FIGHTING HEALTH-RELATED DISINFORMATION IN SOCIAL MEDIA WITH LARGE LANGUAGE MODELS

Moisés Robles Pagán
Electrical and Computer Engineering Department
University of Puerto Rico - Mayagüez
moises.robles@upr.edu

Manuel Rodríguez Martínez
Electrical and Computer Engineering Department
University of Puerto Rico - Mayagüez
manuel.rodriguez7@upr.edu

***Abstract*—Combating disinformation in social media is an important problem, particularly when the disinformation target healthcare. In our work, we are exploring how to finetune Large Language Models (LLM) to counteract health-related disinformation on social media. The base models that were finetuned for this project are Flan-T5, BERT, Mistral, GPT-J, and LlaMa-2. The process of finetuning was divided in two sections: 1) classifying if the text is health related, 2) verifying if the text contains disinformation. Then, we augmented the LLM with RAG to query trusted medical sources that can be used to debunk disinformation. The first part was done by using a dataset of around twelve thousand labeled tweets. Then, to determine misinformation a web scraper was design to gather data from social medias like Truth Social and classified by health professionals. For the final RAG step, we collected data from official health sources like the CDC and PubMed and stored them in a vectorize database, Chroma. Our current experiment on Flan-T5 shows that the system can classify if tweets are health related with a precision of 79%, a recall of 82%, and a F1 score of 80%. This can help health experts combat and rebut disinformation in the different social media platforms.**

***Index Terms*—Large Language Model, social media**

I. Introduction

Lorem

A. Contributions

The contributions in this research

B. Paper Organization

This paper has the following organization. Section II contains the literature review on the transformer and the Large Language Models architectures, and the different use cases of these models for classification. For section III, we can observe the projects pipeline. Later, in section IV we have the methodology and experiment. Section V presents our results. Related works are presented in section VI. Ending with section VII, we have our conclusion with suggestions for future work.

II. Literature Review

A. Transformers

The

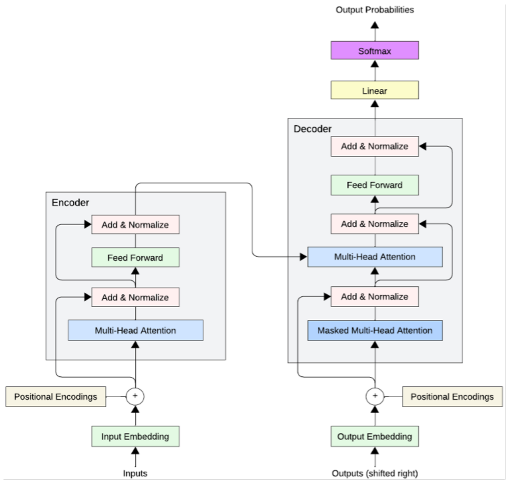

Fig. 1. The transformer architecture

B. Large Language Models (LLM)

The are three different architectures for Large Language Models, encoder-only, decoder-only, and encoder-decoder. Each one has advantages on specific tasks.

1) Encoder-only models: These models are like BERT. This type of model predict by masking specific words in a sentence. The are better for classification and sentiment analysis.

2) Decoder-only models: For these model we have the well known GPT-3, Mistral, and LLaMa. They receive one input and try to predict the entire text. They are good for summarizing and text-generation.

3) Encoder-Decoder models: The encoder-decoder models are T5. They mask entire sequences of text. Good for translation and question and answering. [**?**]

C. Health Misinformation

III. Methodology

A. Fine-tuning

The Large Language Models (LLM) used for this paper are pre-trained, meaning that the model learned how words relates to each other, the model is trained to understand a language. This process is computationally expensive and requires large amount of data. Instead of creating a model from scratch, we can teach an existing model to learn a new task such as text classification, translation, generation, or other. This is called fine-tuning, in this case was used to teach the model to classify two types of texts: health-related and misinformation-related.

The health-related dataset consists of over 12,000 tweets extracted by the previous THS project. Said dataset was classified on three category: related, unrelated, and ambiguous tweets. A second dataset was created with data from different sources such as news, social medias, and blogs classified as misinformation and non-misinformation ***[Insert References]***.

The models used for the classification process were Bert, T5, and LLaMa-2.

1) Health-Related Classification:

2) Misinformation Classification: 2. Misinformation classification

LR, Batch size, seed, and epochs are static.

- Each model was fine-tune twice.

1. Sequence classification: 1, 2, or 3; 1 or 2.

2. Classification with text generation: Related, Unrelated, or Ambiguous; Misinformation or Not Misinformation.

- Weighted average added for the sequence classification.

B. Misinformation Rebuttal Pipeline

- Scraper extract links from PubMed

- Save links in temporary file and upload to a consumer to process link information.

- Data is divided into different tables

- Papers context is broken into chunks and uploaded into a vectored database.

- Chunks are mapped to relational database metadata. + Add ERD Schema

- LLM sends query into vector database.

- Id and context is retrieved and process. Id is sent to relational database to retrieve source.

- Model respond with full rebuttal and offical link. + Add pipeline image

C. Hardware and Software

- V100 machines: 32Gb VRAM, and 80-ish RAM

 Cuda 11.7
-- Python 3.9.19
- Pytorch 2.0.1
- Transformers 4.34.0

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks . . .". Instead, try "R. B. G. thanks. . .". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Please number citations consecutively within brackets [**?**]. The sentence punctuation follows the bracket [**?**]. Refer simply to the reference number, as in [**?**]—do not use "Ref. [**?**]" or "reference [**?**]" except at the beginning of a sentence: "Reference [**?**] was the first . . ."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [**?**]. Papers that have been accepted for publication should be cited as "in press" [**?**]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [**?**].

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.