Deep Learning Approaches in Fake News

Kaiwen Hu **Yijia Xue** **Yuzhuo Kang** **Zhixuan Li**
{kaiwenh4, xueyijia, yuzhuok}@seas.upenn.edu lizhx@sas.upenn.edu

Abstract

Fake news detection algorithms play a pivotal role in safeguarding the integrity of information in our digital age. With the exponential growth of information on social media platforms, the ability to discern between authentic and deceptive content is crucial. In this study, binary classifiers for fake news detection are trained on the LIAR and FNC-1 datasets, and their performance against published baseline works is evaluated. The model with the best results achieved an accuracy of 70.24% and an F1 score of 0.558, by utilizing an ensemble model of BERT and LSTM. GitHub[1]

1 Introduction

In the latter part of the twentieth century, the internet revolutionized how people share information, often without stringent editorial standards. Recently, social media has emerged as a significant news source for many individuals. As reported by Statistica[2], approximately 3.6 billion people worldwide are active social media users. Social media offers evident advantages in news dissemination, including immediate access to information, free distribution, no time constraints, and diverse content. However, these platforms lack substantial regulation and oversight (Raza and Ding, 2022).

Fake news detection falls under the umbrella of text classification, where it is divided into binary classification (distinguishing between real and fake news) or multi-class classification for higher granularity. We are interested in reviewing the state-of-the-art approaches for labeling fake news detection using various datasets. The main goal of our project will be to build a mixed dataset with different news domains.

[1]https://github.com/kevin00hu/CIS-5300-final-project
[2]https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

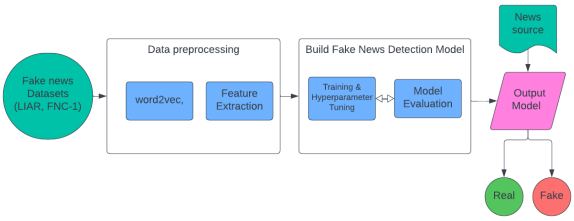

Figure 1: Illustration of Fake News Detection task

2 Literature Review

Much of the previous research on fake news detection has been exploratory with the neural network on labeled data first introduced deep learning method (RNN, LSTM and GRU) to the fake news detection domain in 2016 (Ma et al., 2016). The proposed model utilizes RNNs to learn continuous representations of microblog events, capturing the evolving contextual information over time. Results from experiments on datasets from real-world microblog platforms reveal that the RNN-based method surpasses existing rumor detection models reliant on hand-crafted features. The RNN-based approach demonstrates quicker and more accurate rumor detection than established techniques, including leading online rumor debunking services.

What's more, machine learning approaches are also the popular methods in dealing with Fake news detection (Ahmed et al., 2021). Manzoor et al. (2019) conducts an in-depth analysis of existing research on fake news detection, focusing on traditional machine learning models. The proposed approach involves leveraging tools such as Python's scikit-learn and natural language processing (NLP) for textual analysis. The process encompasses feature extraction and vectorization, employing Scikit-learn's Count Vectorizer and Tiff Vectorizer for tokenization and feature extraction. Additionally, feature selection methods are employed to identify the most suitable features that yield optimal preci-

sion, guided by the results of a confusion matrix.

Yu et al. (2017) proposes a novel method, the Convolutional Approach for Misinformation Identification (CAMI) based on a Convolutional Neural Network (CNN). CAMI can flexibly extract key features scattered among an input sequence and shape high-level interactions among significant features, which help effectively identify misinformation and achieve practical early detection. Experiment results on two large-scale datasets validate the effectiveness of the CAMI model on both misinformation identification and early detection tasks.

Many other models have been successfully built up based on supervised perspective (Ma et al., 2019; Vaibhav et al., 2019). There are also many approaches using weakly-supervised methods. Konkobo et al. (2020) trained a supervised model and an unsupervised model, which utilizes not only the news content information but also the user's comment information on the news as well as the author's credibility information. Mansouri et al. used the LDA method to pseudo-label unlabeled data for better training of unlabeled CNN models with good results (Mansouri et al., 2020).

For the LIAR dataset used by this study, introduced by Wang (2017), previous work had been done in performing both multi-class classification and binary classification. Wang's paper examines multiple methods for multi-class classification, while Khan et al. (2021) performs binary classification on the preprocessed datasets. The models used by Khan et al. (2021) include traditional machine learning methods such as SVM and Naive Bayes, as well as pre-trained language models like BERT and RoBERTa.

3 Experimental Design

3.1 Data

In our project, we are aiming to discover two main datasets and to find out the model performance of combining them.

The first dataset used is the LIAR dataset [3]. It is a dataset created for the purpose of advancing research in the field of fake news detection. It was introduced in 2017 by William Yang Wang (Wang, 2017). The dataset is designed to facilitate the development and evaluation of algorithms for the automatic detection of fake news or misinformation. This is the main dataset to be used for evaluation. It is split into train/dev/test sets, with their respective lengths shown in Table 1. The dataset mainly consists of Statements and Labels. Each statement is labeled with one of six possible labels indicating the degree of truthfulness: False, Barely True, Half True, Mostly True, True, Pants on Fire (for egregiously false statements). For our classification task, we assigned binaary values to the labels, with Mostly True and True being considered true, and the rest being false. After preprocessing the data, among the training set, 35% of the data is labeled as True, while the rest is labeled as False.

Dataset	Length
LIAR_train	10240
LIAR_dev	1284
LIAR_test	1267
FNC_train	39978
FNC_dev	4997
FNC_test	4997

Table 1: Number of senctnces in each dataset after data preprocessing.

The second dataset that we are incorporating is the FNC-1 dataset [4]. It is a well-known dataset in the field of natural language processing and machine learning. The Fake News Challenge is an initiative that aims to explore and improve the automatic detection of fake news. The FNC-1 dataset specifically focuses on the task of stance detection, where the goal is to determine the stance of a body of text (typically a news article) with respect to a headline. The stance is labeled "agree", "disagree", "discuss" or "unrelated", which suggests a focus on distinguishing related from unrelated statements rather than the traditional stance detection. In data preprocessing, labels "agree", "disagree", and "discuss" are mapped to 1, while "unrelated" is mapped to 0, which transforms it into a binary mapping of whether the input sentences are related in content to the headline.

3.2 Evaluation Metric

As a classical classification task, we use four evaluation metrics to observe the model performance. The model output and the true labels are stored on

[3] https://www.cs.ucsb.edu/ william/data/liar_dataset.zip

[4] http://www.fakenewschallenge.org/

the .npy files. The evaluate function loads the true labels (y_true) and predicted labels (y_pred) from the specified file paths. After loading, it computes several key evaluation metrics: accuracy, precision, recall, and F1 score. Each metric is calculated using functions from the sklearn.metrics module. The equations used to calculate the metrics are as follows:

Accuracy: The fraction of correct predictions among the total number of cases evaluated.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Precision: The ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

where (TP) are true positives and (FP) are false positives.

Recall (Sensitivity): The ratio of correctly predicted positive observations to all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

where (TP) are true positives and (FN) are false negatives.

F1 Score: The weighted average of Precision and Recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The results of the test are also visualized by the confusion matrix as a heatmap, which will be also plotted by the evaluation script score.py.

In the domain of fake news detection, while a model may exhibit high classification results, it is necessary to assess the model's ability under varying circumstances. The F1-Score provides a balanced measure in cases where precision and recall may trade off against each other in imbalanced datasets. Precision is also important if falsely identifying real news as fake can have significant impacts on the results (Mridha et al., 2021).

3.3 Simple baseline

The simple baseline is built by a simple and shallow neural network written with Keras, consisting of a stack of densely connected layers (Dense) with the 'relu' activation function. The output layer uses the 'sigmoid' activation, appropriate for binary classification tasks.

We train a 4-layer neural network model with only the LIAR dataset of 10 epochs and a batch size of 256. The accuracy on testset is **64.72%** and the F1 score is **0.40**.

To use the FNC dataset which gives the relation label for news body and news title, we trained a simple neural network. We got ***64.85%*** accuracy and ***0.51*** f1 score for this model. We are assuming this model should have some feature extraction function after training. Then, we feed the LIAR data into this model, we can get a binary label indicating a potential relationship between the first sentence and the rest of the sentence.

With the extra feature obtained from the FNC model, we train a new neural network with the structure as the first model. The f1 score is better than the model using just LIAR dataset - **0.45**, but the accuracy, **62.11%**, is lower than the first model.

We noticed that the best result on the LIAR-PLUS test set is lower than our expectation. The best model achieved a **77.2%** accuracy for binary classification and a **37.4%** accuracy for 6-class classification.

4 Experimental Results

4.1 Published Baseline

LSTMs are used to process word vectors generated from news content (FNC and LIAR datasets), By utilizing the word2vec model to convert text into numerical vectors, these vectors become the input for the LSTM network. Traditional RNNs may have problems with long-term dependencies due to the vanishing gradient problem. LSTMs solve this with the structure of input, forget, and output gates. It allows them to retain information over longer periods, which is crucial for understanding the context of news articles. Unlike feedforward networks, LSTMs can model sequences of variable lengths, which is essential in news content where article lengths can vary significantly. In this project, the LSTM structure is a double-layered LSTM with 150 hidden units per layer, incorporating dropout of 0.4 for regularization. The batch size is 1024. It processes input features through these LSTM layers, followed by a dropout and a linear layer, to perform binary classification of textual data for fake news detection. The dataset is imbalanced

so LSTMs may be biased towards the majority class. We try to solve it by introducing the weight parameter in the initialization of the criterion.

The accuracy is **61.64%**, and the F1-score is **0.52**. The accuracy is lower than other baseline models, but the F1-score is a bit higher. This indicates LSTM model is better than the baseline models. Also, it shows a conflict on two different scoring matrix: accuracy and f1 score.

4.2 Extension 1

We used the BERT (Devlin et al., 2018) (Bidirectional Encoder Representations from Transformers) model for the extension. BERT is a pre-trained language model based on the Transformer architecture, introduced by Google in 2018 (Vaswani et al., 2017). Its principles mainly involve two key concepts: pre-training and fine-tuning. In the pre-training phase, BERT learns deep language representations through large-scale language model tasks such as masked language modeling and next-sentence prediction. In the fine-tuning phase, the BERT model can be fine-tuned based on annotated data for specific tasks to adapt to particular application scenarios.

Compared to simple baseline and strong baseline models, BERT has significant advantages in the task of fake news detection. BERT is a Transformer-based model that utilizes attention mechanisms to capture global information in the text, allowing for a better understanding of contextual relationships. This enables BERT to handle long-distance dependencies more effectively, which is crucial for fake news identification as texts often have complex contextual structures. Additionally, the pre-training mechanism of the BERT model allows it to learn universal language representations from large-scale text data, providing strong generalization capabilities. In contrast, traditional models like LSTMs often require more labeled data for training and are prone to overfitting. In the context of fake news detection, where genuine fake news data is typically limited, BERT's generalization performance makes it excel in small-sample scenarios.

In this extension, we used the "bert-base-uncased" model from HuggingFace. Considering model training time, complexity leading to overfitting, and GPU memory constraints, we set the number of BERT model layers to 2. When constructing the model, we used the AutoTokenizer and AutoModelForSequenceClassification from the transformers library to load the tokenizer and model, respectively. For model training, we employed the AdamW (Loshchilov and Hutter, 2017) optimizer, which introduces weight decay to address potential overfitting issues with Adam (Kingma and Ba, 2014) in certain cases. The learning rate was set to 2e-5, and weight decay was set to 0.01. Additionally, we used a learning rate scheduler, starting with a small learning rate in the early training stages, gradually increasing it, and then maintaining a relatively stable learning rate in the later stages. This helps improve the model's stability in the early training phase and accelerates convergence in the later stages. We set the warm-up steps to 0.1 times the total steps.

To show the training effectiveness, we used mask prediction from the bert-base-uncased" model. The model achieves an accuracy of **41%** and an F1 score of **52%** on the test set. The results of the fine-tuned BERT-based extension are as follows: accuracy is **68.82%**, and F1-score is **0.51**. This shows an improvement compared to both the simple baseline and the strong baseline. Due to the uneven distribution of samples, leading to insufficient learning of positive class data, future considerations may involve adopting data augmentation to expand positive class data.

4.3 Extension 2

Just like in Section 4.2, we employed the "bert-base-uncased" model from Hugging Face, coupled with an LSTM model. The tokenizer was loaded using AutoTokenizer for the BERT model. Initially, the "statement" and "context" fields of fake news were concatenated with [SEP], then tokenized using the tokenizer. The tokenized results were fed into the BERT model, and the output from the last layer was passed into the input of a two-layer LSTM model, incorporating bidirectional encoding. The final hidden layer output from the last time step of the LSTM was fed into a fully connected layer with a ReLU activation function, resulting in a two-dimensional output. After evaluation on the test set, the accuracy of the BERT + LSTM model reached 70.24%, with an F1 score of 0.558.

4.4 Error analysis

Due to the imbalance in our training data, the quantity of samples with label 0 is nearly twice that of label 1. This imbalance causes our model to be more sensitive to fake news, making it more

likely to incorrectly predict real news. Additionally, our model exhibits a non-negligible probability of misclassifying instances with the original label "barely true." This largely indicates that the model has some difficulty in accurately identifying data with unclear intentions of this nature.

5 Conclusions

Overall, our best performing model was the BERT + LSTM model, with an accuracy of **70.24%** and F1 score of **0.558**. Compared to previous works using the same datasets, Khan et al. (2021) was able to achieve an accuracy of 62% and an F1 score of 0.63 for the LIAR binary classification task, with a slightly different binary label-assigning process where "half-true" were also assigned to be true. Our accuracy is slightly lower than the published result, this could be due to different data preprocessing features, such as removing suffices and stemming words performed in the published study (Khan et al., 2021). Furthermore, Mridha et al. (2021) states that real-world fake new detection algorithms would not only use text-based data and metadata but also incorporate surrounding context information such as visual features and social context information. Thus, for this study, more work can be done in terms of both testing out other ensemble models and increased feature engineering.

References

Alim Al Ayub Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. 2021. Detecting fake news using machine learning : A systematic literature review.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Junaed Younus Khan, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. 2021. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Pakindessama M. Konkobo, Rui Zhang, Siyuan Huang, Toussida T. Minoungou, Jose A. Ouedraogo, and Lin Li. 2020. A deep learning model for early detection of fake news on social media. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pages 1–6.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Jim Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*, WWW '19, page 3049–3055, New York, NY, USA. Association for Computing Machinery.

Reza Mansouri, Mahmood Naderan-Tahan, and Mohammad Javad Rashti. 2020. A semi-supervised learning method for fake news detection in social media. In *2020 28th Iranian Conference on Electrical Engineering (ICEE)*, pages 1–5.

Syed Ishfaq Manzoor, Jimmy Singla, and Nikita. 2019. Fake news detection using machine learning approaches: A systematic review. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 230–234.

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

Shaina Raza and Chen Ding. 2022. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13:335 – 362.

Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 134–139, Hong Kong. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection.

Feng Yu, Q. Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. In *International Joint Conference on Artificial Intelligence*.