Fake News Detection in Social Media using Cross and Bi-encoders-based approach

Rayees Ahmad Dar
Department of Computer Science
University of Kashmir, Srinagar, J&K, India
rayees.csscholar@kashmiruniversity.net

Rana Hashmy
Department of Computer Science
University of Kashmir, Srinagar, J&K, India
ranahashmy@gmail.com

Abstract— Despite the ease of access to social media, which has become the primary source of news for most people in today's society, it can sometimes lead to calamitous consequences. It may be exploited in a biased manner for the benefit of certain individuals by spreading or sponsoring misinformation due to its unauthenticated nature. Due to the power of social media, misinformation can spread more rapidly than viruses during pandemics such as COVID-19. Therefore, combating the epidemic of misinformation is equally important as combating the epidemic itself, as misinformation can prove to be lethal and many people tend to believe in it without conducting thorough research. The fight against misinformation has been ongoing for a long time, even before the advent of social media. Manual fact-checking is the most commonly used method on social media and other digital platforms. While reliable and accurate, manual fact-checking is also time-consuming and cumbersome, making it infeasible. As a result, automatic fact-checkers appear to be the only viable solution to this problem. Various machine learning and artificial intelligence-based approaches have been developed for this purpose, with transformer-based models like BERT being the state-of-the-art approach for this problem. These models are based on encoders and decoders. In this paper, we propose a bi- and cross-encoder-based model that has demonstrated promising results when used on the Contraint@AAAI 2021 COVID-19 fake news detection dataset. Our experiments show that this simple technique can be a viable solution for many tasks without the need for pretraining or fine-tuning, which can be computationally costly.

Keywords—fake news detection, covid19, cross-encoders, bi-encoders, semantic search, sentence embeddings, sentence transformers

I. INTRODUCTION

According to many definitions [1][2][3], "fake news" is "a news story or message published and disseminated through the media that contains false information." The spread of fake news is at its worst during major events because people are more prone to accept erroneous information owing to a lack of accessible research and knowledge. False information has the potential to harm people and communities at large, and the unauthenticated nature of OSNs only helps to disseminate such stories further. Since humans cannot keep up with the volume of work involved in manually verifying and labeling data, promising automation solutions based on machine learning and data mining have been introduced.

In this research, we offer a rank-rerank method that utilizes semantic similarity as well as bi-encoders and cross-encoders that are based on the popular sentence transformers. To test our model, we look at English tweets regarding the recent COVID-19 virus from the Contraint@AAAI 2021 COVID-19 false news detection dataset [4]. Our approach uses semantic similarity to rate the top N tweets using bi-encoders and then re-rank them using cross-encoders, bypassing computational bottlenecks in the process. The K tweets with the highest similarities form the basis of our categorization forecast.

The bi-encoder model is an additional pre-trained language model approach that uses a Siamese network and a similarity score [5] to apply self-attention individually to each text before comparing them. One of the most well-liked pre-trained language model-based techniques, the cross-encoder model, encodes both texts in parallel with complete self-attention [5]. To solve the challenge of comparing two sentences, we use a simple architecture that combines cross-encoder and bi-encoder representations. Earlier approaches for identifying fake news that relied on machine learning and data mining are discussed in Section 2. We discuss the dataset we utilized to test our design in Section 3, and then we outline our suggested architecture and experimental setup in Section 4. The final findings and discussion are given in Section 5.

II. LITERATURE SURVEY

Various perspectives and models, ranging from traditional Machine Learning to more advanced Neural Network-based models, have been applied to identify false news. Extensive research in the field of misinformation detection has resulted in a plethora of model- or data-driven strategies-based solutions. For instance, Multinomial Naive Bayes [6] outperforms most classic machine learning models (e.g., Support Vector Machine, Logistic Regression, Decision Trees, AdaBoost, and Naive Bayes). In [7], deep learning methods were evaluated for their ability to detect false news. Several DL models were trained on the Contraint@AAAI 2021 COVID-19 dataset for detecting false news. The models included LSTM, CNN, HAN, bi-LSTM+attention, DistilBERT, and BERT-base. The study approached the problem as a binary classification task, primarily focusing on news content (text). The authors attempted to enhance the pre-trained BERT and DistilBERT models by pre-training them on the tweet corpus related to Covid-19, which resulted in superior performance compared to models trained solely on the dataset. COVID-Twitter-BERT

outperforms other approaches when combined with a BERT-cased model, and HAN also outperformed models without transformers.

The Transformer [8] architecture forms the basis for the majority of cutting-edge approaches to detect false news used today. These models outperform previous non-transformer-based models because they employ a self-attention technique where each word in a sentence is weighted according to its significance and are pre-trained on a large dataset. BERT [9], one of Google's transformer-based pre-trained language models, has 345 million trainable parameters (BERTLARGE) and is a cutting-edge architecture for text classification and other downstream tasks. [10] Combining CNN with BERT to reduce ambiguity, the authors propose fakeBERT, a BERT-based deep learning approach. [11] They also proposed an ensemble of (BERT, ALBERT, and XLNET) and tested this model on the Contraint@AAAI 2021 COVID-19 dataset for the detection of false news. [12] proposes TRANS-ENCODER, an unsupervised sentence-pair model that integrates bi- and cross-encoders to discover more accurate sentence representations. The model employs a pre-trained language model and alternates between the two formulations to enhance performance by generating pseudo-labels. TRANS-ENCODER outperforms cutting-edge unsupervised sentence similarity models. Using word embeddings and LSTM neural networks, the authors of [13] present an approach for detecting false news. The approach is evaluated on two publicly available datasets and outperforms existing models. The authors emphasize the significance of pre-trained word embeddings for the detection of false news.

In [14], the issue of false news on Twitter is examined, and the performance of various embeddings, such as BERT and conventional methodologies, is compared. Using Python, a highly competent forecasting model was created, evaluated, and tested on COVID-19 text data containing false news. [15] proposed various Transformer and recurrent models with contextual word embedding models. The proposed model is evaluated using a different loss function than Binary cross Entropy and is considered a sequence classification task in natural language processing. The use of domain-specific language models and a custom loss function produces the highest average weighted F1-score. Finally, in [16], various supervised text classification algorithms, such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT), are evaluated on the Contraint@AAAI 2021 Covid-19 Fake news detection dataset to combat the problem of fake news on social media. The study also evaluates the significance of unsupervised learning using language model pre-training and distributed word representations derived from the unlabeled Covid tweets corpus; the significance of unsupervised learning is also evaluated.

III. DATASET

The Contraint@AAAI 2021 COVID-19 fake news detection dataset comprises 10,700 social media posts and news articles related to COVID-19, both genuine and fraudulent, that have been carefully annotated. The dataset has been divided into training, validation, and test sets, with proportions of 3:1:1. The class label distribution is well balanced across all sets. Real news items account for 52.34 percent of the samples, while false news items account for 47.66 percent. Genuine news items were sourced from a wide range of trustworthy news sources, while false news items were debunked by third-party fact-checking websites such as NewsChecker and PolitiFact.

IV. METHODOLOGY AND MODEL ARCHITECTURE

In this study, we approach the problem of detecting fake news from a binary text classification perspective, categorizing each news item as either true or fake. Our proposed methodology involves pre-processing the unprocessed text, tokenizing each sentence, encoding them using a bi-encoder model, refining the initial ranking using a cross-encoder model, and finally matching labels and averaging results to obtain a final prediction output.

A. Pre-processing

The unprocessed text is first subjected to rudimentary preprocessing by removing unnecessary URLs, HTML elements, spaces, special characters, emoticons, and stop words. The tweet-preprocessor module in Python is used to eliminate extraneous data from tweets.

B. Tokenization

Each sentence is tokenized using the corresponding tokenization technique for the model that will be used in the future. This is necessary to ensure that the model receives tokens with the correct structure and special tokens.

C. Modelling

We use a bi-encoder and a cross-encoder to classify news items as true or fake. Here's a step-by-step explanation of our approach:

1) *Bi-encoder:* We use the well-known sentence transformer architecture for bi-encoders to produce sentence embeddings quickly. Each sentence is encoded and mapped to a shared embedding space, allowing the gaps between them to be calculated. Bi-encoding is more efficient because encoded sentences can be cached and reused, and bi-encoder outputs can be used as sentence embeddings for subsequent tasks. We experimented with the "paraphrase-mpnet-base-v2" sentence transformer model and applied mean pooling to obtain fixed-size embedding vectors. Bi-encoders were used to rank news items by similarity, although they were not always accurate.

2) Cross-encoder: Cross-encoders combine two sequences and provide them to the sentence pair model, which is typically constructed on top of a Transformer-based language model such as BERT or RoBERTa. By modeling which elements of one sequence correspond to which elements of the other, the attention centers in the Transformer can calculate an accurate classification/relevance score. However, cross-encoders require a lot of computational power because they must generate a new encoding for each pair of input sentences. For tasks like information retrieval and grouping

that require numerous pairwise comparisons, cross-encoding is impractical. Furthermore, cross-encoding of pre-trained language models requires constant fine-tuning on annotated data. In our experiment, we used the "cross-encoder/nli-distilroberta-base" model.

3) Initial ranking: As the first step, we obtain sentence embeddings of the train and test data from the predefined bi-encoder model. These embeddings are compared using cosine similarity, and an initial ranking is established. However, this ranking is not very accurate because bi-encoders do not take into consideration the interactions between sentences.

4) Refining rankings: Therefore, these rankings are refined using a cross-encoder model. The cross-encoder model rescores the top N (N=50) items, and we re-rank them based on the new cross scores, which are either a contradiction, entailment, or neutral score. We average out these three scores to obtain the final ranking.

5) *Matching labels and final prediction*: We finally match the labels of the top K items and average out their results to obtain a final prediction output. We experimented with different N and K values and concluded that they have a slight effect on the final outcome for this specific problem. In our experiment, we set N=50 and K=3.

The general overview is illustrated in Fig. 1, and a pseudo code representation of the approach is shown in Algorithm 1

Algorithm 1. CROSS-BIENCODER MODEL

1. Cosine Similarity:
   Cos_sim = cos(θ) = (A ∗ B) / (||A|| ∗ ||B||)
   where A and B are the sentence embeddings for the train and test data respectively, and ||A|| and ||B|| represent their respective magnitudes.

2. Scoring of top N items (N=50):
   a) For each sentence in the test set:
      top_N = get_top_N(Cos_sim)
      cross_score_N = []
   b) For each sentence in top_N:
      cross_score_N. append (cross encoder (sentence, test_sentence))
      sorted_N = sort (cross_score_N, reverse=True)
   c) final scores = []
      For i in range(N):
      score = (sorted_N[i][0] + sorted_N[i][1] + sorted_N[i][2])/3
      final_scores. append ((top_N[i], score))

3. Matching Labels with top K items(K=3) and Final Prediction Output:
   a) match = 0
      count = 0
   b) For each sentence in the test set:
      true labels = []
      For i in range(K):
      true_labels. append (train data[final_scores[i][0]][1])
   c) if test_data[sentence][1] in true_labels:
      match += 1
      else:
      match -= 1
   d) if match > K/2 and test_data[sentence][1] ==’True’:
      prediction = 'True'
      count += 1
      elif match > K/2 and test_data[sentence][1] ==’False’:
      prediction = 'False'
      count += 1
4. accuracy = count/len(test_data)

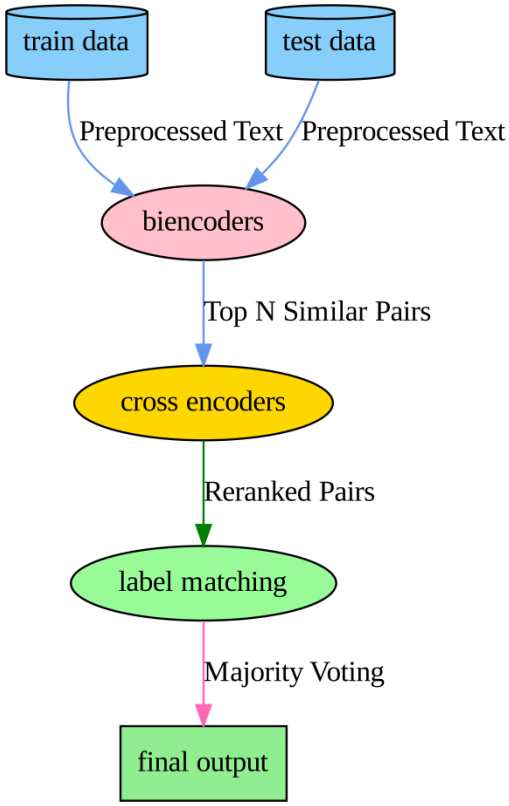

Fig. 1. Methodology of the Proposed model

V. RESULTS AND EXPERIMENTS

We constructed and evaluated a model for detecting false news using data related to COVID-19. The confusion matrix, a tabular representation of the prediction model's results, was used to evaluate the model's performance. The confusion matrix displays the outcomes of a predictive model, including which groups are correctly forecasted, which are erroneously predicted, and what categories of errors are occurring. Figure 2 depicts the confusion matrix for the test data.

To evaluate the proposed architecture, we used metrics such as Accuracy, Precision, Recall, and F1 Score. Accuracy represents the total number of true predictions made by the model out of all predictions made. Precision measures how many of the positive predictions made by the model were actually correct, while Recall measures how many of the true positive cases were correctly identified. F1 Score is a combination of Precision and Recall, and is the harmonic mean of the two metrics.

Table 1 demonstrates that our approach outperformed default models such as Support Vector Machine (SVM). The results highlight the superior performance of our approach, which leverages pre-trained word embeddings to capture the nuanced semantic meanings of textual data. On the validation

set, our method achieved an accuracy of 0.9734, a recall of 0.9753, a precision of 0.9716, and an F1 score of 0.9735, while on the test set, the corresponding values were 0.9677, 0.9675, 0.9679, and 0.9677. In contrast, the best baseline model (SVM) obtained an accuracy of 0.9519 and F1 score of 0.9570. The results demonstrate that our simple and resource efficient method is highly effective at detecting COVID-19-related false news. In this study, we compared the performance of our proposed cross-biencoder approach to that of several baseline models, including logistic regression (LR) and gradient boosted decision trees (GBDT) from [4] and support vector machines (SVM) from [17] that used linguistic features of tweet text.

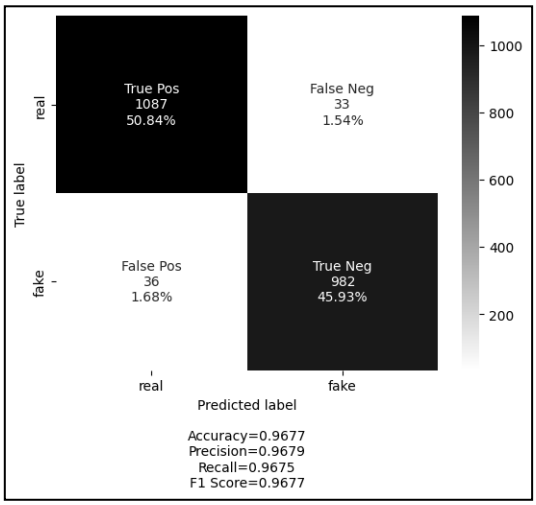

Fig. 2. Confusion matrix of proposed Cross- Biencoder architecture on test data.

LR is a basic linear classification model that is widely employed. GBDT is an ensemble technique that integrates multiple weak decision trees to create a robust classifier. SVM is a prominent kernel-based method that separates classes in a high-dimensional space using a hyperplane. These models have proven effective for various classification tasks, including the detection of false news. We also experimented with a transformer-based classifier, specifically a DistilBERT classifier, whose performance on various NLP tasks was remarkable. DistilBERT is a condensed variant of the BERT model that accomplishes comparable performance but has a quicker inference time and a reduced memory footprint.

Using a bi-encoder and a cross-encoder in a dual-encoder architecture, our cross-biencoder method captures the semantic relationship between tweets more accurately. Using a cross-encoder model that integrates sentence interactions, we further enhance ranking accuracy by refining the initial ranking generated by the bi-encoder. Experiments revealed that our cross-biencoder method outperformed all baseline models, attaining 96.77 percent accuracy. This performance is considerably more accurate than LR (91.96%) and GBDT (86.96%), and marginally more accurate than SVM (95.1%). The DistilBERT classifier attained 96.58 percent accuracy.

Overall, our approach demonstrates that adding a cross-encoder to refine the ranking generated by a bi-encoder can enhance the performance of false news detection. Experiments indicate that the cross-biencoder strategy is a promising approach for this endeavour.

TABLE I. PERFORMANCE COMPARISON OF THE PROPOSED METHOD VERSUS OTHER MODELS ON THE TEST SET OF THE CONTRAINT@AAAI 2021 COVID-19 DATASET.

Model	Accuracy	F1-Score
GDBT [4]	86.96	86.96
LR [4]	91.96	91.96
SVM using linguistic features [17]	95.19	95.70
distilBERT Classifier	96.58	96.73
Proposed approach (Cross and Bi-encoder)	**96.77**	**96.77**

Statistical tests confirmed that the performance differences between our approach and the baseline models were statistically significant, with p-values less than 0.01. The performance can be attributed to the combination of strengths of bi-encoders and cross-encoders. Specifically, we initially used bi-encoders to identify the top N most similar tweets based on semantic similarity, which enabled us to efficiently capture similarities between different tweets. We then used cross-encoders to further refine the results from the bi-encoders by leveraging the full context of the dataset. This allowed us to achieve a more accurate ranking of the top tweets related to COVID-19, while still being computationally feasible.

VI. CONCLUSION

In conclusion, we have presented a novel approach for fake news detection using a resource-frugal technique based on semantic similarity. The proposed approach employs Bi and Cross encoders based on sentence transformers to extract the similarity among the news items. The results demonstrate that our approach outperforms computationally intensive models and is highly effective at detecting COVID-19-related false news. Moreover, we have further improved the performance of our approach by utilizing the last hidden layer of the fine-tuned bi-encoder model as our sentence embeddings. Our future research will be focused on incorporating other contextual features, such as user, engagement, and propagation features, to improve the overall performance of our approach. It would also be interesting to investigate how well our method performs on other generic datasets associated with false news. Overall, our approach presents a promising and efficient solution for detecting false news, which can be easily adopted in real-world applications.

REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media : A Data Mining Perspective,” vol. 19, no. i, pp. 22–36, 2016.

[2] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,”*ACM Comput. Surv.*, vol. 53, no. 5, 2020, doi: 10.1145/3395046.

[3] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News," pp. 7–17, 2016, doi: 10.18653/v1/w16-0802.

[4] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, and I. S. City, "Fighting an Infodemic :COVID-19 Fake News Dataset".

[5] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, "Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring," pp. 1–14, 2019, [Online]. Available: http://arxiv.org/abs/1905.01969

[6] M. Singh, M. Wasim Bhatt, H. S. Bedi, and U. Mishra, "WITHDRAWN: Performance of bernoulli's naive bayes classifier in the detection of fake news," *Mater. Today Proc.*, no. xxxx, 2020, doi: 10.1016/j.matpr.2020.10.896.

[7] N. Aslam, I. Ullah Khan, F. S. Alotaibi, L. A. Aldaej, and A. K. Aldubaikil, "Fake Detect: A Deep Learning Ensemble Model for Fake News Detection," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/5557784.

[8] A. Vaswani, "Attention Is All You Need," no. Nips, 2017.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[10] R. K. Kaliyar, A. Goswami, P. Narang, R. Kumar, K. Anurag, and G. Pratik, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, 2021, doi: 10.1007/s11042-020-10183-2.

[11] S. Gundapu and R. Mamidi, "Transformer based Automatic COVID-19 Fake News Detection System," pp. 1–12, 2021, [Online]. Available: http://arxiv.org/abs/2101.00180

[12] F. Liu, Y. Jiao, J. Massiah, E. Yilmaz, and S. Havrylov, "Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations," pp. 1–18, 2021, [Online]. Available: http://arxiv.org/abs/2109.13059

[13] M. Sahin, C. Tang, and M. Al-Ramahi, "Fake News Detection on Social Media: A Word Embedding-Based Approach," *AMCIS 2022 Proceedings*, Aug. 2022, Accessed: Apr. 16, 2023. [Online]. Available: https://aisel.aisnet.org/amcis2022/vcc/vcc/4

[14] S. P. Devika, M. R. Pooja, M. S. Arpitha, and V. Ravi, "BERT Transformer-Based Fake News Detection in Twitter Social Media," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, no. August, IEEE, 2023, pp. 95–102. doi: 10.1007/978-981-19-6004-8_8.

[15] A. Hande, K. Puranik, R. Priyadharshini, S. Thavareesan, and B. R. Chakravarthi, "Evaluating Pretrained Transformer-based Models for COVID-19 Fake News Detection," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021, pp. 766–772. doi: 10.1109/ICCMC51019.2021.9418446.

[16] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, "Evaluating Deep Learning Approaches for Covid19 Fake News Detection," *Commun. Comput. Inf. Sci.*, vol. 1402 CCIS, pp. 153–163, 2021, doi: 10.1007/978-3-030-73696-5_15.

[17] T. Felber, "Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task," pp. 2–5, Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.03717