Assignment 2

Seyedali Shohadaeolhosseini
Master's Degree in Artificial Intelligence, University of Bologna
s.shohadaeolhosseini@studio.unibo.it

Abstract

Sexism classification, a sub-problem of hate speech has always been important, especially nowadays that companies use data crawled from the World-Wide-Web. In this project, with help of the provided APIs from the Hugging Face, we have worked on two LLMs and tested their performance on classifying a given a tweet as whether it is sexism. Specifically, through defining zero-to-few shot prompts, we have tested our test data and studied the performance of Mistral V2 and Mistral V3. Our experiments show that the LLMs can perform a better prediction if they are provided with some examples of how and what they should generate.

1 Introduction

We start with our two quotes, "your model is a reflection of the provided data!" and "your model learns what you feed it!" these are the fundamentals of any AI models and are what we experienced through out this project. Today the more data you provide to the model, the better model you get. This is one of the reasons that crawling data from the web is being performed more and more. This also is another reason to be more cautious on the data we're crawling. These data are being gathered from a place where people are sharing them openly which arouse a problem, hate speech problem. The sexism detection is not a new problem. (Jhakal et al., 2023) have worked on the same topic. They tested a variety of pre-trained transformers including RoBERTa, BERT, BERTweet and XLNet.

With this motivation, we have studied the performance of LLMs on the sexism detection problem. Specifically we have demonstrated the abilities of the Mistral v2 and Mistral v3 on classifying a given tweet being fed to LLM through a zero-to-few shot prompts.

2 System description

In this project, we have used Hugging Face's API to import two LLMs, Mistral V2 and Mistral V3. After configuring the imported models, we have tested a test data on them. Our task was to provide the test tweet to the model through zero-to-few shot prompting. Having said them, we have been also provided another dataset, demonstrations data, which we used it for few-shot prompting.
In both ways of prompting, the system role is defined as "You are an annotator for sexism detection." and the user provided prompts starts with "Your task is to classify input text as containing sexism or not. Respond only YES or NO.".
In the following, we will briefly discuss the provided datasets. Our experiments and results, discussion and at the end we conclude the project.

3 Data

For this project we are given two datasets, test set and demonstrations set. In the test data we have tweets that are labeled as either 'sexist' or 'not sexist'. These are the tweets that we will use to prompt the model and ask the LLM to classify it.
To conduct a better experiment, after this part, we have provided another prompts for the LLMs in which this time, we are also providing examples to the LLMs as well. We are taking these examples from the provided demonstrations dataset randomly but evenly, to make sure there won't be any biases on the provided examples.

4 Experimental setup and results

Both of the imported models almost follow a same structure of settings. Both of them are quantized and are set to generate only 10 words. Padding for both is set to True and input truncation is disabled, as we are not prompting a large input and we want the model to see the input entirely. The prompt for the models are all passing through a same function.

Having these same conditions, the table 1 shows that corresponding results of prompting. Based on the results, the mistral 2 with zero shot prompting has the best inference accuracy. However, this model has failed on generating the correct format of response as we asked to do for over 30 times. On the other hand, the Mistral 3 with 2 and 3 shots prompting shown performed as well as the mistral 2 and it doesn't have any fail generation. Fail generation is the generation of response which is not like what we defined for the model. Precisely, we expect the models to generate the answer "YES" or "NO" immediately rather than after generating a couple of tokens.

	Accuracy	F1 Score	Fail
MST2-Inf	**0.7466**	**0.7429**	**30**
MST2-2Shots	0.6733	0.6695	2
MST2-3Shots	0.6666	0.6648	0
MST2-4Shots	0.6400	0.6372	2
MST3-Inf	0.5900	0.5164	0
MST3-2Shots	**0.7233**	**0.7195**	0
MST3-3Shots	**0.7233**	**0.7207**	0
MST3-4Shots	0.7033	0.6992	0

Table 1: Three metrics are shown for the four combinations. MST2-Inf refers to the Mistral V2 model with simple inference and shot refers to few shot inferencing. Same applies for the MST3 which is Mistral V3.

5 Discussion

The results show that the mistral 2, from the accuracy and F1-score point of view, is the best model. However, this model for 30 times couldn't generate the exact response we asked it to generate which is the largest fail ratio between all the combinations we tested. On the other hand, the mistral 3 inference, provided 2, 3, 4 shots, is performing as perfect as mistral 2, but with no fail generation. This suggests that the mistral 3 has a better understanding of the input tweet.
Moreover, multi-shot prompting on mistral 2 showed a better response generation where the model generated the correct template of response.

Based on the confusion matrix that we checked, the zero-shot prompts, for both models showed an odd results. Both the Mistral 2 and 3 were biased on the class 1. Mistral 2 has predicted 186 out of 300 tweets as the class 1, 130 of which was only correct. The mistral 3 was even worst in zero-shot. It predicted 267 out of 300 tweets as the class 1, and only 147 of them were classified correctly. This odd behaviour was improved completely after we provided multi-shot prompts.

6 Conclusion

This project proved that prompts can indeed perform an important role in the LLM. A good prompt can help the model to avoid biases. To think and generate response in the same context. This was especially seen at the end when we demonstrated the confusion matrix.

References

Chirayu Jhakal, Khushi Singal, Manan Suri, Divya Chaudhary, Bijendra Kumar, and Ian Gorton. 2023. Detection of sexism on social media with multiple simple transformers. In *CLEF 2023: Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. Chirayu Jhakal and Khushi Singal are Corresponding authors.