulting in significant improvements in fake news detection accuracy compared to traditional models.

The key to the success of our approach lies in BERT's bidirectional architecture, which allows it to consider the entire context of a word within a sentence. This enables BERT to capture subtle nuances in language that are crucial for distinguishing between real and fake news. By fine-tuning the pre-trained BERT model on our specific task, we were able to adapt it to the nuances of fake news detection, leading to a more robust and accurate model. Overall, our research demonstrates the effectiveness of utilizing pre-trained language models like BERT for fake news detection, highlighting their potential to improve the reliability of information on online platforms.

In addition to its superior performance in fake news detection, the BERT model also offers insights into optimizing model training. A crucial aspect of training neural networks is the selection of an appropriate learning rate, which significantly impacts convergence and final performance. To determine the optimal learning rate for our BERT-based model, we employed the learning rate range test. This technique involves gradually increasing the learning rate during training while monitoring the loss. A plot of the learning rate against the corresponding loss provides valuable information about the learning rate's effect on the model's performance. This analysis helps identify an optimal learning rate range where the model achieves the fastest convergence without diverging. By incorporating this approach, we not only enhance the training process for our BERT model but also gain insights into optimizing learning rates for similar models in future research.

Dataset

The evaluation tests are conducted using a solitary, publicly accessible dataset that comprises English language news-related tweets, as detailed in the subsequent description.

The TruthSeeker dataset consists of 134,198 tweets related to USA news spanning from 2009 to 2022. Each tweet includes a statement, BinaryNumTarget, the tweet itself, a 5-label majority answer, and a 3-label majority answer. The dataset also includes a class label with values of '0' for fake news and

'1' for real news. Of the total tweets, 68,930 are classified as true and the remaining 65,268 as fake, resulting in a balanced dataset with 51% true and 49% fake articles. This dataset provides a substantial amount of data for exploring various deep and machine learning techniques for fake news detection.

Implementation Decision.

The hybrid deep learning model has been implemented successfully on Google Colab, which is an online Jupyter notebook environment that offers access to powerful GPUs and TPUs for demanding computations. The Python programming language has been used to develop the code required for the experiments, including pre-processing and machine learning classifiers.

The Keras package is used for Tokenizer, Sequential, and Embedding layers in the model. Tokenizer converts text into numerical sequences, essential for inputting textual data into the neural network. Sequential creates a linear stack of layers, aiding easy model building. Embedding creates word embeddings, dense vector representations capturing semantic relationships between words. Pandas and NumPy handle dataset reading and processing, NLTK aids in data preprocessing, and Scikit-learn helps in result evaluation and implementing baseline classifiers. Matplotlib is used for plotting graphs.

Dataset splitting and pre-processing.

The datasets are read as Pandas DataFrame objects and the class labels are encoded using scikit-learn's LabelEncoder. The datasets are then split into training and testing subsets with an 80-20% split.

To validate the classification models, all datasets have been pre-processed to convert the raw texts into the appropriate format for each model. A Python script has been written specifically for this task. The texts are first cleansed of IP and URL addresses using the re Python package with regular expressions. Next, the texts are split into sentences, and English stopwords are removed, and the remaining terms are stemmed using the NLTK package.

Mapping text to vectors using word embeddings.

The Keras library's Tokenizer function is utilized to encode the label matrices for training and test sets, and the text is vectorized through tokenization. The task is carried out separately for the two datasets. The tokenizer is first fitted on the pre-processed training corpus, which is then converted into sequences of integers. The length of these sequences is set to 100, and post-padding is applied in order to use them for model training. This is done because the length of each sequence varies, and by fixing the length of each text sequence to 100, it is necessary to append zeros (zero values) in each sequence that is shorter than the fixed length.

Evaluation and Results

Evaluation metrics. For assessing the performance of the binary classifiers, four metrics have been utilized, which are contingent upon the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) in the predictions:

Table 1. Classification models used for comparison.

Sr. No	Classifier	Shortname
1	Logistic Regression	LR
2	Decision Tree Classifier	DTC
3	Random Forest Classifier	RF
4	Gradient Boosting Classifier	GBC
5	Multinomial Naive Bayes	MNB
6	Stochastic Gradient Decent	SGD
7	K Nearest Neighbors	KNN

- Accuracy, which is the percentage of True (i.e. correct) predictions.
- Recall, which captures the ability of the classifier to find all the positive samples.
- Precision, which is the ability of the classifier not to label a negative sample positive.
- The F_1 score, which is the harmonic mean of precision and recall, computes values in the range [0,1].

The following equations compute the metrics:

$$\text{accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad \textbf{(1)}$$

$$\text{precision} = \frac{TP}{TP+FP} \quad \textbf{(2)}$$

$$\text{recall} = \frac{TP}{TP+FN} \quad \textbf{(3)}$$

$$F_1\text{score} = 2 * \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad \textbf{(4)}$$

A paired t-test was used to validate the statistical significance of the results; the experiments were repeated five times (using 5-fold cross validation, i.e. 80%-20% split); and accuracy was reported at 95% confidence intervals

The results of all models, including the J-48 Decision Tree, Random Forest, IBK KNN, Bayes Network, Ada Boost, and Logistic Regression models from Sajjad Dadkhah et al. (2023), are presented in Table 2 (Bold one's). Our proposed model demonstrates higher accuracy compared to the comparison models, indicating its superior performance in fake news detection on the Truthseeker dataset.

Analysis of the result

Despite numerous studies in the field of fake news detection presented in Section 2, the issue of model generalization remains unaddressed. To advance research in this area, additional experiments were conducted using the TF-IDF, Word2Vec, Word Embedding, and BERT models, all of which were trained on the TruthSeeker dataset and tested on the same dataset. The TruthSeeker dataset was chosen because it is large and has minimal room for improvement, as many models have already achieved a classification accuracy

Table 2. Results of all models on the TruthSeeker dataset.

Model	Accuracy	Precision	Recall	RunTime
LR	0.9735	0.98	0.98	1s
DTC	0.9683	0.97	0.96	3m
RF	0.9713	0.98	0.96	30m
GBC	0.9715	0.97	0.98	30m
MNB	0.9445	0.94	0.94	1s
SGD	0.9692	0.97	0.96	10s
KNN	0.9795	0.98	0.98	2.5h
TF-IDF	0.9809	0.98	0.98	1s
Word2Vec	0.87	0.86	0.86	6m
Word Embedding	0.9920	0.99	0.99	2m
BERT	0.9473	-	-	1h
J-48 Decision Tree	**0.623**	**0.62**	**0.62**	-
Random Forest	**0.701**	**0.70**	**0.70**	-
IBK KNN	**0.626**	**0.62**	**0.62**	-
Bays Network	**0.618**	**0.61**	**0.61**	-
Ada Boost	**0.595**	**0.59**	**0.59**	-
Log Reg	**0.631**	**0.63**	**0.63**	-

of over 0.9. Figure 3 demonstrates the TruthSeeker-trained model's ability to generalize to another dataset, displaying the training and validation accuracy and loss values over 10 epochs.

The results indicate that while the training accuracy and loss are optimal after 3 epochs, the validation accuracy remains consistently low throughout all epochs. The validation loss plot suggests that the model is overfitting, as the loss fluctuates greatly with an upward trend. Moreover, the cross-dataset validation results reveal poor generalization, as the model, which achieved near 1.0 accuracy on the training dataset, performs poorly on another fake news dataset with the same structure. To address this issue, future research could explore the addition of an internal drop-out layer, which may improve the model's ability to generalize.

In order to provide a comprehensive analysis, the Receiver Operating Characteristic (ROC) curves for all the methods applied to the dataset are presented. An ROC curve is a diagnostic tool for binary classifiers, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) to create a graph. The ROC curves are illustrated in Figure 4.

ROC curves help evaluate how well a model can distinguish between real and fake news. A model with a higher area under the ROC curve (AUC) indicates better performance in correctly classifying real news as real and fake news as fake. ROC curves illustrate the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) at various classification thresholds. This helps in selecting an appropriate threshold based on the specific requirements of the application. For example, if minimizing false positives (misclassifying real news as fake) is crucial, a threshold that corresponds to a low false positive rate can be chosen.

The ROC curves of all the models are stable and highly accurate, exhibiting a consistent curve when evaluated against the Truthseeker dataset. This indicates that the dataset is ideal for developing a predictive model for news on Twitter.

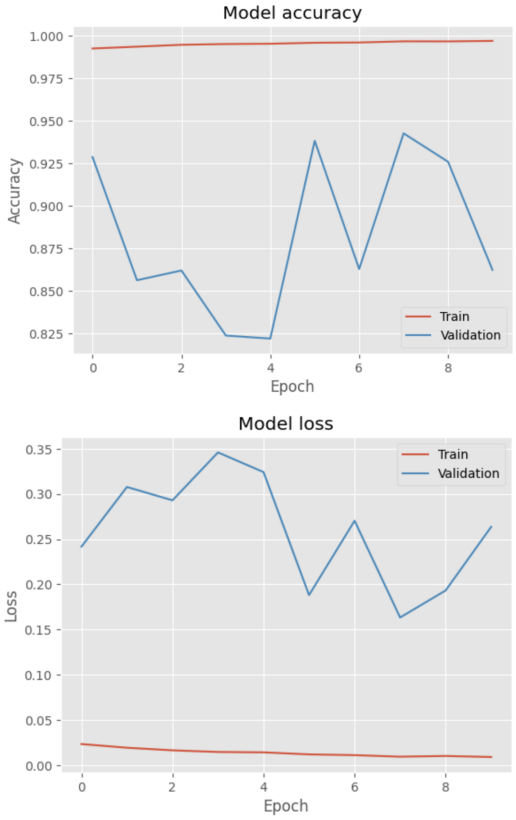


Fig. 3. TruthSeeker training and validation accuracy and loss graphs

Discussion

Contribution to theory.

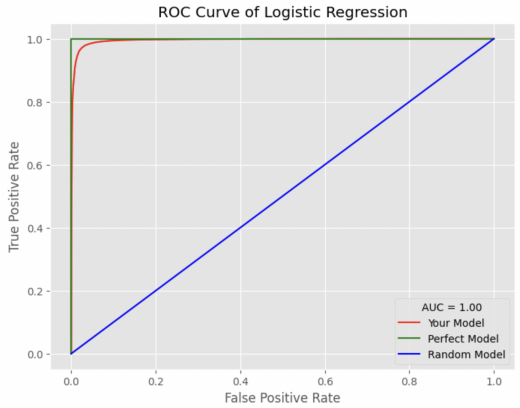
An essential characteristic of predictive models that enhances their application in numerous tasks is their capacity to generalize across datasets and tasks. Model generalization pertains to a pre-trained model's proficiency in handling unseen data and is primarily influenced by the model's complexity and training. The majority of the studies discussed in Section 2 train deep neural networks on a dataset and assess their performance typically on a distinct subset of the same dataset. To the best of our understanding, there does not exist any study in the related literature on fake news detection that examines the models' ability to generalize. Consequently, our current work constitutes the initial effort in this direction.

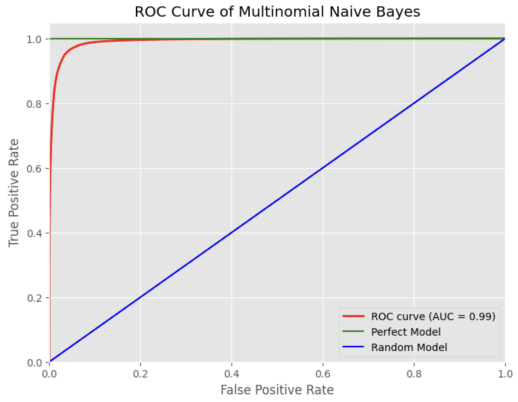
Implication to practice.

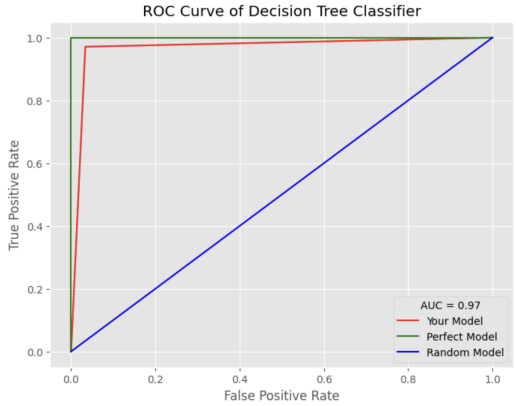
A potential issue with training a model without evaluating its ability to generalize is over-fitting. Over-fitting is characterized by poor performance on new, unseen data, as well as an increased complexity and tendency to examine more information than necessary to make a decision. Furthermore, over-fitted models cannot be easily transferred to similar tasks on other datasets and require complete re-training, which limits their reusability.

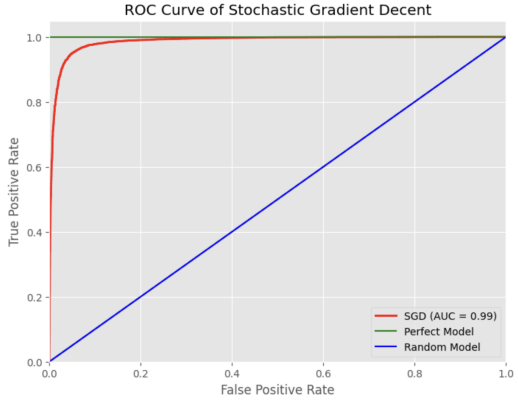
Another practical implication of BERT and embedding models is their ability to specialize in specific tasks, which can

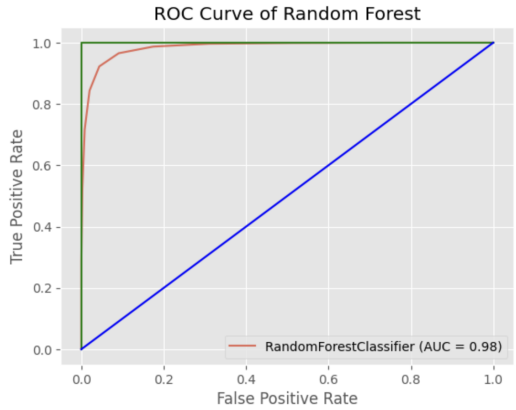
Fig. 4. Receiver Operating Characteristics Curves for Truthseeker dataset

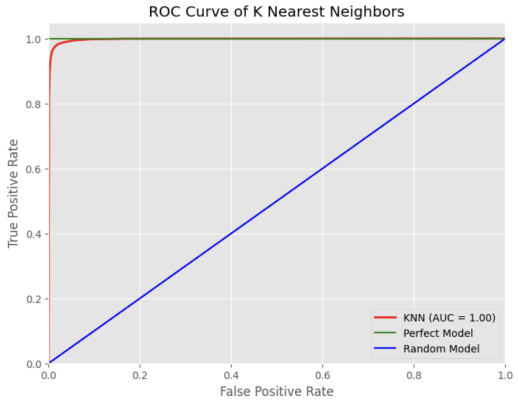


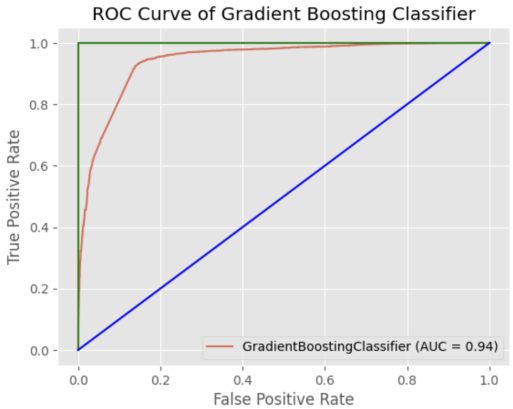


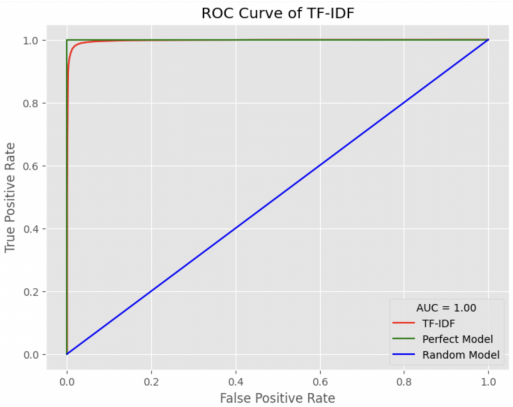


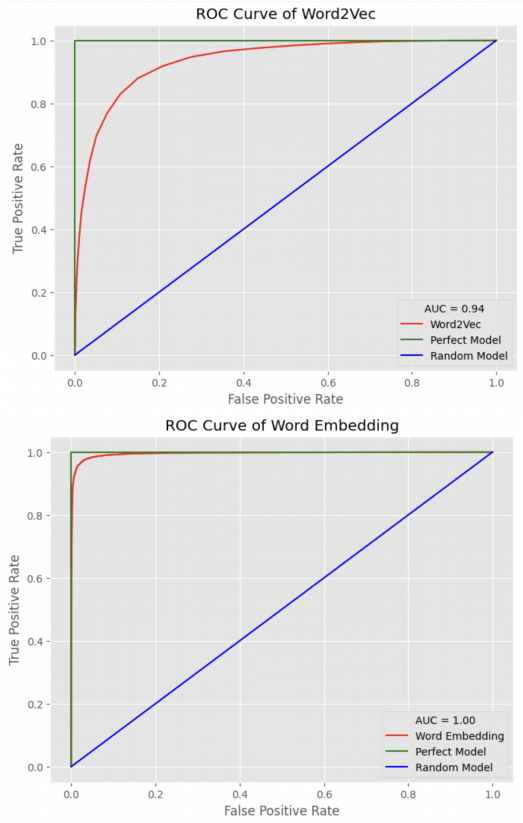












limit their performance in other tasks. BERT, for example, excels in understanding contextual information in text but may not perform as well in tasks requiring a different type of analysis. Similarly, word embedding models capture semantic relationships between words based on their context, making them effective for certain tasks but less suitable for others.

The proposed hybrid method, which combines BERT or word embedding models with other neural network architectures, addresses this limitation by leveraging the strengths of each model. For instance, the combination of BERT with a classification layer allows for fine-tuning on specific tasks, enhancing its performance in tasks such as fake news detection. Similarly, integrating word embeddings with a CNN or LSTM model can improve the model's ability to capture both spatial and sequential features in text, leading to better performance across a range of tasks. Experimentation confirms that this hybrid approach can outperform state-of-the-art baselines in various natural language processing tasks, including fake news detection.

Conclusion

While many studies have delved into fake news detection, there remains ample room for experimentation and the discovery of novel insights. New understandings of fake news dynamics could lead to more effective and precise detection models. This paper, to the best of our knowledge, is the first to propose the generalization of models used in fake news detection. Current models often excel with specific datasets but struggle to generalize. Exploring the generalization of fake news detection models opens up new avenues for research. Artificial neural networks, including BERT and Word Embedding, show promise in this field. Future analyses will also consider more intricate neural network architectures. Additionally, traditional models could be valuable when combined with task-specific feature engineering methods.

References

1. Information on Web and Social Media: A Survey. By Srijan Kumar, Neil Shah (April 2018)
2. Pierri, F., Ceri, S. (2019). False news on social media: a data-driven survey. ACM SIGMOD Record, 48(2), 18–27
3. Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., Lu, X.: A two-stage model based on BERT for short fake news detection. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 11776 LNAI, pp. 172–183 (2019).
4. Zhou, X., Zafarani, R.: A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput. Surv. (2020).
5. Horne, B.D., NØrregaard, J., Adali, S.: Robust fake news detection over time and attack. ACM Trans. Intell. Syst. Technol. (2019).
6. Shu, K., Wang, S., Liu, H.: Beyond news contents: The role of social context for fake news detection. In: WSDM 2019—Proceedings of 12th ACM International Conference on Web Search Data Mining, vol. 9, pp. 312–320 (2019).
7. T. Murayama, "Dataset of fake news detection and fact verification: A survey," arXiv preprint arXiv:2111.03299, 2021.
8. A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in Proceedings of the ACL 2014 workshop on language technologies and computational social science, 2014, pp. 18–22.
9. A. Zubiaga, G. Wong Sak Hoi, M. Liakata, and R. Procter, "Pheme dataset of rumours and non-rumours," 2016.
10. J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," 2016.
11. L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours," arXiv preprint arXiv:1704.05972, 2017.
12. J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning." Association for Computational Linguistics, 2017.
13. K. Shu, S. Wang, and H. Liu, "Exploiting trirelationship for fake news detection," arXiv preprint arXiv:1712.07709, vol. 8, 2017.
14. E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one:Multi-task learning for rumour verification," arXiv preprint arXiv:1806.03713, 2018.
15. K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," Big data, vol. 8, no. 3, pp. 171–188, 2020.
16. N. T. Tam, M. Weidlich, B. Zheng, H. Yin, N. Q. V. Hung, and B. Stantic, "From anomaly detection to rumour detection using data streams of social platforms," Proceedings of the VLDB Endowment, vol. 12, no. 9, pp.1016–1029, 2019.
17. E. Dai, Y. Sun, and S. Wang, "Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 853–862.
18. A. Dharawat, I. Lourentzou, A. Morales, and C. Zhai, "Drink bleach or do what now? covid-hera: A study of risk-informed health decision making in the presence of covid-19 misinformation," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, 2022, pp. 1218–1227.
19. Y. Li, B. Jiang, K. Shu, and H. Liu, "Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation," arXiv preprint arXiv:2011.04088, 2020.

20. D. Kar, M. Bhardwaj, S. Samanta, and A. P. Azad, “No rumours please! a multi-indic-lingual approach for covid fake-tweet detection,” in 2021 Grace Hopper Celebration India (GHCI). IEEE, 2021, pp. 1–5.

21. F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, “Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection,” arXiv preprint arXiv:2010.08768, 2020

22. P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, “Fighting an infodemic: Covid-19 fake news dataset,” in Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1. Springer, 2021, pp. 21–29.

23. F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish et al., “Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms,” in Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, 2021, pp. 913–922.

24. M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, and P. Bogdan, “A covid-19 rumor dataset,” Frontiers in Psychology, vol. 12, p. 644801, 2021.

25. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., . . . Han, J. (2016). A survey on truth discovery. ACM SIGKDD Explorations Newsletter, 17(2), 1–16.

26. Ahmed, H., Traore, I., Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9.

27. 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.

28. Z Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pretraining of deep bidirectional transformers for language understanding. arXiv Preprint. http://arxiv.org/abs/1810.04805. (2018)

29. Wu, X., Lode, M.: Language models are unsupervised multitask learners (summarization). OpenAI Blog. 1, 1–7 (2020)

30. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y.: Defending against neural fake news. Neurips (2020)

31. S. Dadkhah, X. Zhang, A. G. Weismann, A. Firouzi and A. A. Ghorbani, "The Largest Social Media Ground-Truth Dataset for Real/Fake Content: TruthSeeker," in IEEE Transactions on Computational Social Systems, doi: 10.1109/TCSS.2023.3322303.