Ensembling of Various Transformer Based Models for the Fake News Detection Task in the Urdu Language

Sakshi Kalra[a], Preetika Verma[a], Yashvardhan Sharma[a] and Gajendra Singh Chauhan[b]

[a]*Department of Computer Science and Information Systems, Birla Institute of Technology and Science Pilani, Pilani Campus, Rajasthan, India*

[b]*Department of Humanities and Social Sciences, Birla Institute of Technology and Science Pilani, Pilani Campus, Rajasthan, India*

Abstract

The spread of misinformation has become a severe issue affecting society. Inaccurate information has enormous potential to cause real-world impacts. Developing algorithms to detect fake news automatically will be very useful in preventing unnecessary panic and damage caused by rumors. This fake news problem is present for all languages, and it becomes crucial to solve it for languages other than English, with scarce datasets. This paper aims to tackle the problem of automatic fake news detection in Urdu, a low-resource language. FIRE-2021 has provided the Urdu dataset used in this paper. We fine-tuned monolingual and multilingual transformers. After searching for hyperparameters, we tried ensembling our models. We submitted our model for the UrduFake task, and it achieved an accuracy of 0.596 and an F1-macro score of 0.449.

Keywords

Fake News Detection, Natural Language Processing, Label Classification, Various Transformers, Ensemble Techniques

1. Introduction

In 2016, Fake News was such a generally used word that the Oxford Dictionary appended this word to their official list, with the description: "false stories that appear to be news spread on the internet or using other media, usually created to influence political views or as a joke."Fake news dissemination had been a concerning issue since the invention of the printing press in the 15th century. The spread of fake news and misinformation has brought disastrous consequences many times. An example is the recent Facebook post on 19th September 2021, which claimed that the Canadian prime minister and his wife faked their covid-19 vaccinations on live television[1]. This created panic and was widely shared. Later, it was found to be false. Therefore, it is crucial to develop an algorithm to curb the spread of fake news, creating panic and confusion among people.

There have been many attempts for automatic fake news detection in English, Chinese,

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

French, and other high-resource languages. Many Natural Language Processing techniques are used to detect fake news from the English language. Various Transformer based models such as BERT [2], XLnet [3], DistilBERT [4], etc., have been designed to get the word embeddings and word dependencies for the English language text. Very little work has been done on Urdu even though it is spoken as a first language by nearly 70 million people and as second by 100 million people. Urdu is a low-resource language, and there is a scarcity of publicly available datasets for NLP tasks using this language. It is the most popular language in Pakistan, and it has around 100 million speakers across the world [5]and is widely spoken in the Indian subcontinent. It still does not have a lot of language processing tools like parsers and corpora. Many researchers have tried to target the Urdu language. A shared task [6],[7] on fake news detection in the Urdu language has been started to tackle the fake news detection problem.

Fact-checking websites like Politifact [8] have come up for English which checks the accuracy of statements. Researchers have divided fake news into seven categories- false news, polarised content, satire, misreporting, commentary, persuasive information, and citizen journalism [9]. It has also been found that fake news articles are less factual, less grammatically correct, and have more emotionally charged claims. Analyzing the linguistic features of text has also been proved to help in classification.

In this paper, we target the Urdu Fake News Detection task by participating in NEWUrdu-Fake@FIRE2021[10]. We experimented with various multilingual transformer-based models individually and tried to get results by ensembling these different transformer-based models; We are getting better results without the ensembling approach.

2. Related Work

Most of the proposed approaches have been used for the English language [11], and there have been some efforts for Spanish[12], German[13] as well. Data augmentation with machine translation is used to tackle the problem of small datasets[14]. Newly annotated data in Urdu is generated by translating the English dataset introduced. Google Translate is used for this purpose. It is found that the classifier trained on the original Urdu dataset showed better results than the augmented and translated datasets. One reason for this was that the MT translation quality between Urdu and English was inferior. Recent studies have extracted different features from Urdu text and fed them into supervised classification models like logistic regression, k-nearest neighbors, random forests, and support vector machines. These features try to model the news articles mathematically. Linguistic features include the total number of words, frequency of function words and phrases, parts of speech tags, unique word count, syntactic dependencies, clauses, punctuation, etc. Domain-specific linguistic features precisely align with the news domain and include quoted words, external links, etc. Approaches using transformer-based models have been utilized to detect misleading news articles [15] and false covid related news[11].

3. Dataset

The dataset used for this task is 'Bend the Truth.' This binary annotated corpus contains articles from six domains: technology, education, business, sports, politics, and entertainment [5]. This is the only annotated corpus available for detecting false news in Urdu. The real news articles are collected from different mainstream news websites like BBC News, CNN Urdu, Daily Pakistan, urdupoint, etc. A newspaper library is used for web scraping. The data is collected and annotated manually. If a legitimate website is published, the source is mentioned, or the same news is found on other reliable websites, the article is real. Different lengths of texts are collected during the data collection process. For fake news collection, professional journalists are hired to write news articles for all the domains. They are asked to avoid unintentionally introducing any patterns in the fake news articles. After collection, all the articles are reread to remove typing errors and word misuse. Data is cleaned - Latin alphabet characters are removed, Eastern Arabic-Indic numerals are converted to Western Arabic numbers. Paragraphs are ramified into sentences on Urdu end markers. The training dataset had 750 real articles and 550 fake. The evaluation is done on an unknown dataset of 300 articles.

4. Proposed Techniques and Algorithms

Transformer-based implementation involves pre-training, which is followed by fine-tuning. The model is trained on large language datasets(monolingual) or datasets in multiple languages(multilingual) for the first part. All the initialized parameters are fine-tuned using the labeled data from the given dataset. The code is available in the github repository. [1] Only the encoder part of the transformer architecture is used for getting the word-embeddings. One additional output layer is added to calculate the probability for real and fake classes. Various word embeddings models have been used and listed below:

- RoBERTa: This model is trained on Urdu news data from Pakistani newspapers. It is built on BERT, trained with larger mini-batches and learning rates.
- ALBERT: Trained on Urdu datasets, but it gave inferior results.
- XLM–RoBERTa: It is trained in 100 different languages and has the exact implementation as RoBERta. It is pre-trained on more than 2TB of CommonCrawl data. The idea is to map any language to a language-agnostic vector space where all the languages for the same input would point to the same area.
- Multilingual BERT: This is pre-trained in 104 languages. The texts are lowercased and tokenized using WordPiece and a shared vocabulary size of 110,000. The languages with a larger Wikipedia are under-sampled, and the ones with lower resources are oversampled.

4.1. Hyperparameter Description

Raytune library is used for hyperparameter tuning. The original head of the models is removed and replaced with a classification head so the output would be for two classes. The training dataset is used for fine-tuning the pre-trained models available on huggingface. Alberta has

[1] https://github.com/Kalra-Sakshi/URdu-FND.git

Table 1
Eight Trials to Find the Optimum Hyperparameters

Trial no.	Learning rate	No. of epochs	Batch size
1	7.16125e-06	2	2
2	5.24269e-06	4	2
3	8.15284e-06	3	4
4	1.00742e-05	2	4
5	1.35181e-055	2	2
6	2.04561e-05	4	2
7	1.42903e-05	4	4
8	7.91852e-06	3	8

not ensembled due to the poor results. Monolingual RoBERTa is giving the best results. Both the multilingual models are poor at detecting the fake class. For the final run, we submitted roberta-urdu-small, which is pretrained on Urdu news corpus since it alone gives better results than any ensembled combination. The training data has normalized using the normalization module from urduhack library to eliminate the characters from other languages like Arabic. We used the same tokenizer for fine-tuning the models. Eight trials are conducted to find the optimum hyperparameters. Search space is defined, and hyperparameter combinations are randomly selected. Adam optimizer with weight decay is used for the model optimization. Table-1 lists the total number of 8-trials, conducted to find the optimum hyperparameters.

4.2. Ensembling of the Models

We tried to ensemble the three transformer-based architectures[11]. We have not considered Alberta as it gives inferior results. The ensembled model computes the average of all softmax values after extracting the softmax probabilities from each model. In our problem, the results from the three transformers individually are not close to each other. This model performs better than multilingual-bert and xlm-roberta. Monolingual roberta gives a better result on the validation dataset alone than any different model or ensemble combination. Figure-1 shows the transformer based ensemble model architecture.

5. Results and Evaluations

Different combinations are ensembled using soft-voting after the hyperparameter tuning to get the best results. The results on the validation dataset are listed in Table-2. Roberta is giving the best results for the fourth trial and the multilingual models for the seventh trial. Alberta is providing poor results and is not used for ensembling. We tried soft-voting for all combinations.

5.1. Error Analysis

The multilingual models are inferior at identifying fake texts and classified a lot of them as real. The reason would be due to the slight imbalance in the training data. The monolingual model

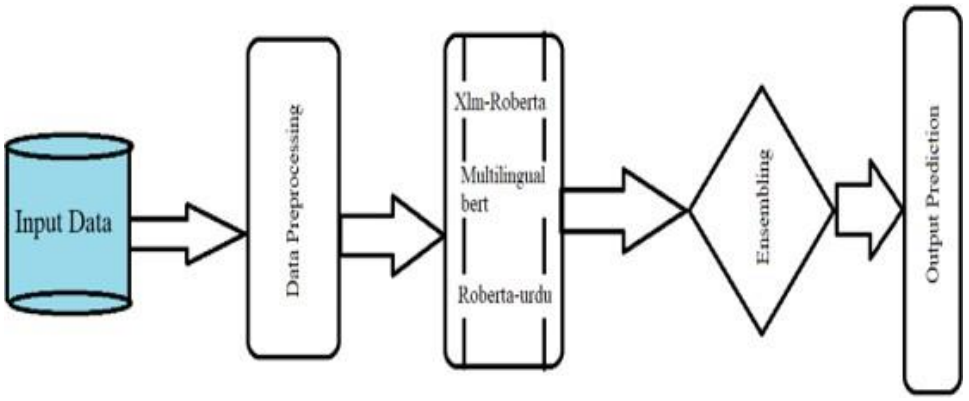

Figure 1 Transformer-based Ensemble Model Architecture

Table 2
Dataset statistics

Language	Training Data	Testing Data
RoBERTa-urdu-small	0.9083	0.9215
XLM-ROBERTa	0.8167	0.8260
bert-based-multilingual-case	0.8282	0.8442
alberta-urdu-large	0.6183	0.6621
ensembling all three	0.8625	0.8758

Table 3
Error Analysis Report

Fake	Fake	Fake	Real	Real	Real		
Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**F1 Macro**	**Accuracy**
0.266	**0.120**	**0.165**	**0.654**	**0.835**	**0.734**	**0.449**	**0.596**

alone performed slightly better than any ensembled combination. For the surprise dataset, our submitted run is unable to detect most of the fake news articles. Table-3 lists the error analysis report.

6. Conclusions and Future Work

Traditional Machine Learning-based approaches produced better results than a transformer-based approach. The training set is small, and even though the results are suitable for the validation set, the model cannot perform well on the surprise dataset. It is also unable to distinguish fake samples and gives very low precision, recall, and F1 score for the fake class.

Future work involves trying to extract features from the intermediate transformer layers. It would be interesting to try transfer learning for different languages—for example, training on an English dataset and testing on Urdu. We can try if the problem of insufficient data can be solved by using datasets in other languages and training on multilingual models. This idea is to turn any incoming language into a language-agnostic vector in a space where all languages for the same input would point to the same area.

References

[1] Politifact, fact checking website, URL: https://www.politifact.com/factchecks/2021/sep/24/facebook-posts/trudeaus-got-their-covid-19-shots/.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv 2018, arXiv preprint arXiv:1810.04805 (2021) 0–85083815650.

[3] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[4] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[5] M. Amjad, G. Sidorov, A. Zhila, H. Gómez-Adorno, I. Voronkov, A. Gelbukh, "bend the truth": Benchmark dataset for fake news detection in urdu language and its evaluation, Journal of Intelligent & Fuzzy Systems 39 (2020) 2457–2469.

[6] M. Amjad, G. Sidorov, A. Zhila, A. F. Gelbukh, P. Rosso, Overview of the shared task on fake news detection in urdu at fire 2020., in: FIRE (Working Notes), 2020, pp. 434–446.

[7] M. Amjad, G. Sidorov, A. Zhila, A. Gelbukh, P. Rosso, Urdufake@ fire2020: Shared track on fake news identification in urdu, in: Forum for Information Retrieval Evaluation, 2020, pp. 37–40.

[8] Politifact, fact checking website, URL: https://www.politifact.com/.

[9] Economic times, seven types of fake news URL: https://economictimes.indiatimes.com/news/politics-and-nation/ seven-types-of-fake-news-identified-to-help-detect-misinformation/ no-message-in-fake-news/slideshow/72106573.cms.

[10] Fire 2021, urdufake2021, URL: https://www.urdufake2021.cicling.org/home.

[11] S. Gundapu, R. Mamidi, Transformer based automatic covid-19 fake news detection system, arXiv preprint arXiv:2101.00180 (2021).

[12] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J. J. M. Escobar, Detection of fake news in a new corpus for the spanish language, Journal of Intelligent & Fuzzy Systems 36 (2019) 4869–4876.

[13] I. Vogel, P. Jiang, Fake news detection with the new german dataset "germanfakenc", in: International Conference on Theory and Practice of Digital Libraries, Springer, 2019, pp. 288–295.

[14] M. Amjad, G. Sidorov, A. Zhila, Data augmentation using machine translation for fake

news detection in the urdu language, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 2537–2542.
[15] H. Jwa, D. Oh, K. Park, J. M. Kang, H. Lim, exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert), Applied Sciences 9 (2019) 4062.

A. Online Resources

The implementation of different pre-trained BERT-models are available at

- Huggingface.