Comprehensive Fake News Classification Leveraging LLM

Arjun Kapoor and Atharv Arya
Department of $SEAS$
University at Buffalo
Buffalo, NY 142603
{akapoor5;aarya2}@buffalo.edu

Abstract

Fake news detection is a critical issue in combating the spread of misinformation across digital platforms. This project integrates traditional machine learning, deep learning, and ensemble models into a unified framework to enhance accuracy and robustness. Using the WELFake dataset—a benchmark combining multiple smaller datasets—this approach aims to achieve better generalization and performance compared to existing methods, to improve fake news detection systems.

1 Introduction

Fake news detection has become a crucial area of research due to its significant societal impacts. The unchecked spread of misinformation, often through clickbait videos, social media platforms, online news outlets, and even mainstream media, serves various agendas such as financial gain, political manipulation, undermining competitors, and more. For example, misinformation about elections has influenced voter perceptions, while false health related news during COVID-19 pandemic jeopardized public safety. With the increased accessibility of online information-sharing platforms, the spread of false or misleading news has become widespread. The absence of accountability allows individuals to propagate content on any topic without bearing responsibility for its consequences, thereby exposing readers to the risk of manipulation by false information.

Addressing this challenge requires advanced technological solutions as it is not possible for humans to read all the content on social media platforms and control it. Advancements in Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) offer promising automated avenues to address this challenge. This project explores recent state-of-the-art methodologies for mitigating the spread of fake news, emphasizing on traditional machine learning algorithms, feature-based classifiers, deep learning algorithms, and the integration of these methods through ensemble models to enhance classification accuracy.

Unlike other approaches discussed in this review, our approach aims to integrate a broader spectrum of methodologies. While some studies explore traditional ML or feature-based classifiers and others focus on combining DL or transformer models, this project combines all these techniques into a unified ensemble framework. The papers discussed above use a variety of datasets some of which serve as benchmarks for Fake News Classification. These datasets include ISOT Fake News Dataset, LIAR, FakeNewsNet, and more from sources like Kaggle. However none of the reviewed papers use WELFake dataset. This dataset stands out because it is made by combining many smaller datasets, giving us a more extensive and diverse dataset. This gives us a possibility of achieving an even higher performance model that is even better at generalization.

2 Related works

2.1 Traditional Machine Learning Algorithms

Algorithms like Logistic Regression, Support Vector Machines (SVM), Naive Bayes, Random Forest, Decision Trees, Multi-layer Perceptron, etc have been foundational for fake news classification tasks. These algorithms are based on features from term frequency-inverse document frequency (TF-IDF) that gives a numeric representation to words highlighting their importance and adjusting according to the frequency of appearance of the words.

Lai et al. [2022] implemented some of these algorithms using TF-IDF features. Random Forest performed the best with 92.877% test accuracy, demonstrating robustness relative to other algorithms on textual dataset. Similarly, Abdullah-Al-Kafi et al. [2022] had Naive Bayes perform the best with accuracy of 85.33%. When compared to other DL techniques and ensemble model approach which are going to be discussed in upcoming sections, shows inability of these ML models to accurately perform on text data.

2.2 Deep Learning Models

Various Deep Learning architectures have been employed for the task of Fake News classification including Long-short term memory (LSTM), Bidirectional-LSTM, Artificial Neural Networks (ANN), Convolution Neural Networks (CNN), Recurrent Neural Network (RNN), and Large Language Models (LLMs) like Bidirectional Encoder Representations and Transformers (BERT), Generative Pre-trained Transformers (GPT). They perform well for the task at hand as they extract features effectively and also maintain hierarchical information.

There have been many approaches of using stand alone DL architectures named above and a hybrid or combination of them. Agarwal and Dixit [2020] archives a test accuracy of 94% by using CNN and 97% using LSTM. Whereas, Lai et al. [2022] applies LSTM, CNN with DeepNetwork, and CNN with GlobalMaxpool and get test accuracies of 93.59%, 92.61% and 97.76% respectively. These findings prove the power of these methods in handling high-dimensional data and effectively capturing salient features and contextual sequences in the datasets.

2.3 Ensemble Learning Approaches

Ensemble Learning is a technique which essentially aggregates predictions from multiple models to effectively leverage their individual strength. It has proved to be an excellent approach for increasing robustness and accuracy for Fake News Classification tasks as it offers improved generalization. Combinations of many models have been used with different approaches, applying them on multiple and different datasets, and different pre-processing techniques.

Choudhary et al. [2021] develops and employs a framework called "BerConvNet" which is essentially a combination of BERT and CNN, and applies them to four different datasets to achieve highest accuracy of 97.45% on one of the datasets. Aslam et al. [2021] makes an ensemble model combining Hybrid-CNN, ANN, LSTM, and Bi-LSTM to achieve an accuracy of 89.8%. Ali et al. [2022] approaches this problem in a different way and uses deep ensemble learning for binary and multi-class classification with average accuracy of 98%. The highest average test accuracy observed was achieved by Mridha et al. [2021], their ensemble model was a combination of Bi-LSTM and RNN with accuracy of 98.75%. This shows that ensemble learning not only increases the accuracy but also the model's performance on noisy and complex datasets.

2.4 Comparison with Previous Work on the Same Dataset

In a previous optional assignment for CSE 574: Introduction to Machine Learning (Fall 2023), we worked with the same dataset but with a significantly different approach. While we retained some preprocessing techniques, such as removing URLs, lowercasing, language detection, tokenization, and stop-word removal, the current project has expanded upon these methods. We have now added special character removal, negation handling, spelling correction, number removal, and Part-of-Speech (POS)-aware lemmatization.

Previously, our analysis was limited to using Support Vector Machines (SVM) and Naive Bayes, trained exclusively on the title column of the dataset, ignoring the richer information in the text column. This led to low accuracy scores of 81.83% with SVM and 84.67% with Naive Bayes. In contrast, this project has adopted a more comprehensive approach, using 12 different models while keeping only SVM from the earlier assignment. This expanded methodology, combined with the enhanced preprocessing pipeline, has resulted in significantly better results compared to the earlier results.

3 Data

WELFake is a labeled dataset that consists of 72,134 news articles with 35,028 real and 37,106 fake news. It consists of four columns :

- Serial Numbers,
- Title of the article,
- Text body, and
- Label where 0 signifies fake news and 1 signifies real news.

This dataset is a combination of popular benchmark datasets including Kaggle, BuzzFeed Political, Reuters, and McIntire.The goal of developing WELFake was to create a larger and more diverse dataset to address the common issue of overfitting in text classifiers, as many existing datasets tend to be comparatively smaller and less varied.

We first conducted Exploratory Data Analysis (EDA) on the raw dataset to understand it better and find trends and patterns in the data. We first explore the language of title and text in the dataset:

(a) Text column (b) Title column

Figure 1: Language Distribution in the dataset

We also analyzed usage of punctuations, frequency of uppercase letters, and use of numeric characters in both categories. Following which we visualized the article length in both fake and real news and found out that real news articles are longer than fake news articles, as shown in the following histogram:

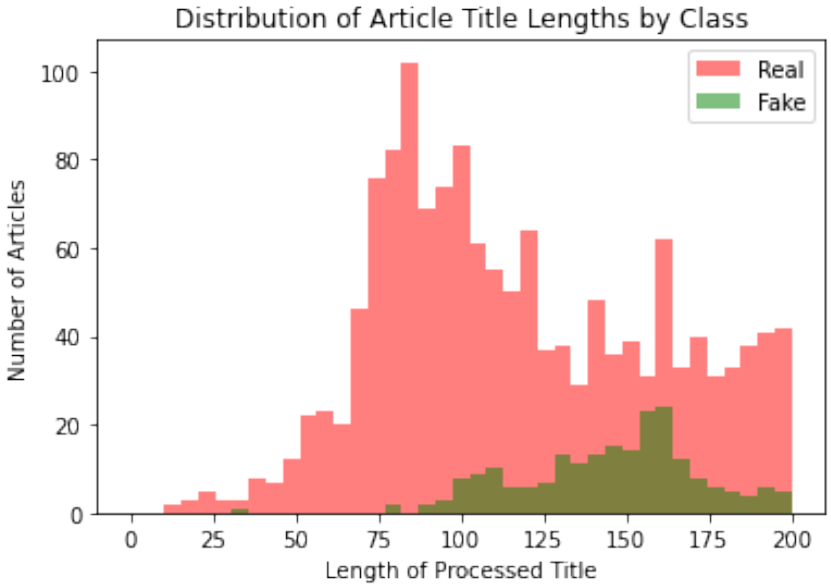
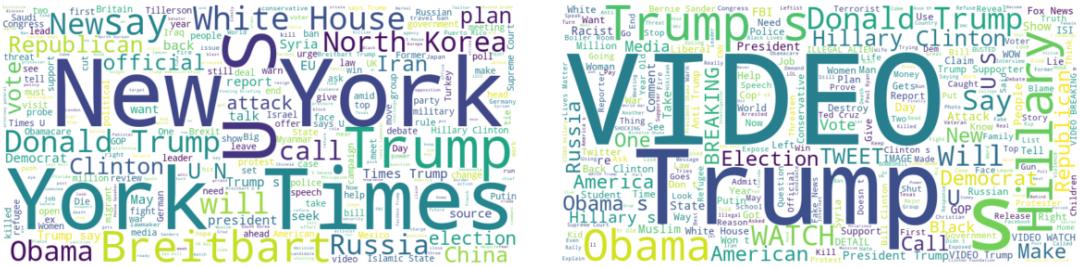


Figure 2: Distribution of title length by class

To understand the general theme of the articles, we visualized commonly used words in both classes of the title by using wordcloud. This was done separately for both the categories and revealed that most of the news articles fell into the categories of elections, political figures, geo-political situations between different countries, social media, and more.

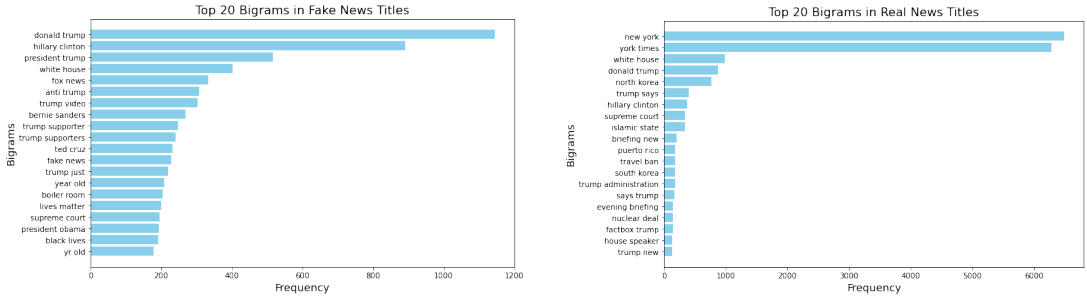


(a) Fake News

(b) Real News

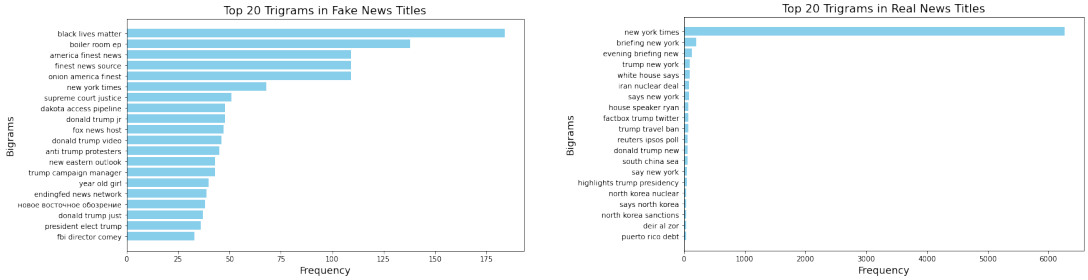
Figure 3: WordCloud of Title column

We then visualized n-grams in the title column for each class separately to further understand the common theme. Through this we could very well understand the theme of the news articles on which the news was most likely to be fake or real. But it also created confusion as there were certain sequences of words that were present in both classes like “Donald Trump”, “White House”, “Supreme Court”, “North Korea”, “South China Sea”. The following plots show the frequency of top 20 bigrams and trigrams:



(a) Fake News (b) Real News

Figure 4: Bi-grams of Title column



(a) Fake News (b) Real News

Figure 5: Tri-grams of Title column

Furthermore we also plotted a graph to get a sense of sentiment polarity within the dataset to understand if the data was biassed or not, the plot shows that the distribution was mostly neutral.

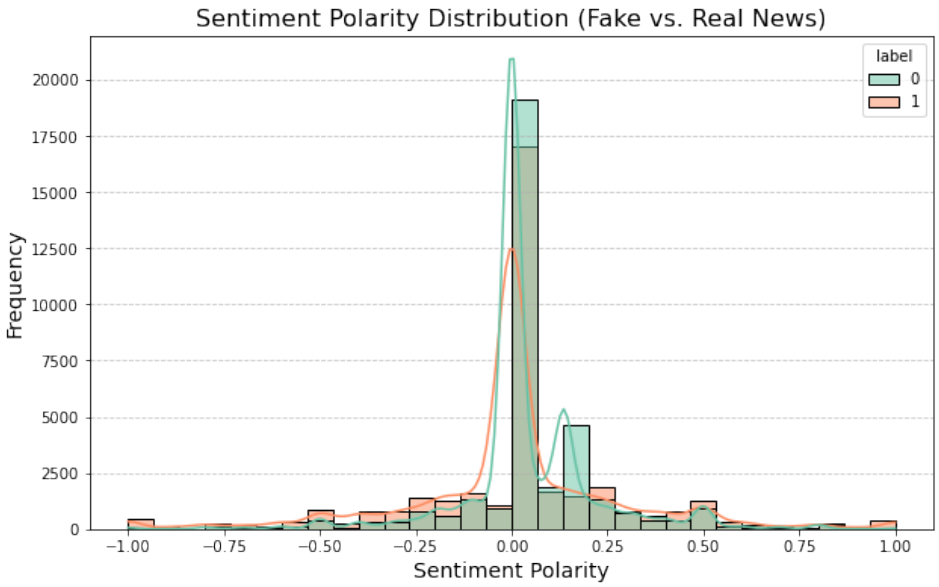


Figure 6: Sentiment Distribution

After getting better understanding on the dataset through EDA, the dataset was preprocessed by removing null values, URLs, contractions, numbers, characters and stopwords. The dataset was also

tokenized, lemmatized, and to mitigate a little bit of bias that was shown above in the sentiment polarity graph above was done through negation handling. The preprocessing took approximately 10 days to run as the dataset was huge.

4 Methods

For this text classification problem, we employed a combination of Machine Learning (ML) algorithms, a Deep Learning model (LSTM), and a transformer-based model (BERT), leveraging feature engineering techniques to enhance performance.

Firstly, we applied Logistic Regression, Support Vector Machine (SVM), Random Forest and XGBoost to establish baseline performance. We then applied Term Frequency-Inverse Document Frequency (TF-IDF) on the dataset which is a feature engineering technique specifically for NLP tasks. It has two parts, Term Frequency which checks the frequency of words in the data and Inverse Document Frequency that measures rarity of words across the dataset. Based on this, it assigns numerical weights to each word and coverts the whole input into 1 vector representation. We applied Random Forest and XGBoost again on feature enginnered data as they were the better performing model.

We then applied a Deep Learning model, Long Short-Term Memory (LSTM), as it offers learning sequence based patterns as it understands text at a deeper level than traditional ML algorithms and maintains context as well. This took over 3 hours to train the model on CPU. Following which we made an ensemble model combining Random Forest, XGBoost, and LSTM. This ensemble model combines all three models to make predictions, which is evident by the results as we got more accuracy using this model. Following this, we developed another ensemble model with weighted average, assigning heavier weights to better performing models, in the ratio 40:40:20 for random forest, XGBoost and LSTM respectively.

We then applied Bidirectional Encoder Representations from Transformers (BERT), which is a state-of-the-art transformer based model and it resulted in even more increased accuracy. BERT, developed by Google in 2018, is a masked language model which uses encoder only transformer architecture. BERT took a little over 4 hours to train on CUDA. We then developed two additional ensemble models, first one was a combination of BERT and LSTM, and the second one was a combination of Random Forest, XGBoost, LSTM and BERT.

5 Results

The results obtained from the models clearly demonstrate the progression in performance from classical machine learning approaches to advanced ensemble and deep learning models as shown in the following table:

Table 1: Accuracy and loss values of all the models on test set.

Model	Accuracy(%)	Loss
Logistic Regression	68.21	0.619
Support Vector Machine (SVM)	78.33	0.4738
Random Forest	83.06	0.3966
XGBoost	81.44	0.4133
Random Forest with TF-IDF	95.77	0.2046
XGBoost with TF-IDF	96.46	0.1009
LSTM	93.65	0.2924
Ensemble Model - 1 (LSTM + RF with TF-IDF + XGB with TF-IDF)	96.89	0.0899
Ensemble Model - 2 (Weighted Average on Ensemble Model - 1)	96.66	0.0899
BERT	99.329	0.0531
Ensemble Model - 3 (BERT + LSTM)	99.329	0.1182
Ensemble Model - 4 (BERT + Ensemble Model - 1)	99.329	0.0423

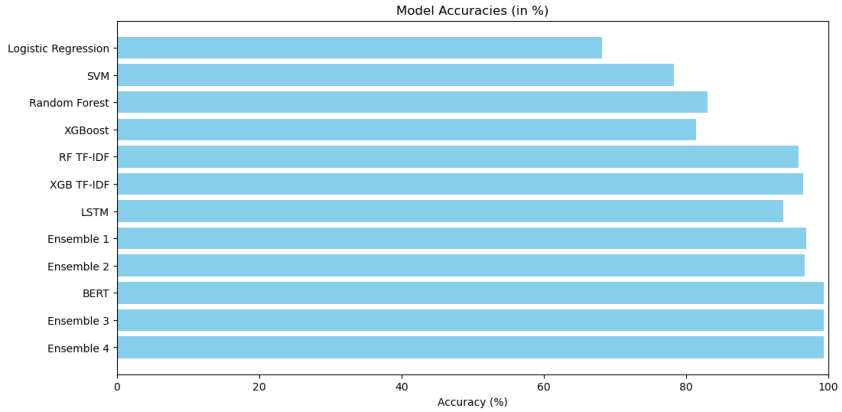


Figure 7: Plot showing accuracy of all models.

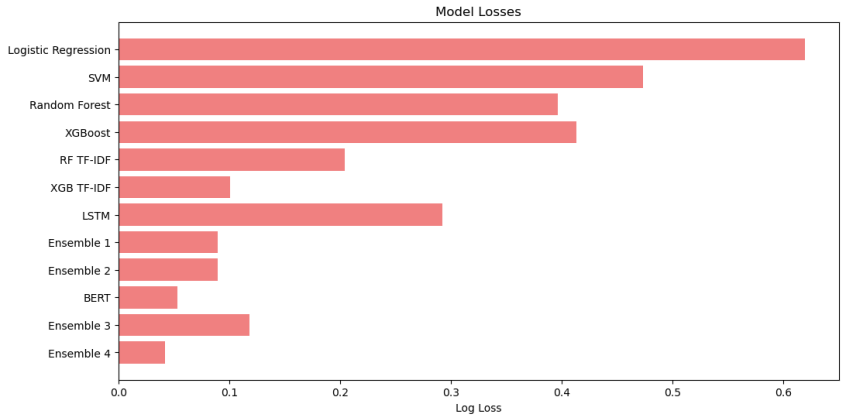


Figure 8: Plot showing loss of all models.

Logistic Regression and SVM do not perform that well and achieve moderate accuracy with high loss. Whereas Random Forrest and XGBoost perform better, and their performance inreases even further when they are applied on feature engineered data through TF-IDF.

LSTM achieved a strong accuracy 93.65%. The performance is further increased in Ensemble Model 1 and 2. BERT and the following models achieve the same and highest accuracy of 99.329% with Ensemble Model 4 achieving the lowest loss of 0.0423.

The following plot compares the performance of our models along with the performance achieved by the research papers mentioned in our literature survey:

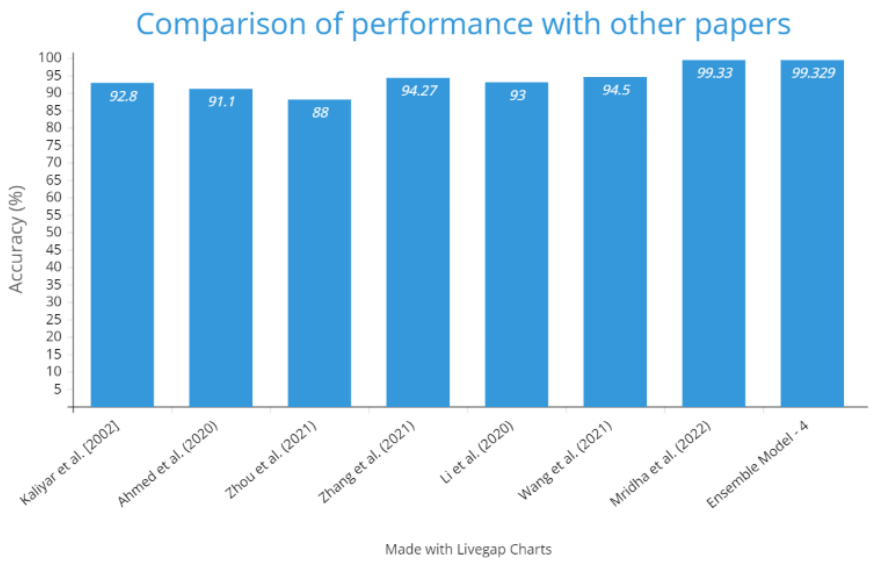


Figure 9: Comparison of performance with papers from Literature Survey

6 Conclusion and future work

The project explored a range of ML and DL models providing valuable hands-on experience with real-world problems and datasets. A key learning outcome was the impact of feature engineering, particularly TF-IDF, which dramatically improved the performance of tree-based models. Additionally, working with LSTM and BERT enhanced our understanding for natural language processing (NLP) tasks. The implementation of ensemble learning further strengthened our ability to combine diverse models for achieving better performance.

To improve the results and expand the scope of this project, following areas can be explored:

- Automating hyperparameter tuning by using Optuna.
- Lack of other languages results in lower generalization, which can be increased by translating the non-english articles.
- Inclusion of other languages.
- LLMs like BERT can be made from scratch to have more control over performance.
- This project was focused on political theme, so exploring other domains like finance, medicine, law, and more is a possibility.

References

Md. Abdullah-Al-Kafi, Israt Jahan Tasnova, Md. Wadud Islam, and Sumit Kumar Banshal. Performances of different approaches for fake news classification: An analytical study. In Isaac Woungang, Sanjay Kumar Dhurandher, Kiran Kumar Pattanaik, Anshul Verma, and Pradeepika Verma, editors, *Advanced Network Technologies and Intelligent Computing*, pages 700–714, Cham, 2022. Springer International Publishing. ISBN 978-3-030-96040-7.

Arush Agarwal and Akhil Dixit. Fake news detection: An ensemble learning approach. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1178–1183, 2020. doi: 10.1109/ICICCS48265.2020.9121030.

Abdullah Marish Ali, Fuad A. Ghaleb, Bander Ali Saleh Al-Rimy, Fawaz Jaber Alsolami, and Asif Irshad Khan. Deep ensemble fake news detection model using sequential deep learning technique. *Sensors*, 22(18), 2022. ISSN 1424-8220. doi: 10.3390/s22186970. URL `https://www.mdpi.com/1424-8220/22/18/6970`.

Nida Aslam, Irfan Ullah Khan, Farah Salem Alotaibi, Lama Abdulaziz Aldaej, and Asma Khaled Aldubaikil. Fake detect: A deep learning ensemble model for fake news detection. *Complexity*, 2021(1):5557784, 2021. doi: https://doi.org/10.1155/2021/5557784. URL `https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/5557784`.

Monika Choudhary, Satyendra Singh Chouhan, Emmanuel S. Pilli, and Santosh Kumar Vipparthi. Berconvonet: A deep learning framework for fake news classification. *Applied Soft Computing*, 110:107614, 2021. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.107614. URL `https://www.sciencedirect.com/science/article/pii/S1568494621005354`.

Chun-Ming Lai, Mei-Hua Chen, Endah Kristiani, Vinod Kumar Verma, and Chao-Tung Yang. Fake news classification based on content level features. *Applied Sciences*, 12(3), 2022. ISSN 2076-3417. doi: 10.3390/app12031116. URL `https://www.mdpi.com/2076-3417/12/3/1116`.

M. F. Mridha, Ashfia Jannat Keya, Md. Abdul Hamid, Muhammad Mostafa Monowar, and Md. Saifur Rahman. A comprehensive review on fake news detection with deep learning. *IEEE Access*, 9:156151–156170, 2021. doi: 10.1109/ACCESS.2021.3129329.

- Link to dataset : https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification
- Optional Assignment - CSE 574 Intosuction to Machine Learning (Fall 2023)
- https://medium.com/@maleeshadesilva21/preprocessing-steps-for-natural-language-processing-nlp-a-beginners-guide-d6d9bf7689c9
- "Attention is all you need" : https://arxiv.org/abs/1706.03762
- https://en.wikipedia.org/wiki/Tf
- https://www.sciencedirect.com/topics/computer-science/bidirectional-long-short-term-memory-network
- https://en.wikipedia.org/wiki/Transformer$_(deep_learning_architecture)$

- https://arxiv.org/abs/1810.04805