Zero-shot fake news detection

Asmita Mukherjee(MT21115), Harsh Vardhan Bhadauriya(MT21122),
Hanzalah Firdausi(MT21027), Shreyansh Jain(MT21089), Sehban Fazili(MT21143)

March 13, 2023

1 Introduction

The potential effect of false information has increased due to the growth of social media. Major media and technology companies have made a sizable investment in the fight against false news, amounting to over $300 million, with help from a number of activists, NGOs, and interest groups. Millions of people worldwide are affected by the spread of fake news and misinformation, emphasising the need for a remedy. An online eradication strategy built on technology is required due to the persistent spread of false information. To solve this problem, we present a tool that verifies the reliability of news reports. The tool must assess the details, incidents, and associated data in a news story.

The popular fake news detection methods are trained on the dataset provided at the training time. The training set does not contain the data points for the current events, so it is hard for the model to predict if a recent event is introduced at the test time. Our model will crawl the web for relevant and popular articles and use them in prediction. Our model will be able to perform in real-time and helps to make sure that we can give a sound prediction for current articles. Another important contribution is to eliminate the training phase. Most of the algorithms in the domain are deep learning and supervised in nature. They have an extended training phase. We aim to design a zero-shot architecture that does not need to see any training data and could directly start with the testing/prediction phase. We will also work on minimizing the prediction time.

We propose a two-step solution: Question-Answer generation and Answer generation task. The Question Answer Generator will be utilized to generate questions and answers from a given input news article which is to be evaluated. The verified news articles will be retrieved by conducting a reverse text search on credible and reputed news websites, focusing on a specific topic. The pre-trained answer generation model will then be used to generate answers to the questions generated by the QA Generator, using the trustworthy news article as context. The answers from the input news article will be compared with the answers from credible sources using evaluation metrics. Based on the results of this comparison, a determination will be made as to whether the news article is trustworthy or not, using a set threshold.

Furthermore, we aim to create a robust generalized model that can work on any test dataset to predict whether a news article is fake or real with greater confidence to improve the application of our work. The metrics that can be used to evaluate the similarity between the answers of the input news article and the verified news article are as follows:

- **Bleu score:** Bilingual Evaluation Understudy score is a metric that helps to understand the differences between a machine-generated text and the ground truth. Even Though it is easy to calculate, however, it does not take different meanings of words into account and does not take the order of words into account. Hence along with Bleu score, we also need to take into account other metrics along with human evaluation.

- **Rouge score:** The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scoring calculates the similarity between a candidate document and a collection of reference documents.

- **METEOR score:** METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric for the evaluation of machine-translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.

- **BERT score:** BERT score tries to adjust the deficiencies of ROUGE and BLEU. Since ROUGE and BLEU scores do not take into account the context, BERTScore takes BERT embeddings of the machine-generated text and the ground truth and finds the cosine similarity between them. Hence BERTScore takes context into account and gives a better idea of the quality of the generated text.

2 Literature Survey

We have summarised few papers related to our work.
Gunasekara et al[1]proposed generative modelling by which we were inspired by. In this paper the author shares the idea of using Question Answering Model and Answer Generator Model. They also present a general framework for training abstractive summarization models to address problems with omitting important facts from sources and including facts that are incongruent with the sources. To summarise papers, they first train a sequence-to-sequence model. This model is then trained further using reinforcement learning with rewards based on question-answering. They use a combination of numerous automatic measurements and human judgement to assess the summaries produced by this framework. Three widely used transformer-based summarization models and two publicly accessible datasets are used to assess the proposed framework.

Imbwaga et al.[6] used tree-based decision tree classifiers and a gradient-boosting ensemble algorithm for fake news detection. Standard preprocessing steps of converting the text to lowercase, and the special characters, dot values, links, and null values were removed. In order to convert the text into feature vectors with which the machine learning algorithm can be trained, TF-IDF was used. In order to create TF-IDF, vocabulary was created from the text after tokenization. For the classification models, the following models were used, Logistic Regression Classifier, Decision Tree Classifier, Gradient Boosting Classifier, and Random Forest Classifier. The paper got the best accuracy using Decision trees and gradient boosting.

Kaliyar et al.[2] proposed an architecture for fake news detection by combining different parallel blocks of the single-layer deep CNN with the BERT. They utilize BERT to get the contextual representation of a sentence. The architecture is based on three parallel blocks of 1-D CNN with BERT having different kernel sizes. This is followed by the max pooling layer after each block. Every document is processed through different CNN configurations having different kernel sizes and filters. The CNN architecture aims at addressing ambiguity which is one of the main challenges of fake news detection. They also show that their model surpasses the current benchmarks for classifying fake news. This method is supervised in nature and is only applicable to the data distribution of the training set. We have planned to overcome the supervised nature of pre-existing methods by designing a zero-shot prediction architecture and eliminating the training phase.

Rodriguez et al. [4] investigate the feasibility of applying deep learning techniques to discriminate fake news on the Internet using only text. Three neural network architectures are proposed, one based on BERT, a modern language model created by Google. The next architecture is based on LSTM. LSTM cells are recurrent neurons that have the capability to remember information from the past. They are composed of gates that maintain a hidden cell state, allowing them to remember more distant information than vanilla recurrent units. This is important in NLP as words from the past often influence the current ones. The next architecture is based on Convolutional Neural Networks (CNN). CNN are computer vision and NLP networks that work by applying a series of filters to their input. These filters are N-dimensional matrices which are slided (convoluted) over the input, and after training the network, they produce activations (known as feature maps) where certain patterns are detected.

The work by Liu, Bang, et al[3] seeks to automatically generate high-quality and varied question-answer pairs from unlabeled text corpora at scale by mimicking a human questioner's style. Just like human asks meaningful questions, the system extracts multiple aspect information from the text. The architecture uses neural network models based on the multiaspect information retrieved to produce a range of queries in a manageable manner. The method essentially transforms the one-to-many mapping problem into a one-to-one mapping problem. As a result, it can be scaled up or down while maintaining high-quality question generation.

Sanh et al [5] proposes DistilBERT, a smaller and faster version of BERT (Bidirectional Encoder Representations from Transformers), a popular language model that achieves state-of-the-art results in various natural language processing tasks. DistilBERT was created by compressing the original BERT model while retaining its performance. The compression technique used in DistilBERT involves distillation, which involves training a smaller model to mimic the behavior of a larger model. In this case, the smaller model is trained to replicate the output of the original BERT model on a set of tasks, but with fewer parameters. This results in a model that is significantly smaller and faster, making it more suitable for deployment on resource-constrained devices or for applications where low latency is critical. Despite its smaller size, DistilBERT is still capable of achieving competitive performance on various natural language processing tasks, including question-answering, sentiment analysis, and text classification. Its smaller size also makes it easier to train and fine-tune on smaller datasets.

3 Baseline

Our objective is to classify the given news article as fake or real correctly. We have modified our objective by training the baseline on one dataset and testing the model on a different dataset to see how well our model generalizes regarding fake news detection. We preprocess both datasets by removing punctuations, spaces, stopwords etc. We used the WELFake dataset to train our model with 14000 train samples having 6700 fake news samples and 7300 real news samples. We tested our models on the Fake News Detection dataset with 2000 samples with 1056 fake samples and 944 real samples. The different methods for baseline evaluations are as follows.

3.1 TF-IDF vectorization with ML classifier

We have vectorized our data using TF-IDF vectorization and used SVD to reduce the dimension to 20 components. We have applied different machine learning models to our dataset, and the results are captured in table 1

Method	Train F1-score	Test F1-score
Logistic Regression	0.86	0.20
KNN	0.91	0.25
Decision Tree	1.00	0.26
SVM	0.89	0.19

Table 1: Baseline result using TF-IDF vectorization and ML models

3.2 Bert Embeddings with ML classifier

We have vectorized our data using Bert embeddings that generated a vector of dimension 768 for each data sample. We have applied different machine-learning models to our dataset, and the results are as follows 2

Method	Train F1-score	Test F1-score
Logistic Regression	0.95	0.22
KNN	0.95	0.31
Decision Tree	1.00	0.27
SVM	0.94	0.19

Table 2: Baseline result using Bert embeddings and ML models

3.3 Bert Classifier

We have also used the pre-trained Distlilbert classifier and fine-tuned on our dataset. We used a pre-trained Distilbert tokenizer to generate train, test and validation embeddings. The model is trained using 3 train epochs, and it took 1120 global epochs to train our model. The training batch size is 32, and the eval batch size is 64. The train and test F1 score as follows 3:

Method	Train F1-score	Test F1-score
Distilbert classifier	1.00	0.19

Table 3: Baseline result using Distilbert classifier

3.4 Conclusion

While experimenting with different baseline methods, we have observed that our trained model needs to generalize better on the test set in the task of fake news detection. Going forward, we intend to train and create model that can generalize well on the test data and achieve a better F1 score.

Equal contribution from all team members.

References

[1] Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Ranit Aharonov, and Sachindra Joshi. Using question answering rewards to improve abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 518–526, 2021.

[2] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.

[3] Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043, 2020.

[4] Álvaro Ibrain Rodríguez and Lara Lloret Iglesias. Fake news detection using deep learning, 2019.

[5] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[6] Uma Sharma, Sidarth Saran, and Shankar M Patil. Fake news detection using machine learning algorithms. *International Journal of Creative Research Thoughts (IJCRT)*, 8(6):509–518, 2020.