INFO 555 - Applied NLP Project 3

Kiwoon Hong

1 Description

In our interconnected digital world, misinformation can have a huge impact on societies around the world. Social networks and on-line platforms enable misinformation to spread rapidly, making it increasingly difficult for individuals to identify reliable sources. Furthermore, the creation of misinformation using AI is not only easy and fast, but also increasingly difficult to detect. As a result, the spread of fake news has become a significant challenge with widespread consequences.
To address this pressing issue, the goal of this project is to create a detection model using fake news data. By building effective fake news detection models, we can help stop the spread of misinformation, promote media literacy, and restore trust in reliable sources of information.
`GonzaloA/fake_news` Dataset [1] on Hugging Face contains the titles of news articles, their contents, and the labels of each article. Thus, this dataset can be used for model augmentation or to understand the performance of a model.

2 Related Work

As the importance of misinformation detection has grown over time, so has the research around it. The paper by Rubin et al. (2016) focused on satirical fake news, intentionally including cues that reveal its deception. They explored the characteristics of satirical news and compared satirical and authentic news across 12 topics. They then built a satire detection model using an SVM-based algorithm. A paper by Pérez-Rosas et al. (2018) also introduced two new data sets in seven news domains, explored the linguistic differences between fake and legitimate news, and used them to develop a linear SVM model that can detect fake news with up to 76% accuracy. The paper by Fung et al. (2021) introduced a new approach to the detection of fake news: cross-media consistency checking, which examines information across multiple media types to identify subtle inconsistencies in content. They also proposed a method for generating fake news data by introducing data synthesis using knowledge graphs. This approach achieved 92% - 95% accuracy.

3 Procedure

3.1 Data

Hugging Face's dataset `GonzaloA/fake_news` is designed for the task of detecting fake news. It contains labeled data that classifies news articles as'real' or 'fake.' It is also divided into three splits: 'train', 'validation', and 'test'. Since there is no training with the data in this project, we sampled from the 'train' part to evaluate the performance of the model. The 'train' part consists of 24.4k samples, of which 45.8%(11,158) of the samples are labeled as 0 (fake).

3.2 Experiment

First, as a high-level overview of this project, OpenAI's `GPT-4o-mini model` was set as the baseline model, and then a simple RAG experimental model using the web search API was used to evaluate the performance of both models. Due to the cost of the model and the limitation in search volume of the Web search API, which will be discussed later, only 500 pieces of data were sampled and used to evaluate the model performance. Since the distribution of labels was even, underover sampling was not used.

3.2.1 Baseline Model

This baseline model, the `fake_news_detect` function, uses the GPT-4o-mini model as the underlying architecture for fake news detection. The function takes the news title and content as input. The model uses OpenAI's `Chat Completions` API to process the input through prompts that explicitly tell the

[1] `https://huggingface.co/datasets/GonzaloA/fake_news`

model to act as an expert in fake news detection. The system prompt asks the model to classify the news content based solely on its title and text, allowing it to predict an immediate label. The response is labeled as “fake” or “real”, reflecting the model’s prediction.

3.2.2 Experimental Model

This experimental fake news detection model incorporates a search augmented generation (RAG) approach to build a baseline. The `fake_news_detect_rag` function incorporates external knowledge using the `search_and_contents` function, which retrieves web search results related to the news title. The `search_and_contents` function utilizes crewAI’s `exa.ai` API[2] to conduct the web search. The retrieved information, up to three web pages per keyword, is passed to the GPT-4o-mini model along with the news titles for classification.
A system prompt instructs the model to determine the authenticity of the news based on the availability and relevance of search results: if no results are found, the article is likely to be fake; if there are verified contents, it is likely to be real, thereby the model is instructed to inspect the contents of the article before making a judgment. This model is a step towards increasing the reliability of fake news detection by incorporating broader contextual information.

3.3 Evaluation & Analysis

For the baseline model, iterate over each row of the sampled dataset, use the `fake_news_detect` function, which takes the title and text as variables, and then save the results in a new column. Similarly, for the experimental RAG model, iterate over the sampled dataset, use the `fake_news_detect_rag` function, and then store the results in new column.

3.3.1 Bootstrap Resampling

Then, I performed the bootstrap resampling to compare the performance of the two fake news detection models. For each iteration (1000 iterations in total), I randomly sampled 10 rows from the sampled dataset. Then, I compared the predicted labels of the two models (label_llm and label_rag) to the actual labels (label). For each sample, I added +1 to the score for each row if the experimental model helps and -1 if it hurts. And then, I labeled the experimental model as the better model for that sample if the score is above 1 (1). Otherwise, I gave it a score of 0.

3.3.2 Evaluation

First, p-value was measured using the ’better’ column in Bootstrap.

```
P-Value = # not better / total samples.
```

Then, the accuracy and F1 scores are calculated for both the baseline model (label_llm) and the experimental RAG model (label_rag). The labels predicted by both models are compared to the actual labels in the sampled_df dataset.

4 Results

The statistics and figures are attached in the appendix.

4.1 Bootstrap Resampling

For the 1000 sampling sets, 972 sets are labeled as 0(not better). Therefore, the p-value is 0.972, which means there is no statistically significant improvement in terms of experiment models’ performance. The scores for the experimental model show a normal distribution with a mean of -3, indicating that it performs worse on average.

4.2 Models

The Baseline Model performed well with an Accuracy of 0.8460 and F1 Score: 0.8432. On the other hand, the results of the Experimental Model were close to random with Accuracy: 0.5160 and F1 Score: 0.5399.

5 Conclusion

In conclusion, the project obtained a significantly lower accuracy of the experimental RAG model compared to the baseline fake news detection model (Open AI’s GPT-4o-mini), with a p-value of 0.972, failing to reject the null hypothesis. From this result, I conclude that further optimization and improvement of the RAG model is needed. One possible improvement could be to give different weights based on the credibility of the news provider.

References

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen

[2] https://docs.crewai.com/tools/exasearchtool

McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

A Appendix

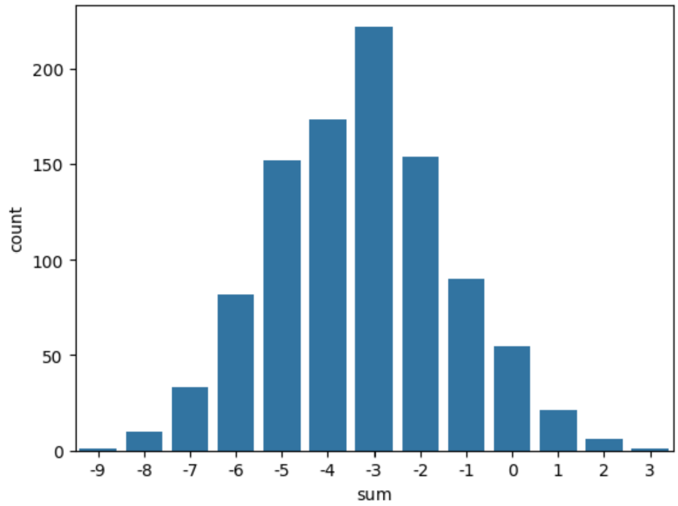

Figure 1: Distribution for scores of bootstrap resampling.

Model	**Accuracy**	**F1 score**
Baseline Model	84.6%	0.843
Experiment Model 2	51.6%	0.540

Table 1: Accuracy & F1 Score for each model.