NewsShield
Detecting Deception with Language Models

Ali Ghorbanpour
STD-ID: 301564668
aga149@sfu.ca

Zhe Wang
STD-ID: 301558386
zwa204@sfu.ca

Tsz Him Leung
STD-ID: 301544249
thl28@sfu.ca

1 Introduction

NLP enables machines to understand, interpret, and generate human language. Large-scale Language Models (LLMs) like GPT-3, trained on extensive text data, have revolutionized NLP by generating coherent and relevant responses for various applications. Combatting fake news is crucial in our information-rich society, as it distorts public opinion, erodes trust, and fosters societal divisions. Traditional methods like rule-based systems and manual fact-checking are inefficient. Advancements in NLP and LLMs present opportunities for automated and scalable fake news detection.

Fake news's widespread impact on individuals, communities, and society is undeniable. Misinformation distorts the truth, sways opinions, and can even undermine democratic processes. Real-world cases, from election misinformation campaigns to spreading false health claims during pandemics, underscore the urgency of detecting fake news. To foster informed decision-making and safeguard information integrity, it is crucial to develop reliable and precise techniques for fake news detection.

Fake news here refers to all forms of false, inaccurate, or misleading information, which now poses a big threat to human civilization.

2 Related Work

Current approaches to the Fake News Detection consists Knowledge-based methods [3], Style-based methods [7], Propagation-based methods [13, 11, 9, 14], and Source-based methods [5, 2]. These methods utilize manifolds of machine learning and deep learning techniques, such as SVM [16], Random Forest [4], Decision Tree [1], Gated Recurrent Unit [6], etc.

Existing benchmarks for evaluating the performance of Fake News Detection Model include FakeNewsNet [10], LIAR [15], PHEME [17]. These benchmarks provide datasets and evaluation metrics to assess the accuracy and effectiveness of different fake news detection methods.

3 Project Objectives

The objective of this proposal is to present a course project on utilizing LLMs for detecting fake news. Fake news has become a pervasive problem in today's information-rich society, and it is crucial to develop effective methods for identifying and combating it. By leveraging the power of LLMs, we aim to create a system that can analyze textual information and discern the authenticity and reliability of news articles, ultimately contributing to the fight against misinformation.

4 Methodology

The development of the project will be performed in Python since it has the most resources and is one of the most widely used languages for machine learning, which will help us to find the resources and information necessary to implement the solution quickly.

For the components in our system, as our goal is to create a system that can identify whether a particular news is fake news or not, the core component of the system will be made using a large language model (LLM) that is fine-tuned for the task of identifying fake news. In addition, a simple interface will also be implemented as a demonstration for the model, however, it is expected that such a model should be integrated into an existing system or a set of models for more comprehensive news analysis. Further, we also aim to collect several fake news datasets off the internet and fine-tune the large language model using them.

In regards to the significant language model, we plan on using the BERT family of LLM models to accomplish the goal of our system, we are choosing this family of models because of its relatively high accuracy for a relatively low number of parameters. This is important as we have limited resources on training models, meaning that fine-tuning an extensive model will be too computationally costly for us to accomplish within a reasonable amount of time. This approach also helps with saving time, as fine-tuning from an existing model consumes far fewer computational resources than training a large model from scratch.

We will also use several techniques for training the LLM, prompt tuning will be used to make the model more likely to output the result we wanted, then fine-tuning will adjust some parameters of the model itself so that it can learn some of the patterns of the specific use case in hand, which is fake news detection.

In our dataset selection process, we will adopt a comprehensive approach by incorporating a diverse range of fake and genuine news articles from multiple sources. This inclusive strategy aims to minimize any potential biases that could arise from relying solely on a single basis. The dataset will encompass crucial components such as the news title, content, and classification indicating whether it is fake or not. To prepare the data for model training, we will merge the title and content, forming a unified sequence of tokens that will be utilized to fine-tune the model based on the acquired classification outcomes. By leveraging this meticulous dataset curation and preprocessing, we aim to enhance the accuracy and robustness of our fake news detection model.

Potential datasets like "Liar, Liar Pants on Fire,"[15] "Fake News Detection on Social Media: A Data Mining Perspective,"[12] and the "Fake news classification - 2023"[8] offer labeled news articles with fake and genuine instances. We will evaluate their quality, diversity, and model performance to determine the most suitable dataset(s) for our project, considering the option of utilizing multiple datasets for comprehensive analysis.

5 Evaluation and Metrics

Evaluation and metrics are essential components for assessing the performance of fake news detection models. To measure the effectiveness of these models, various evaluation methods and metrics have been proposed in the literature. Commonly used evaluation metrics include accuracy, precision and recall, F1 score, the area under the ROC curve (AUC-ROC), cross-validation, confusion matrix, ROC curve, and precision-recall curve:

1. Accuracy:
$$\frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$
2. Precision:
$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$
3. Recall (Sensitivity):
$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$
4. F1 Score:
$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
5. Area Under the ROC Curve (AUC-ROC):
$$\int_0^1 \text{True Positive Rate}(\text{FPR})\, d(\text{FPR})$$

6. Confusion Matrix:

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

7. ROC Curve: Plot of True Positive Rate (TPR) vs. False Positive Rate (FPR)
8. Precision-Recall Curve: Plot of Precision vs. Recall

These metrics offer a comprehensive way to quantify the performance of fake news detection models and facilitate comparison between different approaches. In our project, we will select one or two evaluation metrics to assess the performance of our model and compare it with baseline models. This evaluation process will provide insights into the effectiveness of our model and its potential improvements.

References

[1] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.

[2] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2051–2055, 2021.

[3] Y. Dun, K. Tu, C. Chen, C. Hou, and X. Yuan. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 81–89, 2021.

[4] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE, 2013.

[5] Q. Li, Q. Zhang, and L. Si. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1173–1179, 2019.

[6] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.

[7] P. Przybyla. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 490–497, 2020.

[8] S. Shahane. Fake news classification, Jul 2023.

[9] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[10] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

[11] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 626–637, 2020.

[12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[13] K. Shu, S. Wang, and H. Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320, 2019.

[14] A. Silva, L. Luo, S. Karunasekera, and C. Leckie. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565, 2021.

[15] W. Y. Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[16] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7, 2012.

[17] A. Zubiaga, G. Wong Sak Hoi, M. Liakata, and R. Procter. Pheme dataset of rumours and non-rumours. 2016.