Zero-shot fake news detection

Asmita Mukherjee
MT21115

Harsh Vardhan Bhadauriya
MT21122

Hanzalah Firdausi
MT21027

Shreyansh Jain
MT21089

Sehban Fazili
MT21143

Abstract

With the internet's rise in reach and spread, spreading and disseminating false information has become easier. Millions of people are affected by the spread of misinformation, which creates confusion, mistrust and disharmony and can be utilised to force false narratives and propaganda. The spread of misinformation undermines the ability of people to take correct decisions by spreading uncertainty and requires a remedy to combat the same. Due to the persistent spread of false information, an online eradication strategy built on technology is required. To solve this problem, we present a tool that verifies the reliability of news reports, allowing end-users to make better-informed decisions.

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. Due to the persistent spread of false information, an online eradication strategy built on technology is required. To solve this problem, we present a tool that verifies the reliability of news reports. The tool must assess a news story's details, incidents, and associated data.

1.1. Motivation

Most current fake news detection models are supervised models trained extensively on a dataset set given at the training time. These models involve extensive training to make a reliable judgement during the training phase. The performance may be bounded by the nature of the dataset that was given during the training phase, and models may face an issue in making a sound prediction if a recent news article is given to check for its authenticity. Motivated by this problem, we propose a zero-shot architecture that does not need to see any training data and can directly start with the testing/prediction phase with high reliability in real time. We also aim to minimize the prediction time of our model.

1.2. Problem Statement

The task emphasizes zero-shot fake news detection of the current news topics. Our model can reliably identify current news articles as authentic or fake without any training in a real-time basis to the end-user.

2. Literature Survey

We have summarised few papers related to our work. Gunasekara et al [2]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. [8] 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 were removed, and the special characters, dot values, links, and null values. 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. The following models were used for the classification models: 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. [3] 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 addresses 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 only applies to the training set's data distribution. 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. [6] 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 [5] 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 [7] 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 significantly smaller and faster model, making it more suitable for deployment on resource-constrained devices or for applications where low latency is critical. Despite its smaller size, DistilBERT can still perform competitively 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.

Nan Duan et al [1] addresses the task of generating questions for question answering, which is an important subtask of natural language processing (NLP). The task involves generating a natural language question based on a given passage of text, such that the answer can be found within the passage. The paper proposes a neural network-based model that generates questions by simultaneously predicting the answer spans and question types. The model takes a passage of text as input and generates a question by selecting a relevant span of text from the passage and transforming it into a question form based on the question type. The proposed model consists of several components, a passage encoder, an answer span predictor, a question type predictor, and a question decoder. The passage encoder encodes the input passage into a fixed-length vector representation, which is then used by the answer span predictor and the question type predictor to predict the answer span and the question type, respectively. The question decoder then generates the final question by transforming the predicted answer span into a question form based on the predicted question type.

Lewis et al [4] proposes a new approach to unsupervised question answering based on cloze translation. Cloze translation is a technique used in machine translation where a blank in a sentence is replaced with a word or phrase from the target language. This technique is commonly used in language learning and has been shown to be effective for generating translations. The cloze translation approach to unsupervised question answering works by taking a question and generating a corresponding sentence in the target

language by replacing a blank in the original sentence with the answer to the question. The system uses a neural machine translation model trained on a large corpus of parallel texts to generate these answers. The paper argues that traditional unsupervised question-answering approaches rely heavily on pre-existing knowledge and heuristics, which limits their effectiveness. Instead, the cloze translation approach leverages the power of machine translation to generate answers based on the context of the question without relying on any training data.

3. Novelty

Our zero-shot fake news detection model is novel due to the following points

- Since the method is zero-shot, it can make a classification almost instantly while other deep-learning and supervised methods require an extensive training phase.
- Zero-shot method performs extremely well on unseen classes.
- In our baseline, we have seen that methods trained on a dataset perform well on the test set that is from similar domain. Since our method relies on question-answering, it generalizes well across different domains.

4. Dataset

We have used supervised machine learning and deep learning methods that require training for our baselines. For our baselines, to test how well they generalize, we have trained our models on the WELFake dataset 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. We didn't use any dataset for our zero-shot fake news detection model.

5. Methodology

5.1. Baselines

We have used different baseline models for the fake news detection task.

TF-IDF vectorization with ML classifiers: 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.

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

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.

5.2. Proposed Method

The problem statement of detecting zero-shot fake news necessitates the development of an architecture capable of harnessing the pre-existing state-of-the-art models. Our objective is to create a conclusive model that can accurately determine the credibility of a given news article as either trustworthy or untrustworthy without undergoing any training phase.

We have divided our model into multiple phases. An article will be passed through these stages and its trustworthiness will be decided based on the evaluation score. The comprehensive architecture is illustrated in Figure 1. The following subsections explain each phase and the objectives we aim to achieve.

5.2.1 Question-Answer generation

Question-Answer Generation is Natural Language task that generates questions and answers from a given article or statement. This task requires the model to comprehend the article and target the most important and meaningful sentences which can be used to generate questions and answers which summarise the sentence

The Question and Answers are generated using an answer-aware question generation model, which has been fine-tuned using the t5-base model.

5.2.2 Web Scrapping

We use the headline of the news article, search it on Google, and scrape the most relevant articles from credible sources. The idea here is that the articles on the same headline from credible sources would be trustworthy. We make a list of the most relevant articles and store them.

5.2.3 Answer Generation Model

Answer Generation is a subtask of NLP that aims to generate natural language responses to a given question within the context of the article. The Answer Generation task requires the model to understand the context of the input question, interpret it, and generate a response that is grammatically correct and semantically meaningful.

We use a DistilBERT model, fine-tuned on the SQuAD dataset, to generate answers to the questions using the verified news articles as context. We give the questions generated from the QA generator on the input news article as input and the verified news articles as context to the answer generation model to generate answers.

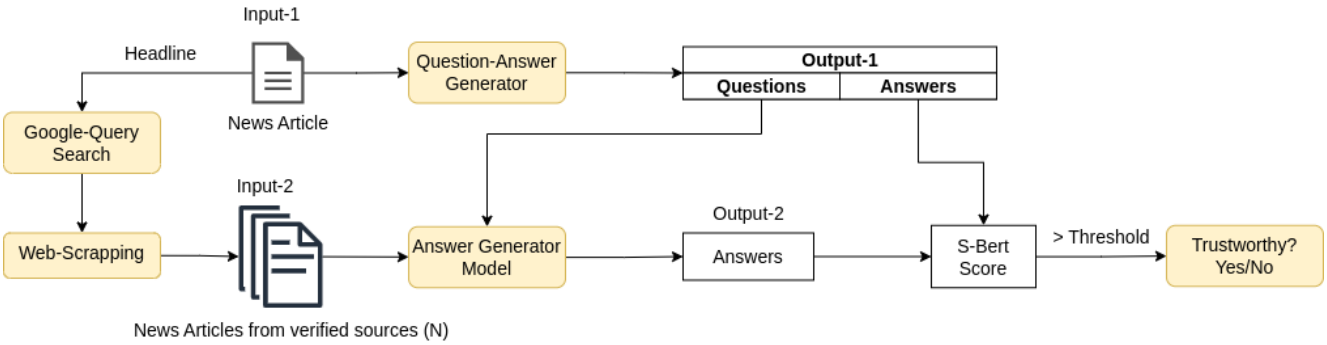

Figure 1. The pipeline for Zero-Shot fake news detection

5.2.4 Evaluation

To label an article as trustworthy or untrustworthy, we will match the answers generated for the input article with the answers generated for the relevant articles scrapped from the web. Here we can use different evaluation metric such as BLEU score, Rougue score, METEOR score or BERT score. Depending on the distribution of the score, we will assign a threshold. A score above the threshold will mark the article as trustworthy and a score below the threshold will mark it as untrustworthy.

6. Experiments

6.1. Baselines

6.1.1 TF-IDF vectorization with ML classifiers

The result of TF-IDF vectorization with ML classifier is 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

6.1.2 Bert Embeddings with ML classifier

We have applied different machine-learning models to our dataset, and the results are captured in table 2

6.1.3 Bert Classifier

We used a pre-trained Distilbert tokenizer to generate train, test and validation embeddings. The model is trained using

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 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 captured in table 3:

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

Table 3. Baseline result using Distilbert classifier

We have also tried using pre-trained Distilbert and fine tuned on our dataset and the results are captured in table 4

Method	Train F1-score	Test F1-score
Distilbert classifier	0.32	0.35

Table 4. Baseline result using fine-tuned Distilbert classifier

6.2. Proposed Method

In our work, we used a Question-Answering generation module to generate questions on the input data and generate answers. This is followed by searching the news article's title on the web and scraping the web for the top three news articles. We then use the question previously generated and find the answer via majority voting among the three news articles. The answers are compared between the two methods with S-Bert similarity score, and if they are above the

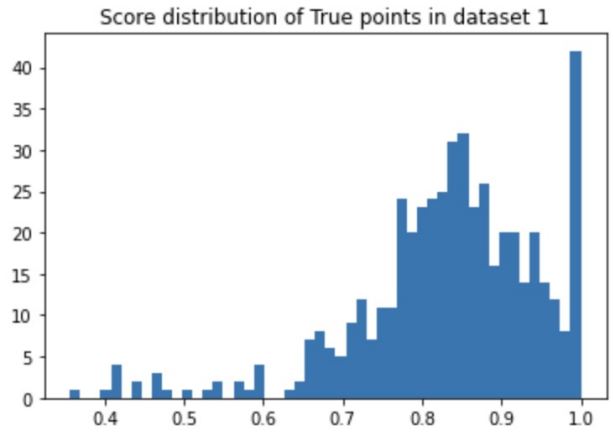

Figure 2. The score distribution of true points on dataset 1

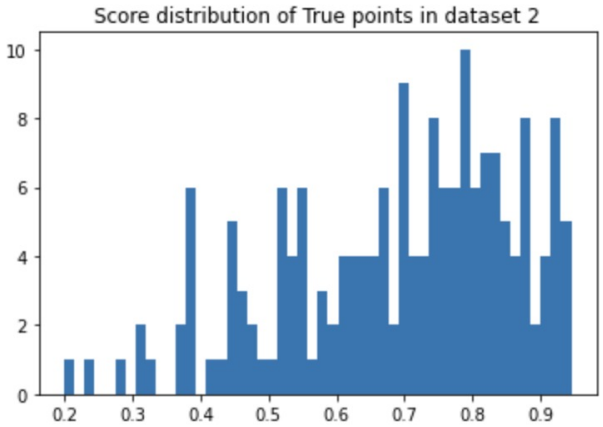

Figure 3. The score distribution of true points on dataset 2

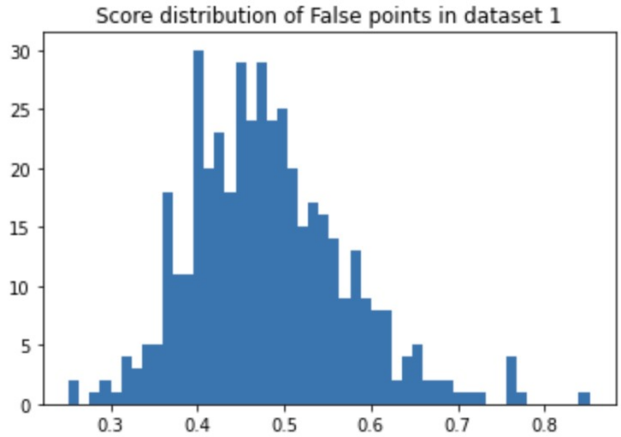

Figure 4. The score distribution of false points on dataset 1

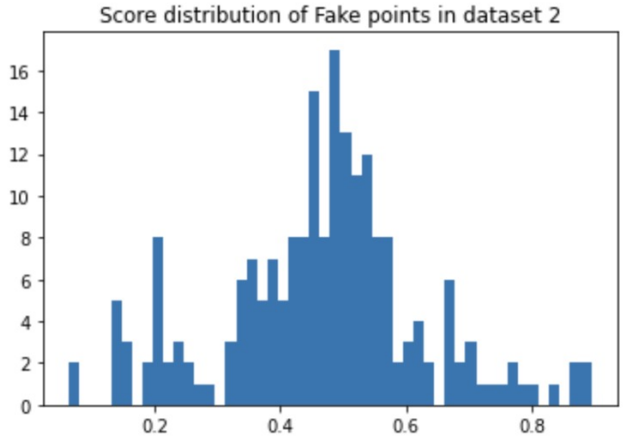

Figure 5. The score distribution of false points on dataset 2

threshold of 0.65, we classify the news article as true. Else we classify it as fake. We have tested our model on a combination of the train set and test set used at the time of baseline evaluation. This is due to the fact that our Zero-shot classifier has no training phase. The distribution score of the true and false news articles on the combined testing data is shown in Figure 2, 3, 4 and 5.

The combined test f1 score is as follows: As we can see,

Method	Combined test F1-score
Zero-shot classifier	0.8998

Table 5. Final result using Zero-shot classifier

our model performs far better than our baselines which can only attain a test F1 score of 0.35 after extensive training, which is far less compared to the combined test f1-score of nearly 0.90. Our model generalizes well on an unseen dataset which was not the case with baselines trained in a supervised manner.

7. Code

The code can be found here. The website can be accessed from here

8. Limitations

We have faced the following challenges while implementing our project

- When scraping data from websites, we found a limit to the number of requests that can be made before the website blocks further requests. This can hinder data collection and slow down the research process.
- Deploying websites with pre-trained models can make the website slow and less responsive. This is because these models can be quite large and computationally

expensive, which can cause delays in processing and returning results to users.

- The approach is limited to text input only.

9. Conclusion

We have found through experimentation that supervised methods do not generalize well when there is a difference between the distribution of train and test samples and, therefore, do not perform well on the task of fake news detection. The supervised machine learning and deep learning methods involve annotated data and extensive training time. On the other hand, Zero shot fake news detection model can perform very well on recent news articles without any training phase. The model generalizes well on the unseen classes without requiring annotated data and shows a big improvement in F1 scores with comparison to baselines. Our robust model makes real-time predictions within thirty seconds and thus has practical application.

10. Future Work

Some of the possible future work that could be done to improve upon the current approach are as follows:

- Predicting news based on the article's URL rather than just the headlines and article text could be a valuable addition to the model. This would involve extracting data from the URL, such as the source or keywords, to gain further insight into the nature of the article and improve prediction accuracy.

- As visual media becomes increasingly prominent in news reporting, it is important to be able to distinguish between authentic and fake images. By training a CNN on image datasets, it could be possible to accurately identify images being used misled or deceptively. These potential areas of future work would involve expanding the scope of the research to include additional data sources and types of analysis. By doing so, we could improve the accuracy and comprehensiveness of our predictions and contribute to developing more robust and reliable news analysis models.

11. Contribution

Every team member made equal contribution.

References

[1] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question generation for question answering. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 866–874, 2017. 2

[2] 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. 1

[3] 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. 2

[4] Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. Unsupervised question answering by cloze translation. *arXiv preprint arXiv:1906.04980*, 2019. 2

[5] 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. 2

[6] Álvaro Ibrain Rodríguez and Lara Lloret Iglesias. Fake news detection using deep learning, 2019. 2

[7] 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. 2

[8] 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. 1