HAMiSoN-Generative at ClimateActivism 2024: Stance Detection using generative large language models

Jesús M. Fraile-Hernández and **Anselmo Peñas**
NLP & IR Group at UNED
jfraile@lsi.uned.es and anselmo@lsi.uned.es

Abstract

CASE in EACL 2024 proposes the shared task on Hate Speech and Stance Detection during Climate Activism. In our participation in the stance detection task, we have tested different approaches using LLMs for this classification task. We have tested a generative model using the classical seq2seq structure. Subsequently, we have considerably improved the results by replacing the last layer of these LLMs with a classifier layer. We have also studied how the performance is affected by the amount of data used in training. For this purpose, a partition of the dataset has been used and external data from posture detection tasks has been added.

1 Introduction

CASE in EACL 2024 is a shared task focusing on Climate Activism (Thapa et al., 2024). This task consists of three subtasks, the first two are focused on Hate Speech detection, a task that is important for peace and harmony in society (Parihar et al., 2021). The last subtask consists of the posture detection of tweets on this topic.

Stance detection seeks to determine the author's point of view - usually in favour, against or neutral - on certain topics, using textual analysis (AlDayel and Magdy, 2020; Küçük and Can, 2020). Due to the large amount of information that is processed daily on social networks, stance detection has become an important task that facilitates the understanding of social and political changes (Darwish et al., 2017).

Due to the large amount of information that large Language Models (LLMs) receive during their training and their good results in many benchmarks, they are being used for tasks such as posture detection in text (Cruickshank and Ng, 2023; Mets et al., 2023). In which models such as ChatGPT, GPT-NeoX (Black et al., 2022), Falcon 7B and 40B (Almazrouei et al., 2023) and Llama 2 7B and 13B (Touvron et al., 2023) were used. All of them were used as Sequence-to-Sequence models.

In this paper we will compare the performance of different LLama 2 model structures for stance detection tasks. It also seeks to study how the performance is affected by the amount of data used in training. For this purpose, a partition of the dataset will be used and external data from stance detection tasks will be added.

The rest of this paper is structured as follows: Section 2 describes the datasets to be used along with the task to be solved. Section 3 describes the methodology followed including the models, the data processing, the model inputs and the training dataset. Section 4 presents the results, which will be discussed in Section 5. Finally, conclusions and future work are given in Section 6.

2 Dataset and task

The dataset on climate activism (Shiwakoti et al., 2024) has been used, focusing on the subtask of stance detection. This dataset has a collection of tweets labeled according to their stance about climate activism. Henceforth, we will refer to it as CASE.

Additionally, the dataset from (Mohammad et al., 2016) has been employed, which is related to the stance detection task too. This dataset was used in the International Workshop on Semantic Evaluation (SemEval-2016). It includes tweets labeled with stances about various targets such as climate change, atheism, feminism, etc.

3 Methodology

This document aims to make a comparison between different Llama 2 model approaches, in addition to studying how the performance is affected by the amount of data

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 79–84
March 22, 2024 ©2024 Association for Computational Linguistics

3.1 Models

Four different approaches have been used, always based on the auto-regressive language model, Llama 2 7B.

- Llama 2 7B Chat (**7B Chat - seq2seq**). This model is specially trained to be used as a chatbot, for this reason the prompts that will be the inputs to the model will follow the guide proposed in (Touvron et al., 2023). These prompts will be described in section 3.3. In our case we are looking for a response with a word that indicates the stance of the entered text.

- Llama 2 7B Chat with a final classification layer and using prompts formatted (**7B Chat - clf prompt**). In this model we start from the Llama 2 7B Chat model and eliminate the last linear layer to add another linear layer that has as input the last hidden state of the model and as output 3 neurons, one for each stance label. With this model, text formatted following the prompt guide mentioned above has been used as input.

- Llama 2 7B Chat with a final classification layer and using raw prompts (**7B Chat - clf no Prompt**). It is a model with the same architecture as the previous one, however the text without formatting has been used as input.

- Llama 2 7B with a final classification layer (**7B - clf**). In this model we start from the Llama 2 7B model and carry out the same process as the two previous models. The non-chat model has not been trained to be used as a chatbot so the text used as input will not have any specific format.

3.2 Dataset preparation

Each approach use the text of the tweet in the input. However, a pre-processing of the text has been performed, consisting of the following steps:

- Remove all urls from tweets.

- Remove all users in the form @user.

- Separate hashtags into individual words. For this we have used the wordninja library, which uses a probabilistic division of concatenated words using NLP based on the frequencies of unigrams from the English Wikipedia.

Four experiments have been performed varying the dataset used for training. Two of them are using only the CASE dataset and the other two are using the whole SemEval dataset together with the CASE data. Regarding the CASE dataset. One of the experiments uses a stratified partition for each label of the training set with a size of 70% for training and 30% for validation (hereafter referred to as part or partition) and another experiment uses the training set for training and the development set for validation (hereafter referred to as full).

3.3 Model inputs

Since models that have not been trained to have conversations are being used, a particular input format has been used for each model. For the **7B Chat - clf no Prompt** and **7B Chat - clf** models the input is the processed text as shown in 3.2.

For the models **7B Chat - seq2seq** and **7B Chat - clf prompt** the prompt guide proposed by Meta has been used together with a description of the task as shown below.

> <s>[INST]«SYS»
> Classify the stance of the following text. If the stance is in favour of ***stance-target***, write FAVOR, if it is against of ***stance-target*** write AGAINST and if it is ambiguous, write NONE. The answer only has to be one of these three words: FAVOR, AGAINST or NONE.«/SYS»
> ***Processed Text*** [/INST]

.

Where ***stance-target*** is the target that the tweet is talking about. In the case of the CASE dataset this would be Climate Activism. In the case of the SemEval dataset tweets have targets such as climate change, atheism, feminism, etc.

3.4 Training phase

To train each of the proposed models, a Fine Tuning has been performed using the LoRA technique: Low-Rank Adaptation of Large Language Models (Hu et al., 2021) together with a 4-bit Quantization (Dettmers et al., 2023). As hyperparameters for training we have selected a range $r = 64$, an $\alpha = 16$, and a dropout of 0.1.

With this configuration, it is possible to train around 350M parameters, which is a 95.5% reduction of the total number of parameters of the original models.

4 Results

This section presents the results, evaluated on the test set, of all the experiments that have been carried out.

Table 1 shows the F1 macro value of the 8 different runs. The results are split into part if the 70-30 partition was used or full if the whole dataset was used for training, as explained in section 3.2. Models marked with * indicate that they have been trained with the CASE and SemEval dataset. In addition, the results of the Baseline model used in (Shiwakoti et al., 2024) are included. This Baseline model, named ClimateBERT (Webersinke et al., 2022), is an adaptation of a BERT model, a language model trained on a corpus sourced from climate-related news, abstracts, and reports.

Hereafter, model (1) is the 7B Chat - clf prompt model trained with the partition of the data and model (2) is the 7B Chat - clf non prompt model trained with the total data.

Approach	**part**	**full**
Baseline	0.545	
Best model leaderboard	0.7483	
7B Chat - seq2seq	0.7043	0.7062
7B Chat - seq2seq *	0.6986	0.6845
7B Chat - clf prompt (1)	**0.7246**	0.6958
7B Chat - clf prompt *	0.7102	0.7009
7B Chat - clf no prompt (2)	0.7068	**0.7366**
7B Chat - clf no prompt *	0.7231	0.7300
7B - clf	0.7245	0.7189
7B - clf *	0.7190	0.7160

Table 1: Results for the test set (trained on the 70-30 % CASE partition or the full CASE train set). Models marked as * indicate that they have been trained with the CASE and SemEval dataset.

Table 2 shows the percentage of misclassified and well-classified instances for each number of systems. For example, the first value of 7.9 % in the second row indicates that 7.9 % of the instances have been misclassified by two systems and the other 6 systems have classified them correctly.

Some metrics for the best performing models using the partition (1) and with the total data (2) will be shown below.

Figure 1 shows the normalised confusion matrix over the rows for model (1). Similarly, Figure 2 shows the normalised confusion matrix over the rows for model (2).

		part		**full**	
		Wrong	Right	Wrong	Right
Number of systems	**1**	13.3 %	5.4 %	16.5 %	6.2 %
	2	7.9 %	5.1 %	9.9 %	5.3 %
	3	6.3 %	4.2 %	6.9 %	4.9 %
	4	5.3 %	5.3 %	6.6 %	6.6 %
	5	4.2 %	6.3 %	4.9 %	6.9 %
	6	5.1 %	7.9 %	5.3 %	9.9 %
	7	5.4 %	13.3 %	6.2 %	16.5 %
	8	9.9 %	42.6 %	8.1 %	35.9 %

Table 2: Percentage of misclassified instances per number of systems (trained on the 70-30 % CASE partition or the full CASE train set).

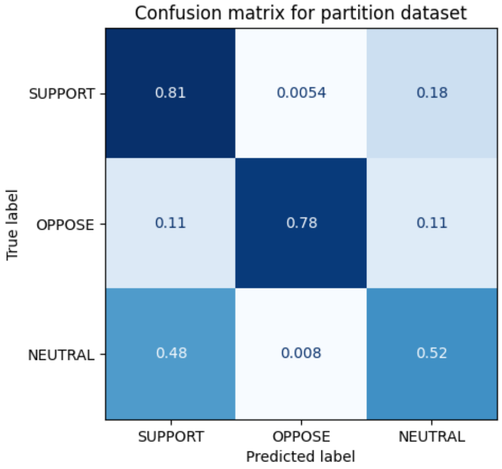

Figure 1: Confusion matrix for (1) model.

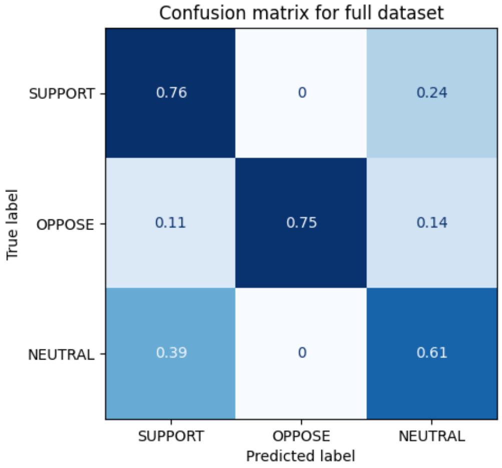

Figure 2: Confusion matrix for (2) model.

Furthermore, if we limit ourselves to studying only the misclassified instances, Table 3 shows the percentage of misclassified instances for each class for model (1). For example, the value of

37.53 % in the first row means that 37.53 % of the misclassified instances were Support and have been classified as Neutral. In the same way, Table 4 shows the percentage of misclassified instances for each class for model (2).

		Predicted label		
		Support	Oppose	Neutral
True label	Support	-	1.12 %	37.53%
	Oppose	3.60 %	-	3.37 %
	Neutral	53.48 %	0.90 %	-

Table 3: Percentage of instances misclassified by model (1) per class, over the set of misclassified labels.

		Predicted label		
		Support	Oppose	Neutral
True label	Support	-	0 %	48.66 %
	Oppose	3.35 %	-	4.46 %
	Neutral	43.53 %	0 %	-

Table 4: Percentage of instances misclassified by model (2) per class, over the set of misclassified labels.

5 Discussion

From the results shown in Table 1 it can be seen that all models outperform the Baseline model by quite some distance. This could be expected since the Baseline model has far fewer parameters than the Llama 2 7B model. Moreover, our best model obtains the 7th position in the leaderboard, only 0.0117 behind the leading model for this task.

As for using the CASE partition or the total data we see that although using all the data is how the best result is obtained, only 3 of the 8 models improve. In particular the Chat models improve with a classification layer at the end and without using the prompts system. However, the difference in performance is quite small.

Regarding the addition of SemEval data, if we look model by model, we see that the performance is only improved in 2 of them. The difference between adding the data at most worsens 0.0217 and at most improves 0.0163. This could be because the SemEval dataset contains about 3k examples of various topics compared to 7k in CASE. Of the SemEval dataset, only 13.5% of the data was related to climate activism and 86.5% was related to another topic. This data distribution may add noise to the training. This is why the models do not specialise in stance detection on Climate Activism as much as using only Climate Activism data. However, when adding the data, there is a considerable increase in training times.

As for the 4 different approaches to the Llama 2 model that have been used. As the seq2seq results are the lowest, we can conclude that it is better to remove the layer that allows to obtain a text sequence and replace it by a classifier layer. In addition, although the 7B Chat classifier models were the best performers, model 7B shows results with less variation when more data is added.

Looking at the results in Table 2 we can see that in the partition 9.9 % of the instances are incorrect for all models compared to 8.1 % if all data is used. However, when using partitioning we see that 42.6 % of the instances are classified well by all models, compared to 35.9 % if all data is used. All systems as a whole classify better if partitioning is used than if all data is used. This is consistent with the previous discussion as only 3 of the 8 models improve when using all data.

Comparing the confusion matrices in Figure 1 and Figure 2 we can see that for the Support and Oppose instances the model trained with the partition classifies better than the model trained with all. However, the latter classifies better the Neutral instances, thus obtaining the F1 difference between both.

Regarding the percentages of misclassified instances per class collected in Table 3 and Table 4 both models have little tendency to misclassify end-to-end (real label Support and predicted label Oppose or vice versa). Almost all misclassified instances are due to the Neutral label.

6 Post-competition analysis

Since this is a generative model, we could use a zero-shot approach. However, using this approach Llama 2 7B Chat model obtained an F1 result of 0.5685. This result is somewhat higher than the Baseline model proposed by the organisers, but significantly lower than the Fine-Tuned models.

In addition, adding the SemEval collection to the models caused a decrease in the performance of the models. One of the reasons could be due to the use of data not related to climate change. For this reason, the best architecture (7B Chat - clf no prompt) was re-trained by adding only the SemEval climate change related data. This model obtained an F1 of 0.7346, only 0.002 below the model that not use additional data. Looking more closely at these two models we could see that there was only a difference of two misclassified instances. By carefully studying the structure of the SemEval-2016 dataset and the CASE dataset, we realise that there is a temporal difference between the instances of both datasets. The CASE dataset contains terms such as Greta Thunberg or the Ukrainian-Russian war that SemEval does not. In addition, there are hashtags such as #FridaysForFuture or #ClimateStrike which are movements started in 2018. Therefore both datasets contain different lexical fields.

7 Conclusions and Future Work

In this paper, we have reported our participation in CASE in the framework of EACL 2024 in the stance detection subtask. For this task we have compared the performance of several variants of Llama 2 models and studied the effect of adding more data to the models.

Our results are significantly better than the proposed Baseline model and we have found that for this classification task it is better to dispense with the seq2seq structure of Llama 2 and use a classifier layer. We have also seen that adding more data tends to make the models behave worse.

As lines of future work it would be interesting to make an ensemble of all the models and analyse the performance of the models by training with different percentages of the CASE dataset. As the smaller Llama 2 model has been used, it would also be interesting to test these architectures with the larger Llama 2 models, 13B and 70B. In addition, to be able to use several related collections. If they are spaced in time, more robust semantic dimensions could be studied or datasets close in time could be used.

Limitations

The models described have been trained using only English text. For this reason, if a different language is used, good results may not be obtained. Additionally, the number of GPUs, the time required for training and inference, and the energy needed are resources that not everyone may have access to.

Acknowledgements

This work was supported by the HAMiSoN project grant CHIST-ERA-21-OSNEM-002, AEI PCI2022-135026-2 (MCIN/AEI/10.13039/501100011033 and EU “NextGenerationEU”/PRTR).

References

Abeer AlDayel and Walid Magdy. 2020. Stance detection on social media: State of the art and trends. *CoRR*, abs/2006.03644.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Iain J Cruickshank and Lynnette Hui Xian Ng. 2023. Use of large language models for stance classification. *arXiv preprint arXiv:2309.13734*.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. hillary: What went viral during the 2016 US presidential election. *CoRR*, abs/1707.03375.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2023. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *arXiv preprint arXiv:2305.13047*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3945–3952.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. *Preprint*.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Shuvam Shiwakoti, Hariram Veeramani, Raghav Jain, Guneet Singh Kohli, Ali Hürriyetoğlu, and Usman Naseem. 2024. Stance and hate event detection in tweets related to climate activism - shared task at case 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models, 2023.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text.