

LLM-Based Multi-Agent Models for Multiclass Classification of Strategic Narratives

Alberto Caballero¹, Roberto Centeno¹ and Álvaro Rodrigo¹

¹NLP & IR Group, Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain

Abstract

This paper details the deployment of a Large Language Model (LLM)-based multi-agent system for Task 2 of DIPROMATS at IberLef 2024, focusing on the multiclass multilabel classification of tweets into predefined international narratives. Despite challenges derived from complex narrative structures and limited data, our proposed approach, integrating a Signal Builder Agent and a Classification Agent, performed well in both English and Spanish. The effectiveness of this model in handling intricate narrative data demonstrates the potential of agent-based LLM architectures in multilingual narrative analysis, contributing significantly to the advancement of NLP to help actors navigate international relations contexts. Our research tackles the complexities of identifying and categorizing strategic narratives expressed through social media, a task essential for understanding geopolitical dynamics and influencing public opinion. By leveraging the advanced capabilities of LLMs, our system enhances the detection and interpretation of narrative elements within multilingual tweets. The Signal Builder Agent refines narrative signals through techniques such as keyword extraction and synthetic example creation, thereby improving the model's ability to generalize from limited data. Concurrently, the Classification Agent employs these enriched signals to accurately classify tweets into one of 24 distinct narratives, each representing nuanced geopolitical themes. Our model demonstrated significant improvements over baseline systems, achieving higher precision and recall across both English and Spanish datasets. This was evident in metrics such as F1-Strict, F1-Lenient, and F1-Average scores, which showed superior performance in narrative classification tasks. The successful integration of signal enhancement and decision-making processes in our multi-agent architecture underscores the robustness and adaptability of LLMs in complex, real-world applications.

Keywords

Large Language Models (LLMs), Multi-Agent Systems, Narrative Analysis, Strategic Narratives, Signal Enhancement

1. Introduction

Narrative analysis is crucial for understanding the strategic maneuvers of international actors through their public communications. These actors craft narratives to orchestrate shared meanings of past, present, and future events, influencing both domestic and international policy landscapes [1]. In the digital age, social media platforms, such as X, serve as battlefields for

IberLEF 2024, September 2024, Valladolid, Spain


✉ acaballer382@alumno.uned.es (A. Caballero); rcenteno@lsi.uned.es (R. Centeno); alvarory@lsi.uned.es (Á. Rodrigo)

🌐 <https://scholar.google.com/citations?user=30xziTkAAAAJ> (A. Caballero);

<http://nlp.uned.es/~rcenteno/indice.php> (R. Centeno); <https://sites.google.com/view/nlp-uned/people/> (Á. Rodrigo)

🆔 0000-0001-9095-4665 (R. Centeno); 0000-0002-6331-4117 (Á. Rodrigo)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

disseminating and contesting these narratives, which makes the ability to automatically classify such narratives critically important.

The IberLEF forum [2], particularly DIPROMATS [3], provides a platform to tackle this challenge. This year’s competition featured two tasks: the first focused on identifying and analyzing narrative elements in multilingual texts, while the second, which we participated in, centered on the complex problem of multiclass multilabel classification. Specifically, participants were required to identify predefined narratives constructed by various international actors that individual tweets supported. The multilingual nature of the dataset, spanning English and Spanish, and the limited examples provided for model training, added to the task’s complexity.

Our research aimed to address the second task using an LLM-based multi-agent model designed to handle the nuanced requirements of narrative classification. Our model needed to distinguish between 24 distinct narratives—six for each major international actor as defined in the contest guidelines—while managing the linguistic and contextual diversity of the narratives. Each narrative encapsulated complex geopolitical themes, ranging from power dynamics and historical accounts to cultural significance and political ideologies. For example, narratives like “The West is immoral, hostile, and decadent” under Chinese narratives, or “Russia leads an alternative system to that sponsored by the West” under Russian narratives, demonstrated this complexity.

This paper details our approach to the contest, focusing on the development and deployment of our LLM-based multi-agent system [4]. Our system leverages advanced natural language understanding capabilities to efficiently interpret and classify multi-thematic and multilingual data. By integrating a Signal Builder Agent with a Classification Agent, our model enhanced both the detection and classification accuracy of narrative elements across various languages. Notably, this was achieved without the need for fine-tuning, typically required for such tasks. Instead, we effectively used a limited training dataset (54 samples in English and 48 samples in Spanish) as referential context for the system, allowing us to distill relevant information and guide the decision-making process of the agent efficiently.

2. LLM-based Multi-Agent Approach

Our methodology for addressing the classification challenge in Task 2 employs a novel LLM-based multi-agent approach. This architecture segments the task into distinct processes, each managed by specialized agents, as described in [5]. Specifically, the Signal Builder Agent and the Decision-Making Agent each play differentiated roles, effectively handling the complexities of multiclass multilabel classification across multilingual datasets.

A general overview of the architecture can be gathered by looking at Figure 1.

2.1. Signal Builder Agent

The primary function of the Signal Builder Agent is to enhance the detectability of narrative elements within the tweets [6]. Given the few-shot learning nature of the task, where only a limited number of examples are provided for each narrative, this agent employs advanced natural language processing techniques to extrapolate and amplify narrative signals available within the originally provided dataset.

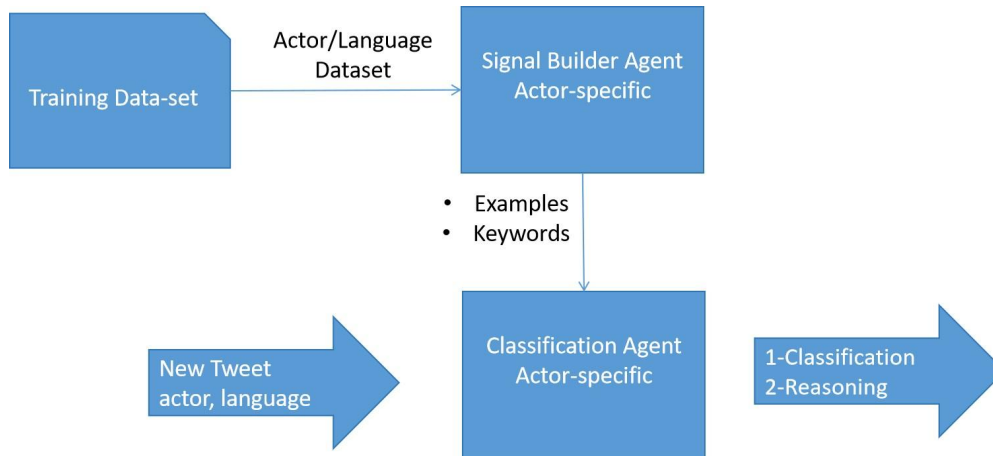


Figure 1: System Architectural Overview

Signal Enhancement Techniques Used:

- Keyword Extraction

This technique involves identifying and extracting significant words or phrases that are strongly associated with each narrative. These keywords serve as condensed representations of the narratives, helping to bridge the gap between limited examples and the model's understanding of the narrative context.

- Synthetic Example Creation

The agent utilizes techniques such as paraphrasing and semantic similarity to expand the initial set of examples, producing a synthetic example for each different narrative.

The output from the Signal Builder Agent consists of enriched narrative signals, enhancing the original data and making it more robust against the variability in new, unseen tweets.

2.2. Decision Making Agent

Following signal enhancement, the Decision-Making Agent takes over to perform the actual classification of tweets into the respective narratives. This agent integrates the enriched signals with the LLM's capabilities to make informed classification decisions.

Classification Process Implemented:

- Input Integration

The agent receives both the original tweet and the enhanced signals as input. This dual-input strategy ensures that the decision-making process benefits from both the raw data and the processed insights.

- LLM Utilization

Leveraging the natural language understanding capabilities of the LLM, the agent uses contextual cues from the enhanced signals to classify tweets. The model is prompted with narrative-specific queries that include both the tweet and the associated signals to determine the most likely narrative classifications.

- Decision Logic

The agent employs a Chain-of-Thought prompting strategy (CoT) [7] to reason through the different potential narratives. This reasoning forms part of the output generated by the agent, alongside the predicted label.

3. Analysis of Results

This section provides a detailed comparison of our proposed LLM-based multi-agent approach, against other participating models in the competition. The comparison focuses on the final evaluation results across various models, including open-source, zero-shot, and few-shot learning models. [8]. The performance metrics are evaluated using three types of F1 scores: F1 Strict, F1 Lenient, and F1 Average, across both English and Spanish. 'F1 Strict' measures precision and recall at a stricter criterion, requiring an exact match between the predicted and actual data. 'F1 Lenient' allows for partial matches, thus providing a more forgiving assessment. 'F1 Average' calculates the mean of the 'Strict' and 'Lenient' scores, offering a balanced view of overall performance.

The results presented in Table 1 below illustrate the performance of all models participating in the competition, including the Mixtral 8x7B model, which serves as the baseline established by the organizers. For simplicity, and because it is the main metric used in the competition, this table exclusively utilizes the F1-Strict score, providing the most demanding reference for comparison.

Model	F1-Strict En	F1-Strict Es
LLM Multi-Agent Model (GPT-4)	0.5643	0.4831
Mixtral 8x7B (Baseline)	0.3769	0.2875
umuteam-Zephyr	0.3046	0.3149
umuteam-TuLu	0.2729	0.2303

Table 1

Model Results Comparison.

Our proposed model significantly outperformed all competitors in both languages and across all evaluated metrics (F1-Strict, F1-Lenient, and F1-Average). This superior performance can likely be attributed to two key factors: the use of the advanced LLM, GPT-4, and the robustness of our multi-agent approach, which will be discussed further in the development phase. Specifically,

the integration of the Signal Builder Agent and the Decision Making Agent has effectively enhanced narrative signals and adapted to the complexities of the multilabel classification task.

The dynamic signal processing and decision-making capabilities of our proposal have proven essential in achieving its high performance. These results highlight our model's ability to not only recognize and classify narratives accurately but also adapt to the nuanced variations across different languages and narrative styles. This adaptability is particularly noteworthy when compared to other models in the contest, which often relied on less dynamic, open-source versions of LLMs or simpler few-shot learning methodologies.

4. Discussion

The participation of our team using the LLM-Based Multi-Agent Model underscores a significant advancement in the field of narrative analysis within international relations. Our model's performance across multiclass, multilabel, and multilingual classification tasks in both English and Spanish demonstrates the effectiveness of this methodology and shows how LLMs can be used for these tasks even in scenarios where very limited data is available.

During our internal experiments using the initially provided training dataset, we conducted a comparative analysis of various models, all based on the same Generative Pre-trained Transformer (GPT-4) and employing the Chain-of-Thought (CoT) prompting strategy. This analysis revealed that specific signal enhancement strategies significantly influenced performance outcomes. Model configurations that incorporated keywords and customized examples—referred to as the proposed model—achieved the highest F1-Strict scores, recording 0.86 in English and 0.81 in Spanish. These results underscore the critical importance of dynamic input processing in maximizing the efficacy of large language models (LLMs), as detailed in Table 2.

In contrast, models based on the initial examples provided in the narrative descriptions (Original Examples) and an alternative model that searched for semantically similar examples in the provided dataset (Semantic Search Examples) exhibited lower performance metrics.

Tested Models	F1_Strict-English	F1_Strict-Spanish
CoT - Original Examples	0.75	0.64
CoT - Semantic Search Examples (n=4)	0.71	0.69
CoT - Key Words (n=10)+ Customized Examples (n=1)	0.86	0.81

Table 2
Performance of GPT-4 Model on Different Architectural Designs

These findings not only validate the superiority of using a sophisticated, proprietary LLM in complex classification scenarios but also underscore the necessity of integrating advanced signal processing techniques to amplify the LLM's inherent capabilities. The LLM-Based Multi-Agent Model sets a benchmark for future developments in automated narrative analysis systems and opens new avenues for further research into enhancing LLM performance through strategic signal manipulation and multi-agent architectures.

4.1. Conclusion and Future Work

Looking ahead, integrating more complex multi-agent systems and advanced signal processing techniques presents a promising direction for developing AI systems capable of mimicking human reasoning but with greater scalability and explainability. By enhancing the model's architecture to include additional NLP tools such as semantic reasoning engines and context-aware processing units, we can approach the subtlety and depth of human cognitive processes. This progression will allow AI systems not only to detect and classify data but to understand and interact with it in a fundamentally human-like way, albeit at a scale and speed unmatched by human analysts.

Such developments could lead to breakthroughs in how AI systems manage the vast and nuanced streams of data in global discourse, making them indispensable tools for real-time decision-making in complex scenarios such as diplomatic negotiations and international policy-making.

In conclusion, our research demonstrates that the strategic enhancement of narrative signals, coupled with the dynamic capabilities of modern LLMs like GPT-4, can significantly improve the accuracy and efficiency of narrative detection and classification. This approach promises to revolutionize the field of NLP by providing more precise, adaptable, and effective tools for understanding and managing the flow of narratives in international discourse.

Acknowledgments

This work has been partially funded by the Spanish Research Agency (Agencia Estatal de Investigación), through the DeepInfo project PID2021-127777OB-C22 (MCIU/AEI/FEDER, UE) and the HOLISTIC ANALYSIS OF ORGANISED MISINFORMATION ACTIVITY IN SOCIAL NETWORKS project (PCI2022-135026-2).

References

- [1] B. O'Loughlin, A. Miskimmon, L. Roselle, *Strategic Narratives: Communication Power and the New World Order*, 2013. doi:10.4324/9781315871264.
- [2] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024)*, co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.
- [3] P. Moral, J. Fraile, G. Marco, A. Peñas, J. Gonzalo, Overview of DIPROMATS 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers, *Procesamiento del Lenguaje Natural* 73 (2024).
- [4] Z. Wang, Y. Yu, W. Zheng, W. Ma, M. Zhang, Multi-agent collaboration framework for recommender systems, *arXiv preprint arXiv:2402.15235* (2024).
- [5] S. Han, Q. Zhang, Y. Yao, W. Jin, Z. Xu, C. He, Llm multi-agent systems: Challenges and open problems, *arXiv preprint arXiv:2402.03578* (2024).

- [6] Y. Zhang, Q. Yang, A survey on multi-task learning, *IEEE Transactions on Knowledge and Data Engineering* 34 (2022) 5586–5609. doi:10.1109/TKDE.2021.3070203.
- [7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* 35 (2022) 24824–24837.
- [8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.