LTG at SemEval-2025 Task 10: Optimizing Context for Classification of Narrative Roles

Egil Rønningstad
Department of Informatics
University of Oslo, Norway
egilron@uio.no

Gaurav Negi
Data Science Institute
University of Galway
gaurav.negi@insight-centre.org

Abstract

Our contribution to the SemEval 2025 shared task 10, subtask 1 on entity framing, tackles the challenge of providing the necessary segments from longer documents as context for classification with a masked language model. We show that a simple entity-oriented heuristics for context selection can enable text classification using models with limited context window. Our context selection approach and the XLM-RoBERTa language model is on par with, or outperforms, Supervised Fine-Tuning with larger generative language models.

1 Introduction

Shared task 10, subtask 1 for SemEval 2025 (Piskorski et al., 2025) presents annotated news articles in Bulgarian, English, Hindi, (European) Portuguese, and Russian. For each article, a number of entities are identified and annotated for narrative roles. These entities are in the text framed as playing one of three main narrative roles; protagonists, antagonists or innocent. For each main role, there is a number of fine-grained roles, 22 in total. Each entity may be given multiple fine-grained roles available for the assigned main role. Further details on the task can be found in the technical report (Stefanovitch et al., 2025). By classifying how entities are framed in news articles, we can better understand reporting angles and identify bias variations between different news sources.

Our contribution focuses on the following challenges in modeling the narrative roles of the annotated entities:

a) Each article may contain much irrelevant text with respect to a given entity.

b) A sentence mentioning an entity may also mention other entities and frame each differently.

c) The text segment(s) contributing to the entity framing may span multiple sentences.

As seen in Table 1, the provided training examples are in the hundreds and the thousands for each language. We therefore hypothesize that XLM-RoBERTa-large (XLM-R), a multilingual masked language model, could be a cost-effective starting point. Due to this, we pose the following research questions:

- **RQ1:** Can we find rule-based approaches that mitigate the above mentioned challenges, including irrelevant text and documents that will not fit in the XLM-R context window?
- **RQ2:** Can a fine-tuned XLM-R-based model be outperformed by larger language models where the entire document fits well within the context window?

To answer these questions, our paper presents results from experiments regarding text preprocessing where we fine-tune XLM-R on various text segments and evaluate each pre-processing strategy. These results are compared against zero-shot prompting of a large language model (LLM), and Supervised Fine-Tuning of LLMs with 7-8 billion parameters.

1.1 Context Optimization

Extracting only the relevant text segments for a given task can be named “context optimization”. Studies have shown that for long-context LLMs, irrelevant or distractive text as part of an input, reduces model performance (Shi et al., 2023; Wu et al., 2024; Cai et al., 2024; Wang et al., 2024). This task is also being described as “context rewriting” (Wang et al., 2024) or “prompt compression” (Liskavets et al., 2024), but for this paper we prefer the term “context optimization”. For MLMs this task is imperative due to the limited context window of, as for XLM-R, 512 subword tokens, which is not enough for many texts longer than microblog messages and short user-submitted reviews.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 440–447
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

1.2 Supervised Fine-Tuning

Supervised Fine-Tuning (SFT) has proven to to be a viable path for creating text classification models. Depending on the training examples and compute resources available, this approach may lead to better classification than in-context learning where pairs of example text and labels are fed to a LLM before requesting the classification of a new example (Mosbach et al., 2023).

1.3 Dataset

The data for this shared task were released sequentially. The experiments reported here, are trained with the finalized train split and evaluated on the labeled dev split. The number of labeled entities for training and evaluation is found in Table 1.

Language	train	dev
BG	625	30
EN	686	91
HI	2331	280
PT	1245	116
RU	366	86
all	5253	603

Table 1: Annotated entities per language in the train and dev splits for the experiments reported on in this paper.

2 Dataset Pre-Processing

Our contribution to the shared task was trained exclusively on the provided data. To address the above mentioned challenges for narrative role classification and answer RQ1, we experimented with how to best prepare the data for fine-tuning and classification.

2.1 Text Span Extraction

We hypothesize that the text relevant to the entity framing would be located in proximity to the entity mention. As the provided texts are split in sentences and paragraph, and the entities in question are pre-identified within the text, we compare a selection of rule-based context extraction approaches, and measure their performance against a simple LLM-generated baseline. To assess the value of such text span extraction, we performed experiments with the following alternative text extraction heuristics:

a) **Single sentences** For each annotated entity, provide only the sentence where the entity is mentioned.

b) **Single paragraph** For each annotated entity, provide the paragraph in which the entity occurs.

c) **Entire text** For each annotated entity, provide the entire document. The model consumes as much text as possible, ignoring the rest.

d) **Entity-to-entity (ent2ent)** For each annotated entity, provide the sentence where the entity is mentioned, and all subsequent sentences until a new entity occurs.

e) **GPT-extracted** For each annotated entity, provide the replies through the api of ChatGPT with gpt-4o, queried to extract the text span(s) containing information regarding the narrative role of the entity.

2.2 Merging languages

When it comes to classification tasks with a multilingual pretrained model, there is the tradeoff between getting a larger training set, and introducing “noise” from the language variety (Conneau et al., 2020; Rønningstad, 2023). We therefore prepared one dataset per language, and one merged dataset containing all languages.

2.3 Preparing data for SFT

We prepared one prompt per text document for Supervised Fine-Tuning.

Prompt Template The prompt template (see Appendix B) used for making the predictions consists of the following segments: (i) Annotation Instructions, (ii) Taxonomy of the primary and secondary entity roles, (iii) The definitions of primary roles, (iv) Document input along with the entities, and (v) Output format.

It should be noted that the entities in question, occur multiple times in a few documents. To localize the correct instance of these entities an entity tag (<entity> </entity>) is placed around the occurrence using the index values provided in the annotations. This processed document is then included in the prompt.

3 Modeling

We here present the various modeling approaches tested, and their evaluation results. The focus is on answering the research questions by applying the

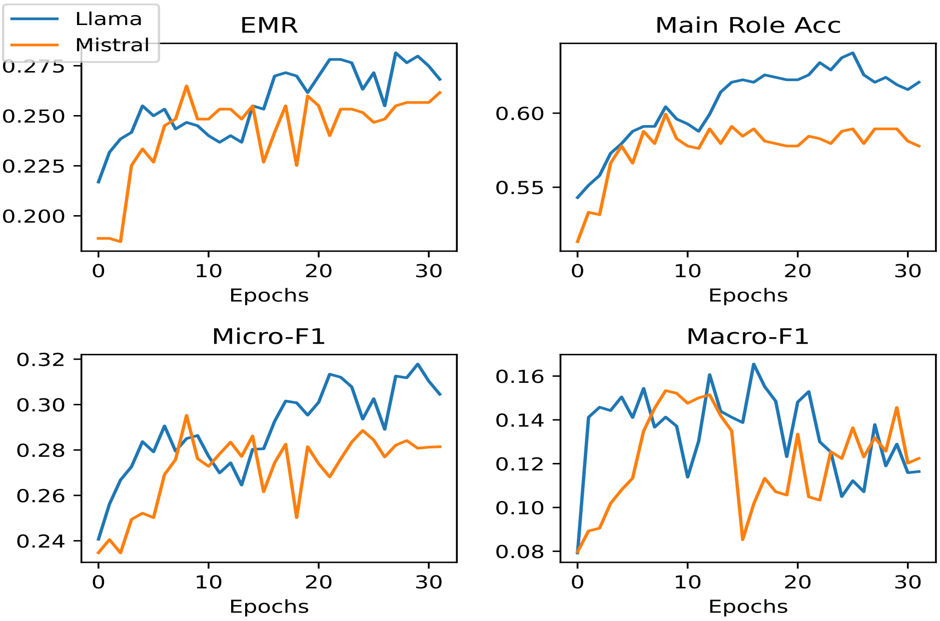

Figure 1: SFT learning trends per epoch as evaluated on dev dataset

various text pre-processing approaches as input for XLM-R fine-tuning. These approaches are compared with classification based on larger language models.

3.1 Modeling with XLM-R

XLM-R-large[1] was used for all XLM-R experiments. No hyperparameters were altered from the standard settings for text classification in the Transformers library. Models were trained for 10 epochs.

Prepend Entity Mention. The text spans provided to the model may be identical when classifying with respect to two different entities, and we therefore prepended the extracted segment with the prefix *"Regarding <entity> :\n"*, where <entity> is placeholder for the entity mention as written in the task annotations. This was done for all XLM-R experiments except for *ent2ent_noprefix*.

We see in Table 2, how the micro F_1 scores were dramatically reduced when there was no prefix presenting the entity in question.

Monolingual vs Multilingual Fine-Tuning. As can be seen in Table 3 and Figure 3 in Appendix A, fine-tuning on all languages improved results noticeably. For the languages with the smallest training set, there were no measurable Micro F_1 results. For subsequent experiments, the training data is understood to consist of all languages in the shared task.

Modeling Main Roles First As there are 22 fine-grained roles in total, and an entity may be labeled with multiple fine-grained roles, we attempted to employ the best-so-far method (Entity-to-entity with prefix) in a two-step modeling and inference approach (main2fine). We first trained a classifier for the three main roles, and predicted main roles for the dev set (one per entity). After separating the dev set into each predicted main role, we trained one multilabel classifier per predicted main role. This reduced the number of possible fine-grained roles to 6 for the Protagonist main role, 12 for Antagonist and 4 for Innocent. The results are found in Table 2, the row *ent2ent_main2fine*.

3.2 Prompting an LLM for Classification

The results from the XLM-R-based models were compared against the labels provided by ChatGPT-4o when queried for classifying one entity at a time, including the entire article and the label definitions in the prompt, but not any examples. The results are found under *gpt-inference* in Table 2.

3.3 Modeling with LLMs and LoRA

For these experiments, we used the instruction-tuned versions of Llama-3.1-8B[2] and Mistral-7B[3] models. These models were further fine-tuned with the training dataset that was provided.

[1] FacebookAI/xlm-roberta-large

[2] meta-llama/Llama-3.1-8B-Instruct

[3] mistralai/Mistral-7B-Instruct-v0.3

Model	Method	BG	EN	HI	PT	RU	All
XLM-R	ent2ent	25.40	31.25	46.49	70.00	**47.73**	**47.75**
	ent2ent_noprefix	15.87	11.46	27.07	58.92	26.29	30.11
	ent2ent_main2fine	25.40	25.00	41.75	**74.69**	40.68	44.51
	sentence	31.75	20.73	**46.88**	70.83	42.46	46.06
	gpt-extracted	25.40	17.71	45.36	64.73	40.00	43.14
	paragraph	25.40	16.75	40.00	67.50	38.64	40.79
	fulltext	28.57	9.42	40.53	62.50	37.29	38.96
ChatGPT-4o	gpt-inference	38.96	**36.95**	36.85	65.64	34.65	41.78
llama	SFT	**43.48**	22.33	20.57	50.00	29.70	31.78
mistral	SFT	36.36	30.77	18.71	49.17	34.83	29.52

Table 2: Evaluation results for the various text extraction and modeling strategies for each test language. Evaluation metric is Micro F_1 for the 22 fine-grained roles. We see that fine-tuning XLM-R on prefixed text segments using the *ent2ent* segment extraction strategy yields the best overall results. All XLM-R experiments contain prefixed segments except the *ent2ent_noprefix* ablation experiment. All experiments classify the 22 fine-grained roles directly, except the *ent2ent_main2fine* experiment.

language	all	in-lang	samples
RU	52.33	0.00	366
BG	26.67	0.00	625
EN	29.67	8.79	686
PT	69.83	44.83	1245
HI	43.21	22.50	2331
all	46.77		5253

Table 3: A comparison of test results (Micro F_1) with XLM-R and the ent2ent context optimization, when training either only on the training data of the test split language (in-lang) with their samples count (samples), or training on the entire train set (all). More results are presented in Figure 3 in Appendix A.

We use Low-Rank Adaptation (LoRA) for the parameter-efficient fine-tuning using the causal language modeling objective. The values prescribed for LoRA parameters in the introductory work were used ($r = 32$ and $\alpha = 64$). The parameters of the models (Llama and Mistral) were loaded with a 4-bit precision configuration. Since the models have multi-lingual capabilities, we combined the training dataset across all languages for fine-tuning.

We fine-tuned LLMs for 32 epochs and selected the checkpoint with the highest Exact Match Ratio (EMR) to make the predictions. We track the metrics on the language-combined development dataset in Figure 1. Per-language final evaluation results are found in Table 2.

4 Analyses

We here present our reflections on the results reported in the paper.

4.1 Finding the Best Text Segments

Table 2 shows that modeling with XLM-R and the ent2ent-approach for segment extraction yields the best overall results. Although this approach is the best alternative for only one of the languages (Russian), we submit these results for all languages. Among our experiments, the ent2ent-approach yields competitive results for all languages except for Bulgarian where the experiments based on larger models are noticeably better. Our contribution ranked from fifth (Hindi and Portuguese) to tenth (English) on the official SemEval results on the test set.

4.2 Performance vs train split size

All our approaches yielded large variations in results between languages. The training data as presented in Table 1 are very different in size per language. Although models were trained on the entire dataset, having data from the same language as the test data (in-language training data) has from experience proven to be important. We therefor compare the results with the size of the in-language train split. Figure 2 shows no clear trend, as Portuguese yielded the best results, while Hindi had almost twice the amount of training data. We see that the XLM-R-based approach is clearly best for the two languages with the most training data. The

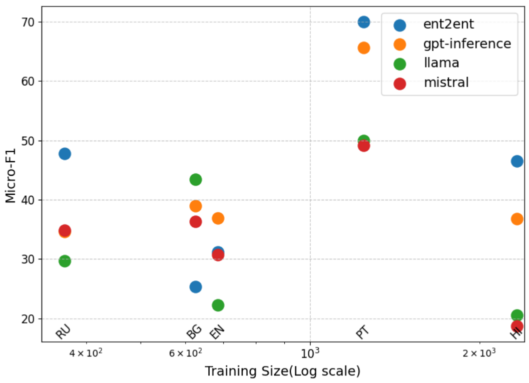

Figure 2: Model performance as a function of in-language training data

need for more than a thousand in-language, in-domain training samples when fine-tuning XLM-R, is in line with previous experience. But again, XLM-R performs best of our approaches for Russian as well, with the fewest training samples. We can assume that languages' presence in the original model pretraining also contributes much to the quality of the resulting inference. A surprise therefore, is all models' poor performance on English data. We can speculate that Portuguese hits a "sweet spot" between language similarity to much of the pretraining data, and size of the task's in-language dataset. But further inspection of the provided dataset's complexity across languages would be required before drawing any conclusion.

4.3 Supervised Fine-Tuning

We were surprised to see how hard it was to train better models than the XLM-R using SFT and Llama or Mistral. These fine-tuning cycles are resource-demanding, and in our prompts, as shown in Appendix B we provide one prompt per text, requiring the model to return all roles for all entities. To simplify the task, one could request classifications only one entity per prompt. Additionally, applying context optimization might prove beneficial here as well. Testing these options with SFT was beyond our resource allocations.

5 Conclusion

We have created a XLM-R-based multilingual model for Entity Framing as a part our contribution to the SemEval 2025 shared task 10 on Multilingual Characterization and Extraction of Narratives from Online News, subtask 1. This model was tested against Supervised Fine-Tuning of Llama and Mistral, and against zero-shot prompting of ChatGPT-4o. We found it imperative to extract entity-oriented text segments in order to effectively utilize XLM-R with long documents containing multiple entities each.

Acknowledgments

Parts of the work documented in this publication has been carried out within the NorwAI Centre for Research-based Innovation, funded by the Research Council of Norway (RCN), with grant number 309834.
Computations were performed in part on resources provided through Sigma2 – the national research infrastructure provider for High-Performance Computing and large-scale data storage in Norway.

References

Hongjie Cai, Heqing Ma, Jianfei Yu, and Rui Xia. 2024. A joint coreference-aware approach to document-level target sentiment analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12149–12160, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Barys Liskavets, Maxim Ushakov, Shuvendu Roy, Mark Klibanov, Ali Etemad, and Shane Luke. 2024. Prompt compression with context-aware sentence encoding for fast and improved llm inference. *arXiv preprint arXiv:2409.01227*.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Jorge Alípio, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on*

Semantic Evaluation, SemEval 2025, Vienna, Austria.

Egil Rønningstad. 2023. UIO at SemEval-2023 task 12: Multilingual fine-tuning for sentiment classification in low-resource languages. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1054–1060, Toronto, Canada. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

Jinlin Wang, Suyuchen Wang, Ziwen Xia, Sirui Hong, Yun Zhu, Bang Liu, and Chenglin Wu. 2024. Fact: Examining the effectiveness of iterative context rewriting for multi-fact retrieval. *arXiv preprint arXiv:2410.21012*.

Zijun Wu, Bingyuan Liu, Ran Yan, Lei Chen, and Thomas Delteil. 2024. Reducing distraction in long-context language models by focused learning. *arXiv preprint arXiv:2411.05928*.

Appendices

A Language-Wise Fine-Tuning

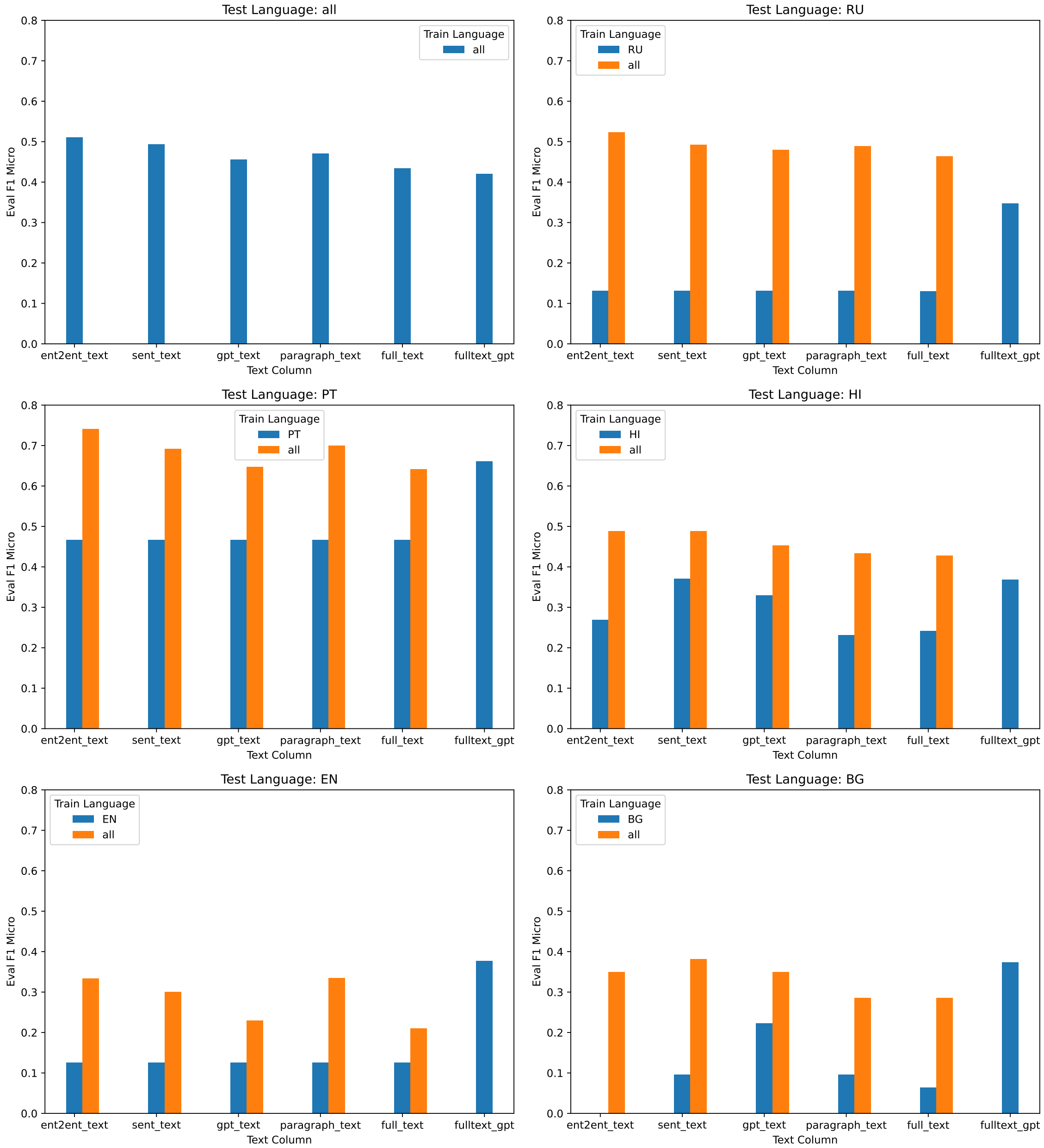

Figure 3: Benefits from training on all languages. For each language, the results improve substantially when fine-tuning the XLM-R on the entire dataset, as opposed to fine-tuning on the test language only.

B SFT Prompt Template

~~~
### Annotation Instructions:
You are given a document that includes various entities along with descriptions of events and actions. Your task is to analyze the text and determine the roles each entity plays according to the taxonomy provided below.

### Taxonomy:
**Protagonist**
- Guardian
- Martyr
- Peacemaker
- Rebel
- Underdog
- Virtuous
**Antagonist**
- Instigator
- Conspirator
- Tyrant
- Foreign Adversary
- Traitor
- Spy
- Saboteur
- Corrupt
- Incompetent
- Terrorist
- Deceiver
- Bigot
**Innocent**
- Forgotten
- Exploited
- Victim
- Scapegoat

### Definitions
- **Protagonist**: A central character or force in a positive role.
- **Antagonist**: A character or force in opposition to the protagonist.
- **Innocent**: Entities that are marginalized or victimized without any active role in the conflict.

### New Input:
<<variable input start>>
**LANG: EN**
**Document:**
According to the Gospel of the Global Warming Hoax, 1850-1910 was the coldest period of the past millennium. Yet glaciers were retreating rapidly. Now that the planet allegedly has a fever, the retreat has slowed dramatically and even reversed:
Our moonbat rulers canceled the Medieval Warm Period and Little Ice Age for failing to comply with climate ideology. But preventing glaciers from growing is more difficult than doctoring the historical record to support climate con man <entity>Michael Mann</entity>'s spurious hockey stick graph.
Nonetheless, prophet of doom <entity>Al Gore</entity> shouts that "we could lose our capacity for self-governance" if we don't surrender still more freedom to Big Government so that it can fix the supposedly broken weather.
On tips from Lyle and Wiggins.
Anyone can join.
Anyone can contribute.
Anyone can become informed about their world.

**Query entities:**
<entity>Michael Mann</entity>
<entity>Al Gore</entity>

### Now for this new document, extract the roles for the following entities:
["Michael Mann", "Al Gore"]
<<variable input end>>
### Output Format
```json
[["entity1", "primary role", ["secondary role 1", "secondary role 2"]],
["entity2", "primary role", ["secondary role 1", ..]]
..]```
"""
~~~