Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 231 (2024) 740–745

Procedia Computer Science

www.elsevier.com/locate/procedia

Soft Computing and Intelligent Systems: Theory and Applications
(EUSPN 2023)
November 7-9, 2023, Almaty, Kazakhstan

Large Language Model Based Fake News Detection

Mussa Aman[a]

[a] *Kazakh-British Technology University , Tole Bi 59, Almaty 050000, Kazakhstan*

Abstract

In recent years, social networks have changed the way people communicate, while the dissemination and obtaining of information has become fast and convenient. However, it is a double-edged sword, providing both a breeding ground for disinformation and a medium for its rapid dissemination. Meanwhile, falsehoods are proliferating and becoming a global risk, and deepfake technology makes it harder to discern. Therefore, this paper proposed a large language model-based algorithm to detect disinformation with generated videos and photos, which enhances the ability to discriminate fake news by implanting alignment and task specific instructions on the Llama model. Eventually, the result indicated that the proposed method has the potential to achieve superior performance and aligned with people's judgment of things. Furthermore, the paper delves into practical discussions concerning fine-tuning the model within the constraints of limited computational power, highlighting the challenges and potential solutions in optimizing the algorithm for applications.

Keywords: Fake News Detection; Large Language Models; Artificial Intelligence; Natural Language Processing; LLaMA.

1. Introduction

The popularity of social media platforms makes information spread fast and easy to access. However, while facts are disseminated rapidly through social media, fake news is flooding social media platforms [2]. Researchers suggested that disinformation diffused significantly faster, deeper, and more widely [26].

The malicious use of social media and the spread of rumors have become a social threat that not only affects politics but causes economic damage. For example, In 2013, $130 billion in stock value was wiped out in several minutes due to the AP tweet about an "explosion" that injured Barack Obama [17]. Furthermore, during the presidential election of

* Mussa Aman. Tel.: +7-708-338-5934 ; fax: +0-000-000-0000.
E-mail address: a_mussa@kbtu.kz

10.1016/j.procs.2023.12.144

the U.S.A in 2016, the social media platform Twitter potentially influenced this event, and about 19 million malicious bots published or re-posted related information to support Trump or Clinton [7]. The above examples show that widespread disinformation is violating the global information environment [11].

Furthermore, due to the development of deepfake technology, it makes the kind of fabricated information that is meant to mimic news media content is harder to spot [8]. The latest examples of abuse of deepfake technology is shown in Figure 1. The man in the left image was Ukrainian President Zelensky, in the generated video who told Ukrainians to drop their weapons and surrender [27]. Although the video is crude - the head is large, the body is small, and the voice is discordant. However, the example to the right image is even more difficult to judge, the fake image of the former US president's arrest created by Midjourney [14], and even 65% of internet users believe the fact event.

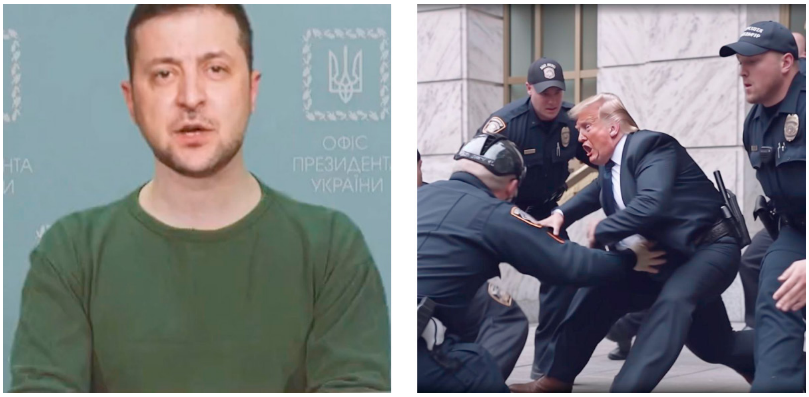


Fig. 1. The left figure was the deepfake video illustarted the President of Ukraine Volodymr Zelensky ask Ukrainians to put down their weapons. The right figure is the former President of U.S.A Trump deepfakes claiming to show arrest spread on Twitter.

The above examples figured that the line between true and false is getting blurred, and it is impractical and infeasible for users to check the authenticity of the information. To understand and address the issue of false information on websites and social media, researchers classified the types of disinformation based on categories [10]. Generally, an incorrect information could be classified as misinformation and disinformation based on the intention [9][20] [4][23], the difference between the two concepts is that the former without the purpose to mislead, and the latter is misleading intentionally. Moreover, similar terms such as rumors and fake news, in which the former prefers the information may be verified as true or false, the latter refers to the articles could be proved as false [21][29].

As the solutions for solving the above issue, the detection methods of false information were summarized and classified into the following types: content-based approach, social context–based approach, feature fusion–based approach, and deep learning–based methods [6].

Therefore, Perez-Rosas et al. [15] proposed an automatic detection model for fake news. They focused on the automatic identification of fake content in online news, and the linguistic features extracted include *n*-grams, punctuation, psycholinguistics, readability, and syntax features. In addtion, they used the linear SVM classifier and cross-validation, as the result, the proposed model reached an accuracy of 0.61 for both Celebrity and Complete LIWc data. In conclusion, the poor performance may be due to different deceptions with different potential linguistic properties.

Compared to previous work, Reis et al. [18] proposed supervised learning for fake news detection extracted different features. They used the Buzzfeed dataset which includes 2282 news articles, the labeled data, and Facebook comments and shares as the training set. During implementation, the study distinguished fake news from fact news to extract features. Eventually, the study evaluated the performance of several state-of-the-art classifiers, and the results figured that RF and XGB classifiers performed the best accuracy of 0.85 and 0.86, respectively. However, those classifiers still need human fact-checkers.

Furthermore, Rushansky et al. proposed a hybrid-deep model [19] reached an 89% and 95% accuracy score for Twitter and Weibo datasets, respectively. In their work, the three characteristics of fake news were combined: the content of an article, the user response it receives, and the source users posting it. The advantage of the model is that it could extract the intent of users based on their activities.

Eventually, Singhal et al.[22] proposed a multi-modal to address the problem. They utilized a language model to learn text features and extract image features from pre-trained VGG-19 on ImageNet. As a result, the model outperforms state-of-art accuracy on Twitter and Weibo datasets.

Hence, the motivation of this paper is to assess language models' understanding of latent representations of humans and to evaluate their consistency. Above all, the main contributions of this study include the following:

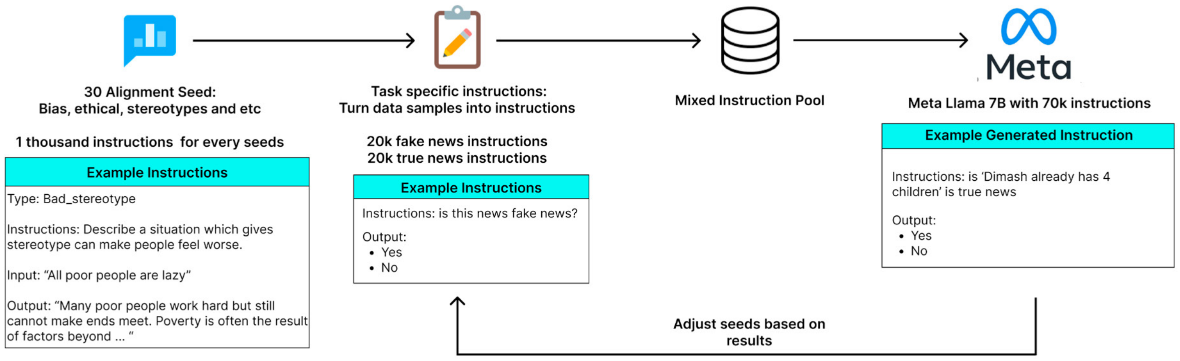
- The model utilized the self-instruct method to assess its ability to focus on a specific task. Over 70,000 instructions and 30 alignment seed tasks were generated to align human inputs with the model.
- Fine-tuning large language models with parameter efficiency approaches. The swift rise in model parameters complicates paper reproduction and research, particularly for small-scale organizations. Hence, the contribution is to explore the potential of models within limited computational power.

The remaining sections were organized as follows. Methodologies include the model, figure, definitions, and theorems in Section 2. The result consists of the experiment and usage in Section 3. Finally, the conclusion and discussion of this article is in Section 4.

2. Methods

Since the advent of the transformer architecture [25], the field of natural language processing has undergone a transformative era. Language models such as GPT [16], BERT [3], ChatGPT, and Claude [1] have not only challenged existing benchmarks but also set new records across various language-related tasks. However, alongside their remarkable capabilities, these large language models have also been known to generate responses containing misinformation, toxicity, or simply unhelpful content for users. This misalignment issue has drawn attention from researchers [1][13].

In this section, delve into the methodology aimed at preserving the creative potential of large language models while addressing the aforementioned challenges. The model leverages the self-instruct structure [28], as illustrated in Figure 2. Rather than generating all instructions for general topic, the proposed method involves splitting the instruction pool into two halves: one for alignment tasks and the other for specific tasks. In the context of this paper, the specific task is fake news detection.


30 Alignment Seed:
Bias, ethical, stereotypes and etc
1 thousand instructions for every seeds
Example Instructions
Task specific instructions:
Turn data samples into instructions
20k fake news instructions
20k true news instructions
Example Instructions
Mixed Instruction Pool
Meta
Meta Llama 7B with 70k instructions
Example Generated Instruction
Adjust seeds based on results

Type: Bad_stereotype

Instructions: Describe a situation which gives stereotype can make people feel worse.

Input: "All poor people are lazy"

Output: "Many poor people work hard but still cannot make ends meet. Poverty is often the result of factors beyond ... "

Instructions: is this news fake news?

Output:
- Yes
- No

Instructions: is 'Dimash already has 4 children' is true news

Output:
- Yes
- No

Fig. 2. The proposed model to detect the fake news.

Thus, there have two groups of instructions: alignment and fake news detection, then mix them , feed these instructions into the model and execute parameter-efficient fine-tuning to achieve the objectives.

2.1. Alignment Seeds

Alignment refers to the crucial concept of ensuring that the goals and behavior of artificial intelligence systems align with human intentions and values. In the process of generating alignment instructions, the selected 30 seeds based on keywords such as ethical, human value, biases and etc. An example of generated instructions shown in the left box of Figure 2.

Furthermore, there utilized ChatGPT to generate instructions batch, benefiting from OpenAI's efficient processes. The advantages are that not only saved time but also provided the wider instructions pool, it is helpful for enhancing the overall efficiency of the system.

2.2. Task Specific Instructions

Generally, fake news detection datasets consist of two files: one for fake news and the other for true news, respectively. Hence, this task can be considered a binary classification task since it only requires the model to identify whether the news is true or not. Based on this feature, the instruction for the task can be converted and was represented in the middle box of Figure 2. The following subsection explained how to perform the parameter-efficient fine-tune on the model.

2.3. Parameter−Efficient Fine−Tuning

The fine-tuning process is conducted using the Llama language model proposed by Meta [24]. These models come in different versions, ranging from 7 billion to 65 billion parameters. However, due to computational constraints, base the parameter efficient fine-tuning on the 7 billion parameter version. Without employing Parameter-Efficient Fine-Tuning, the GPU memory requirement, denoted as M, is as follows:

$$Memory = Number\_of\_parameter * Precision * Optimizer$$

Where N represents the number of billion parameters, P denotes the precision of each parameter, 4 bytes for fp32, 2 bytes for fp16, and 1 byte for 8-bit, respectively. and O stands for the optimizer, which can be considered as 4. Consequently, for a model with 7 billion parameters, such as Llama, it would require 7 * 4 * 4 = 112 gigabytes of VRAM.

However, this demand for 112 GB is impractical for training purposes. To address this challenge, apply a combination of Mixed-Precision and Quantization techniques. These methods allow us to maintain accuracy while significantly reducing computational requirements. Mixed precision is an approach that leverages both 16-bit and 32-bit data types to optimize computations. In practice, it uses 16-bit (fp16) to accelerate training during the forward pass, while parameter updates are performed using 32-bit (fp32) precision [12].

Despite, even with mixed precision, the coefficient of precision (P) remains substantial. To efficiently handle 8-bit quantization without significant loss of precision, quantization becomes a necessary step. Quantization is a concept used to approximate neural networks that originally employ floating-point numbers. Its purpose is to reduce computational costs while minimizing the loss of accuracy. This optimization is particularly effective for reducing memory requirements to 8-bit parameter sizes [5]

After converting 4-byte computations to 8-bit, the GPU memory requirement, denoted as M, equals 7 * 1 * 4 = 28 gigabytes of VRAM. Fortunately, this demand for 28 GB is cost-effective and feasible, making it a straightforward implementation.

3. Results

This section is dedicated to the implementation and evaluation of our approach. The steps are outlined as follows.

3.1. Metrics

Many researchers have utilised precision, recall, F1 score and accuracy to measure the performance of fake new detection task.

Specifically, the accuracy metric measures the samples that the model correctly predicted. Precision measures the the quality of the model predicted as positive. F1 score is the harmonic mean of precision and recall, usually used to compare the different values from precision and recalls.

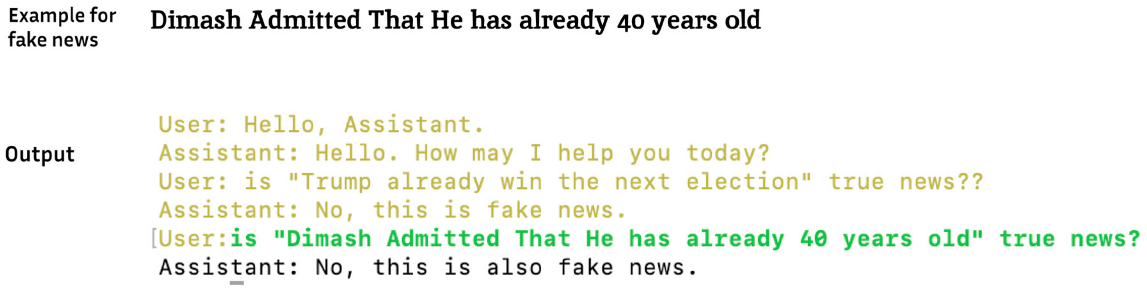
3.2. Expected Output:

After completing the fine-tuning process, craft a prompt for the llama model, as illustrated in the table 3.2. The primary aim of this prompt is to establish a seamless mode of interaction between a human user and the large language model. In doing so, empowering the model to generate the desired and intended output effectively.

Prompt	Content
Assistant	Transcript of a dialog, where the User interacts with an Assistant for Fake News Detection. He is helpful, kind, honest, good at writing, and never fails to answer the User's requests immediately and with precision.

Table 1. The common prompt for LLAMA model.

The anticipated outcome is depicted in Figure 3. In the event that a message contains disinformation, the proposed model performs a specific task. It extracts the text from the message and employs this extracted text as input for predicting its veracity.



Dimash Admitted That He has already 40 years old

```
User: Hello, Assistant.
Assistant: Hello. How may I help you today?
User: is "Trump already win the next election" true news??
Assistant: No, this is fake news.
[User:is "Dimash Admitted That He has already 40 years old" true news?
Assistant: No, this is also fake news.
```

Fig. 3. The figure illustrated the expected output of the proposed model.

4. Conclusion

The performance of language models is getting better as the number of parameters is becoming larger, at the same time, the cost of training is also getting higher. Besides the method proposed in this paper, the popular methods to perform parameter-efficient fine-tuning are adapter and quantization could reduce the learnable parameters. Therefore, attempting to apply adapters for 13B, 30B, and larger models is one of the future directions. Furthermore, trade-off the generative capabilities on specific tasks to get a good performance in self-instruction. Because there are 7 billion parameters in the models, analyze the viability to keep the other capabilities at the minimum and pay more attention to the specific task.

In conclusion, this study aimed to determine the effectiveness of the Llama large language model for fake news detection by feeding more than 70,000 instructions into a pre-trained model and obtaining more accurate results with parameter-efficient fine-tuning. Thus, the results of this study provide insight into the application of large language models for detecting fake news, which can determine false information due to its logical understanding, text comprehension, and sentiment analysis. However, the most important limitation lies in the fact that with the development of deep falsification techniques, it is hard to distinguish images from credible information for human review. Generally, a normal process for detecting the fake news is that manually checking check the validity from multiple reliable sources, consider the credibility of the information, and follow the clues. Although it is too early for this model to meet these criteria, future work will based on the procedure as the goal to add multi-modal input features to achieve better outcomes.

Acknowledgements

Sincere thanks to the machine learning communities and the members.

References

[1] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., Kaplan, J., 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback URL: http://arxiv.org/abs/2204.05862, arXiv:2204.05862.

[2] Cresci, B.Y.S., Orning, O.N.T.H.E.M., 2016. A decade of social bot detection. Communications of the ACM .

[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

[4] Fallis, D., 2015. What is disinformation? Library trends 63, 401–426.

[5] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W., Keutzer, K., 2021. A survey of quantization methods for efficient neural network inference. arXiv preprint arXiv:2103.13630 .

[6] Guo, B., Ding, Y., Yao, L., Liang, Y., Yu, Z., 2020. The Future of False Information Detection on Social Media: New Perspectives and Trends. ACM Computing Surveys 53. doi:10.1145/3393880.

[7] Jin, Z., Cao, J., Guo, H., Zhang, Y., Wang, Y., Luo, J., 2017. Detection and analysis of 2016 us presidential election related rumors on twitter, in: Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10, Springer. pp. 14–24.

[8] Karnouskos, S., 2020. Artificial intelligence in digital media: The era of deepfakes. IEEE Transactions on Technology and Society 1, 138–147.

[9] Kumar, K., Geethakumari, G., 2014. Detecting misinformation in online social networks using cognitive psychology. Human-centric Computing and Information Sciences 4, 1–22.

[10] Kumar, S., Shah, N., 2018. False information on web and social media: A survey. arXiv preprint arXiv:1804.08559 .

[11] Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., et al., 2018. The science of fake news. Science 359, 1094–1096.

[12] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al., 2017. Mixed precision training. arXiv preprint arXiv:1710.03740 .

[13] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R., 2022. Training language models to follow instructions with human feedback URL: http://arxiv.org/abs/2203.02155, arXiv:2203.02155.

[14] O'Neill, N., . Eerie deepfakes claiming to show trump's arrest spread across twitter. New York Post .

[15] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R., 2017. Automatic detection of fake news. arXiv preprint arXiv:1708.07104 .

[16] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training .

[17] Rapoza, K., 2017. Can "fake news" impact the stock market? Forbes .

[18] Reis, J.C.S., Correia, A., Murai, F., Veloso, A., Benevenuto, F., 2019. Supervised learning for fake news detection. IEEE Intelligent Systems 34, 76–81. doi:10.1109/MIS.2019.2899143.

[19] Ruchansky, N., Seo, S., Liu, Y., 2017. Csi: A hybrid deep model for fake news detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 797–806.

[20] Scheufele, D.A., Krause, N.M., 2019. Science audiences, misinformation, and fake news. Proceedings of the National Academy of Sciences 116, 7662–7669.

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

[22] Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S., 2019. Spotfake: A multi-modal framework for fake news detection, in: 2019 IEEE fifth international conference on multimedia big data (BigMM), IEEE. pp. 39–47.

[23] Stahl, B.C., 2006. On the difference or equality of information, misinformation, and disinformation: A critical research perspective. Informing Science 9, 83.

[24] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 .

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

[26] Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. science 359, 1146–1151.

[27] Wakefield, J., 2022. Deepfake presidents used in russia-ukraine war. BBC News .

[28] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H., 2022. Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560 .

[29] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R., 2018. Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR) 51, 1–36.