Fake News Detection Using Transformer-Based NLP Models

Jas Dinh

DATASCI 266 : Section 1

Abstract

This study investigates the effectiveness of BERT and RoBERTa, two transformer-based models, in the task of fake news detection. We evaluated these models on two datasets: Political Dataset which consists of articles from a world and political news, and General Dataset, which consists of articles from diverse sources. On Political Dataset, both BERT and RoBERTa achieved 100% F1-score, while logistic regression achieved 96%, suggesting that the models outperform baseline. When applied to General Dataset, the transformer models achieved F1-scores of 0.98 and 0.99, respectively, while logistic regression obtained 0.94. However, when the models trained on General Dataset were tested on Political Dataset, both transformers showed a significant drop in performance, with F1-scores of 0.57 for BERT and 0.36 for RoBERTa. These results highlight the challenges of model generalization across datasets and underscore the need for more diverse and representative training data. The study demonstrates the limitations of large pre-trained models in cross-dataset generalization, pointing to the importance of domain adaptation and robust validation for real-world applications of fake news detection systems.

Introduction

The widespread misinformation has become an increasing problem in our era of social media where fake news can travel quickly and have far reaching consequences. With NLP, the detection of fake news can be automated and stopped quickly before reaching a larger audience. However, this can be a challenge due to the subtle nuances between fake and real news that even humans have trouble differentiating. Not only does the model have to identify differences in tone and phrasing but also the presence of factual distortions.

While traditional machine learning approaches have been used for this task, transformer-based models like BERT have allowed natural language processing advanced contextual understanding of language. This study aims to utilize transformer-based models to classify news articles as real or fake.

Literature Review

Devlin et al. (2019) first introduced BERT which uses bidirectional attention to model context, allowing for text classification tasks such as fake news detection. Similarly, Liu et al. (2019) optimized BERT with RoBERTa using more robust training to improve its performance on downstream tasks.

In the application of transformers to fake news detection, Kaliyar et al. (2020) demonstrated the efficiency of BERT to capture nuanced patterns in social media contexts. This study aims to expand upon this research by focusing on richer data of news articles compared to short social media posts.

However, there are many challenges in fake news detection as shown by Zhou & Zafarani (2020) who provided a comprehensive survey on the many obstacles machine learning models

face in identifying misinformation. They emphasize a need for more robust models that can handle the subtle linguistic cues and factual inaccuracies in fake news.

Methodology

The two models used for comparison are BERT, a transformer model pre-trained by HuggingFace, and RoBERTa, an optimized version of BERT with improved pretraining for better contextual understanding. The evaluation metric chosen for fake news detection is F1-score which shows how the model performs in false positives and false negatives. The F1-score can also handle imbalanced datasets better than accuracy.

Regarding datasets, the models were trained on two datasets. They were evaluated on individual dataset performance and cross dataset performance. Both datasets contain news articles labeled fake or real and are available on Kaggle. The first dataset had less variability and focused more on world and political news while the second dataset had more diversity in topics. From here on, they will be referred to as Political Dataset and General Dataset respectively.

Basic preprocessing was performed to remove irrelevant columns, balancing the datasets by under sampling the majority class, and tokenizing the text of the articles using BERT and RoBERTa tokenizers.

In regards to basic model parameters, the models had:

Max Length Sequence: 128	Learning Rate: 0.00001
Batch Size: 16	Layer Freezing: 2
Epochs: 2	Optimizer: Adam

The structure of the experiment was to test the hypothesis that BERT and RoBERTa should outperform the baseline of a logistic regression model due to their ability to capture contextual information, but the RoBERTa's optimizations should provide slight performance gains over BERT.

The first experiment was to train the models on the Political Dataset and the General Dataset and evaluate their performance on the datasets individually. The second experiment was to evaluate the models trained on the General Dataset to be tested with the Political Dataset. The theory is that the models can train on diverse articles before testing its performance with a more niche category of articles. This will show the models' ability to generalize information.

Results

Political Dataset Results

Logistic Model Results:

	precision	recall	f1-score	support
0	0.98	0.95	0.96	470
1	0.95	0.97	0.96	428
accuracy			0.96	898
macro avg	0.96	0.96	0.96	898
weighted avg	0.96	0.96	0.96	898

BERT Model & RoBERTA Model Results:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	470
1	1.00	1.00	1.00	428
accuracy			1.00	898
macro avg	1.00	1.00	1.00	898
weighted avg	1.00	1.00	1.00	898

The results are the classification reports of the models trained on the Political Dataset. In the figures above, we can see that the logistic regression model performed at an F1-score of 0.96 and the BERT and RoBERTa models had an extremely high performance of 100%. This supports the hypothesis that BERT and RoBERTa can outperform the baseline logistic regression model, but it is not enough information to a difference between the two transformer models.

General Dataset Results

BERT Model Results:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	2078
1	1.00	0.97	0.98	1963
accuracy			0.98	4041
macro avg	0.98	0.98	0.98	4041
weighted avg	0.98	0.98	0.98	4041

RoBERTa Model Results:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	2078
1	1.00	0.99	0.99	1963
accuracy			0.99	4041
macro avg	0.99	0.99	0.99	4041
weighted avg	0.99	0.99	0.99	4041

These results are the classification reports of the models trained on the General Dataset. The logistic regression (which is omitted) had a performance of 0.94 F1-score while BERT had 0.98 F1-score and RoBERTa had 0.99 F1-score. These results are consistent with the previous and show the transformer models outperform the baseline but no significant difference between the two transformer models.

Models Trained on General Dataset and Tested on Political Dataset

BERT Model Results:

	precision	recall	f1-score	support
0	0.65	0.35	0.45	428
1	0.58	0.83	0.69	470
accuracy			0.60	898
macro avg	0.62	0.59	0.57	898
weighted avg	0.62	0.60	0.57	898

RoBERTa Model Results:

	precision	recall	f1-score	support
0	0.47	0.89	0.61	428
1	0.44	0.08	0.14	470
accuracy			0.47	898
macro avg	0.46	0.48	0.38	898
weighted avg	0.45	0.47	0.36	898

Here, these classification reports show the transformer models trained on the General Dataset but tested with data from the Political Dataset. A significant performance drop can be seen with cross dataset testing as BERT has an F1-score of 0.57 while RoBERTa has an F1-score of 0.36. Surprisingly, the BERT model shows better generalization and performs better than RoBERTa when tested with data from a different source.

Discussion

The results of this study reveal several important insights into the performance and generalization capabilities of logistic regression, BERT, and RoBERTa models in the task of fake

news detection. By analyzing their performance across two datasets—Political Dataset and General Dataset—we can draw meaningful conclusions about their effectiveness and limitations.

The baseline logistic regression model overall achieved high F1-score in both datasets which is surprisingly high for a simple model. The datasets likely have certain keywords or styles of writing that the model can easily capture. Both transformer-based models achieved high performance for both datasets and higher than the baseline. This shows the transformer-based models are able to evaluate nuanced distinctions in the articles that characterize fake news better than the logistic regression model which does not take in contextual understanding.

In the cross dataset testing, the models revealed stark differences in their ability to generalize. Both received significant reduction in performances compared to individual dataset testing and also showed a significant difference from each other. The BERT model performed better than RoBERTa on cross dataset testing despite both datasets focusing on fake news detection. The poor cross-dataset performance raises questions about the robustness of advanced models in real-world applications, where the training and test data are often not identical in structure, style, or content.

The differences in performance are most likely contributed to the differences in topic with the Political Dataset having more niche articles than the General Dataset so the types of language and patterns may differ. The significant performance drop is indicative of overfitting to the nuances of the dataset and therefore show their limited ability to adapt to new data.

On the General Dataset, RoBERTa showed a slight improvement over BERT, consistent with its design optimizations. However, the generalization gap was more pronounced for RoBERTa than for BERT. This raises an interesting question: could RoBERTa's enhanced ability

to learn dataset-specific nuances actually make it more prone to overfitting when fine-tuned? RoBERTa's pretraining on a larger corpus might make it better at capturing dataset-specific patterns but less flexible when faced with entirely new data distributions. The marginal improvements seen in the F1-scores on Political Dataset suggest that RoBERTa's advantages are task-specific and may not always generalize across datasets.

Conclusion

While BERT and RoBERTa demonstrate superior performance on single datasets, their inability to generalize across datasets highlights a critical limitation in current approaches to fake news detection. The results emphasize the challenges of cross-dataset generalization in fake news detection. It is crucial to use datasets that are diverse and representative of the real-world variability in fake and real news. Advanced models like BERT and RoBERTa, while powerful, require careful fine-tuning and evaluation to avoid overfitting. In real-world applications, fake news detection systems must generalize across different sources, topics, and writing styles. The observed performance drop underscores the need for robust validation strategies.

Future work should focus on addressing this generalization gap through improved dataset diversity, domain adaptation techniques, and rigorous cross-dataset evaluations. By doing so, we can ensure that these models are not only high-performing in controlled settings but also robust in real-world scenarios.

References

Bozkus, E. (2022, December 7). *Fake news detection datasets*. Kaggle. https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets?resource=download

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT https://arxiv.org/pdf/1706.03762

Kaliyar, R.K., Goswami, A. & Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed Tools Appl 80, 11765–11788 (2021). https://doi.org/10.1007/s11042-020-10183-2

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/pdf/1907.11692

William Lifferth. Fake News. https://kaggle.com/competitions/fake-news, 2018. Kaggle.

Zhou, X., & Zafarani, R. (2020). Fake News Detection: A Survey. ACM Computing Surveys (CSUR).https://dl.acm.org/doi/pdf/10.1145/3395046?casa_token=_azJaVW2lQ0AAAAA:r4gBZEbzq3G8xzPt4VETvYk0DxkjFA57xtJM5dCJLtus7nPpHYWlc6Wj1e1XYXI1UswSLzWv09rl