NewsShield: Detecting Deception with Large Language Models

Ali Ghorbanpour [1] Zhe Wang [1] Tsz Him Leung [1]

[1]Simon Fraser University, School of Computing Science

Abstract

NewsShield employs Large Language Models (LLMs) to detect deceptive content in news articles.

Using advanced NLP techniques, the model identifies linguistic patterns associated with deception. Promising results show its potential to combat misinformation and promote credible information sources. In our project, "NewsShield," we harness the power of the BERT [1] model to detect deception in news articles. Our results demonstrate the effectiveness of BERT in combating misinformation and promoting credibility in news sources.

Dataset

For our experiments, we utilize the **WELFake** [2] dataset, which comprises 72,134 news articles from diverse sources. This collection includes 35,028 authentic articles (labeled as 1) and 37,106 designated as fake (labeled as 0), ensuring a balanced representation. Curated from multiple datasets, this compilation prevents overfitting and enhances the effectiveness of training. The dataset's varied sources prepare models to handle genuine and deceptive news. It serves as a versatile tool for researchers combating misinformation, fostering a proactive approach to tackling its proliferation. A representative set of example rows is provided below:

Index	Title	Text	Label
1	NYC MAYOR DEBLASIO SAYS...	Why would anyone expect anything else from a m ...	1
2	Britain says expects most EU citizens...	LONDON (Reuters) - The British government said...	0
3	Hurdles high for Merkel in...	BERLIN (Reuters) - Germany s conservative chan...	0
...			

Table 1. **Sample Table Using WELFake Dataset**

Approaches

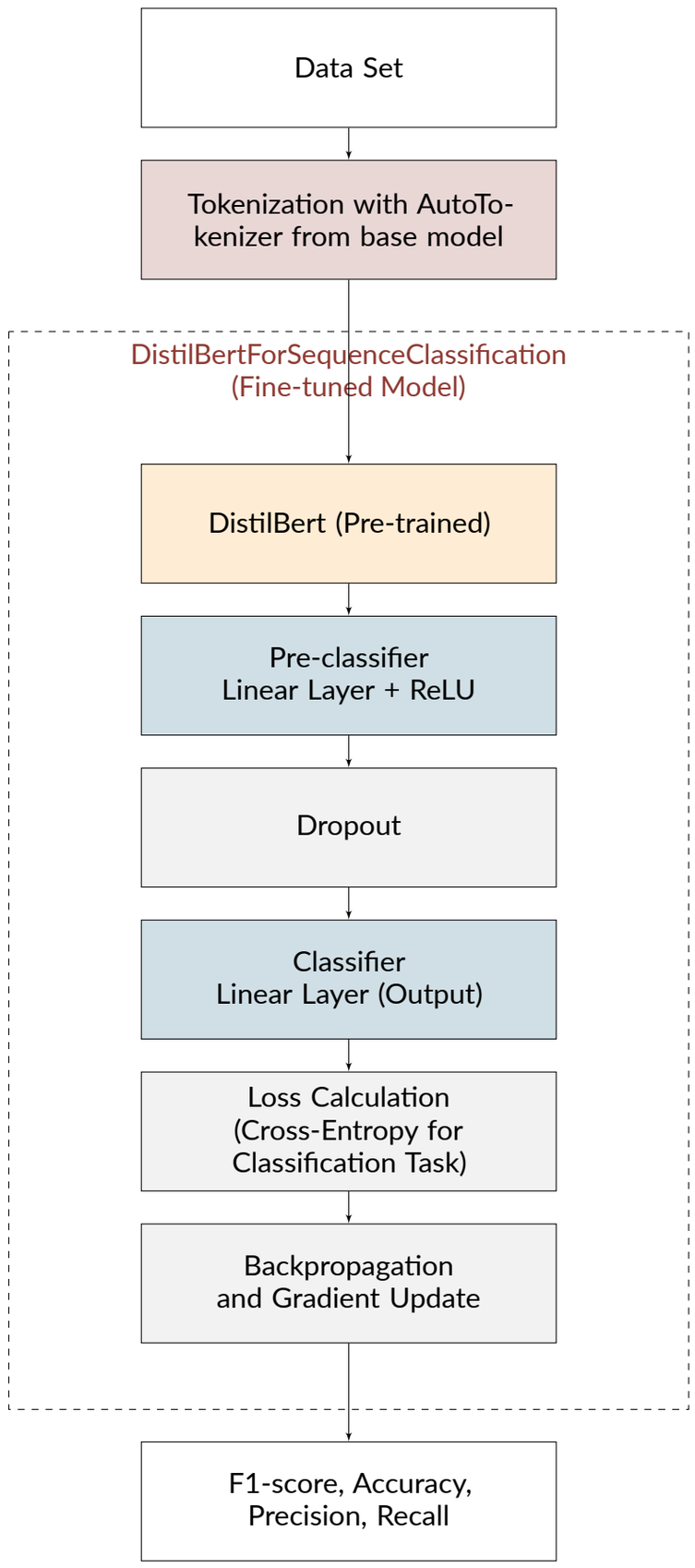
- Fine-tuning a large language model for classification of fake news.
- 3 epochs were used, with each epoch going through the whole training set.
- Validation set was used after the training to check the accuracy of the model. F1 score is also compared with the model before fine-tuning.

Models

- Pre-trained: **DistillBERT** [3]
- Fine-tuned: **DistilBertForSequenceClassification**

```
class DistilBertForSequenceClassification
(DistilBertPreTrainedModel):
    def forward(
        self,
        input_ids = None
        attention_mask = None,
        # ... (other arguments)
    ) -> SequenceClassifierOutput:
        # ... (function body)
```

Framework



Experiments

- Dataset used: A dataset of 72,134 news articles with 35,028 real and 37,106 fake news. 80/20 split, 80% used for training set, 20% used for testing set.
- Task: classification on whether the news is fake or not
- Fine-tuning: Used the whole training set for fine-tuning

Evaluation Metric & Result

Our model underwent a rigorous evaluation employing common classification metrics, yielding the following results:

Evaluation Metric	Fine-tuned Model	Pre-trained Model
Accuracy	0.951632	0.502272
Precision	0.951078	0.497532
Recall	0.951954	0.498211
F-1 Score	0.951078	0.534324

Our fine-tuned BERT model has garnered remarkably high scores across a spectrum of evaluation metrics, especially when contrasted with the lackluster performance of the pre-trained base model – a performance akin to random guesswork.

This experiment distinctly underscores the capacity of fine-tuning within the realm of fake news detection, showcasing its potential to elevate accuracy and efficacy significantly.

Conclusion

Our project harnesses BERT, a powerful Language Model, to combat misinformation. Utilizing a fake news dataset, we employed BERT classifier with remarkable success in distinguishing between authentic news and deceptive content. This highlights the potential of advanced NLP models in addressing misinformation challenges.

Key Takeaways:

- BERT classifier effectively detects fake news.
- Our work contributes to NLP advancements.
- Technology safeguards information integrity.
- AI empowers a trustworthy information ecosystem.

Through this project, we illuminate the transformative impact of AI-driven solutions in safeguarding the veracity of news sources.

References

[1] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[2] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893, 2021.

[3] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.