DRAGON at FIGNEWS 2024 Shared Task: A Dedicated RAG for October 7th Conflict News

Sadegh Jafari,[1*] **Mohsen Mahmoodzadeh,**[2*] **Vanooshe Nazari**[1*]
Razieh Bahmanyar,[1*] **Kathryn Burrows**[3*]
[1]Iran University of Science and Technology
[2]Ferdowsi University of Mashhad [3]Madonna University
{sadegh_jafari@comp,V_nazari@ind,bahmanyar_r99@alumni}.iust.ac.ir
mahmoodzadeh.mohsen@mail.um.ac.ir kateburrows1975@gmail.com

Abstract

In this study, we present a novel approach to annotating bias and propaganda in social media data by leveraging topic modeling techniques. Utilizing the BERTopic tool, we performed topic modeling on the FIGNEWS Shared-task dataset, which initially comprised 13,500 samples. From this dataset, we identified 35 distinct topics and selected approximately 50 representative samples from each topic, resulting in a subset of 1,812 samples. These selected samples were meticulously annotated for bias and propaganda labels. Subsequently, we employed multiple methods like KNN, SVC, XGBoost, and RAG to develop a classifier capable of detecting bias and propaganda within social media content. Our approach demonstrates the efficacy of using topic modeling for efficient data subset selection and provides a robust foundation for improving the accuracy of bias and propaganda detection in large-scale social media datasets.

1 Introduction

In response to the evolving landscape of media representation and discourse surrounding the Gaza-Israel 2023-2024 war, the FIGNEWS Shared task (Zaghouani et al., 2024) has been initiated to delve into the intricate nuances of bias and double standards prevalent in news articles. This task aims to explore diverse perspectives, cultures, and languages, fostering a comprehensive understanding of these events through the lens of major news outlets across the globe. The overarching objective is to establish a shared corpus for comprehensive annotation across various layers, crafting annotation guidelines shaped by the diverse range of conflicting discourses around this sensitive topic. This endeavor seeks to highlight both challenges and commendable aspects within the data, fostering a collaborative community and nurturing the growth of the next generation of NLP researchers.

Ding et al. (2022) demonstrated the potential of GPT-3 for cost-effective data annotation, suggesting that models trained on GPT-3 annotated data can perform comparably to those trained on human-annotated data. Zhang et al. (2021) provided a comprehensive survey on machine learning techniques for automating the labeling of video, audio, and text data, addressing the high cost of manual data annotation and proposing future research directions. Conforti et al. (2020) emphasized the importance of qualitative data for sustainable development projects and introduced the UPV Classification task for annotating qualitative interviews. Anglin et al. (2020) compared complex and simple human annotation schemes, finding that complex schemes are more efficient and precise. Perry (2021) discussed the commercial success of LightTag, a text annotation platform that optimizes the NLP process. Recent research by Šuppa et al. (2024) explored using large language models for classifying tweets on climate activism, comparing zero-shot and few-shot learning approaches. Liu et al. (2021) introduced KATE, a strategy that improves model performance by using similar examples as context for large language models. Zhang et al. (2023) proposed a RAG [1] module to enhance financial sentiment analysis by providing richer context from trusted sources.

Our approach aligns with the goals of the FIGNEWS Shared-task by utilizing topic modeling, specifically the BERTopic (Grootendorst, 2022) tool, to select a well-generalized and suitable subset of social media data for annotation. From an initial dataset of 13,500 samples, we identified 35 distinct topics and selected approximately 50 samples from each topic, resulting in a final subset of 1,812 samples. These samples were annotated for bias and propaganda labels. We then used multiple

[*] These authors contributed equally to this work.

[1] Retrieval Augmented Generation

methods, including RAG, KNN [2] (Cover and Hart, 1967), SVC [3] (Hearst et al., 1998), and XGBoost (Chen and Guestrin, 2016), to create a classifier for these labels. Our study demonstrates the efficacy of using topic modeling for efficient data subset selection and provides a robust foundation for improving the accuracy of bias and propaganda detection in large-scale social media datasets.

2 Annotation Methodology and Examples

In this section, we describe the pipeline for selecting suitable samples for human annotation. After annotating this subset of data, we use it to train multiple models. These models, along with the RAG technique, are then employed to annotate the rest of the dataset using a voting mechanism. The implementation details can be found on the GitHub page [4].

2.1 Development of Annotation Guidelines

Label	Definition
Unbiased	The news presents information neutrally without favoring any particular side.
Biased against Palestine	News that blames the Palestinian side for breaking peace talks and ignoring international laws preventing attacks against non-militant citizens.
Biased against Israel	News that criticizes Israeli attacks against non-militant citizens and accuses them of war crimes.
Biased against both Palestine and Israel	News questioning the activities of both sides and expressing opinions on the loss of peace due to their irresponsibility.
Biased against others	News containing complaints against other countries, people, and organizations.
Unclear	News indicating its stance non-clearly or exhibiting ambiguity.
Not applicable	Topics not directly related to war conflicts or bias annotations.

Table 1: Definitions of Bias Tags.

The creation of annotation guidelines was a meticulous process designed to ensure clarity and consistency across the annotation tasks. We started by defining the objectives for annotating bias and propaganda in news articles associated with the Israel-Gaza war. For bias, the goal was to establish a corpus that facilitates the study of potential biases, categorized into 7 distinct labels. For propaganda, the objective was to prepare a corpus to analyze various forms of propaganda within the news, categorized into 4 labels. To achieve this, we crafted detailed definitions and examples for each label. In Table 1 we have definitions for the Bias task, and also in Table 2 we have definitions for the Propaganda task. These guidelines included label-specific examples for both subtasks, ensuring that annotators had a clear reference for each label. This approach was intended to standardize the annotation process and minimize subjectivity.

Label	Definition
Propaganda	Information, especially biased or misleading, are used to promote a particular political cause or point of view.
Not Propaganda	Information or communication that is factual, neutral, or unbiased.
Unclear	Information that is ambiguous, vague, or difficult to determine its intended purpose or bias.
Not Applicable	Information that does not fall under the category of propaganda or its opposite; it is unrelated to the promotion or criticism of a political cause or viewpoint.

Table 2: Definitions of Propaganda Tags.

2.2 Data Annotation Process

The annotation process involved several steps to ensure accuracy and consistency. Initially, we selected a subset of 1,812 samples from the FIGNEWS Shared-task dataset, using the BERTopic(using e5-base (Wang et al., 2024) sentence transformer as feature extractor) tool to identify 35 distinct topics and choosing approximately 50 samples from each topic. The distribution of Bias and Propaganda annotations is shown in Tables 3 and 4.

Tag name	Train	Test
Unbiased	1224	159
Biased against Palestine	118	22
Biased against Israel	99	13
Not Applicable	73	8
Biased against others	48	5
Unclear	30	5
Biased against both Palestine and Israel	6	2
Total	**1599**	**213**

Table 3: Distribution of Bias labels in Train and Test datasets.

One of the problems with this subset dataset

[2] K Nearest Neighbors

[3] Support Vector Classifier

[4] https://github.com/mohsenMahmoodzadeh/dragon

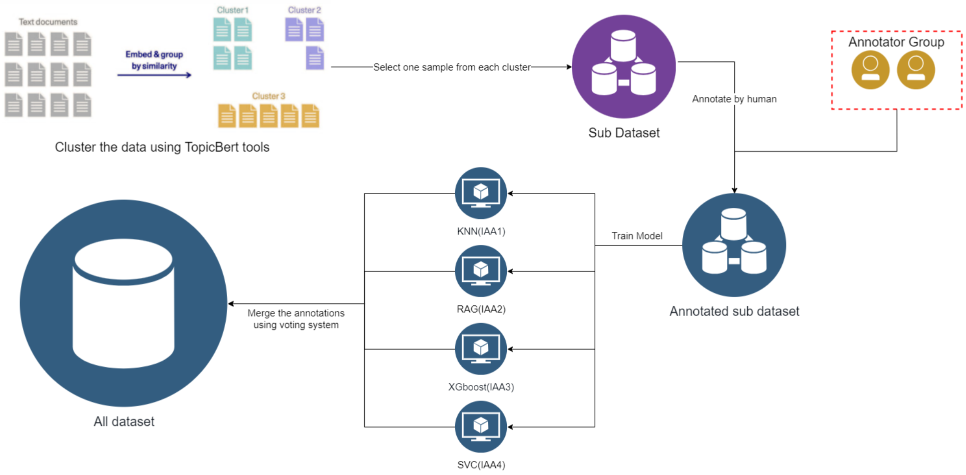

Figure 1: Overview of the annotation pipeline: First, select the subset of data to annotate. Then, annotate this subset and use it to train four different models. Finally, use these models to annotate the rest of the dataset using a voting mechanism.

is that for some tags, like "Biased against both Palestine and Israel," we have very few samples for training and testing different models. As a result, our outcomes can be unstable and unreliable in some cases. Using cross-validation techniques can provide more reliable and stable results.

Tag name	Train	Test
Propaganda	427	107
Not Propaganda	799	200
Unclear	156	40
Not Applicable	66	17
Total	**1448**	**364**

Table 4: Distribution of propaganda labels in Train and Test datasets.

2.3 Inter-Annotator Agreement (IAA) Analysis

In this section, we detail the IAA process, utilizing four different classification methods: KNN, RAG, XGBoost, and SVC. Each of these classifiers serves as an IAA tagger to assess the correlation and quality of data annotation in our process. Using a voting system, we curate these four methods to determine the final label for each new sample. Specifically, KNN serves as IAA1, RAG as IAA2, XGBoost as IAA3, and SVC as IAA4.

The `Vote` function aggregates predictions from multiple machine learning models by initializing a vote count for each possible label. It iterates through each model to predict a label for a given text input and updates the corresponding vote count in a dictionary. In cases where multiple labels receive the highest number of votes, it selects the final label based on a criterion involving `f1_scores`. Otherwise, it simply chooses the label with the highest vote count. This approach effectively leverages ensemble techniques to enhance prediction accuracy by combining insights from diverse models, making it particularly useful in scenarios where individual models may specialize in different aspects of the data.

Algorithm 1 Vote Function

```
function VOTE($models, f1\_scores, text, labels$)
    $votes \leftarrow$ {label : 0 for label in $labels$}
    for $model$ in $models$ do
        $label \leftarrow$ model.predict($text$)
        $votes[label]$ += 1
    end for

    if we have multiple maximum votes then
        $final\_label \leftarrow$ maximum_labels[$f1\_scores$].argmax()
    else
        $final\_label \leftarrow votes$.argmax()
    end if
    return $final\_label$
end function
```

3 Team Composition and Training

First, we define the Bias and Propaganda tags and share these definitions with the annotators. Our team includes two human annotators and several artificial annotators trained on a subset of data chosen using BERTopic, as explained in section 2.2. One of the artificial annotators employs the RAG technique, integrating label definitions into the LLM [5] model's prompts, which our two human annotators also use for their annotations. However, other classic classification models do not utilize these

[5] Large language model

label definitions during their training or inference processes.

3.1 Handling of Ambiguities and Difficult Cases

To handle ambiguities and difficult cases, if we find instances where the human-assigned label differs from the model-assigned label, we discuss these conflicts with our annotators. Sometimes, despite the LLM setting the correct label, the human annotator may assign a different one. In cases where conflicts arise between human results and model annotations, we revise our definitions and attempt to resolve the conflict. If the human annotation is incorrect, we correct it accordingly.

4 Task Participation and Results

The team's performance was bolstered by employing a sophisticated classification approach combining embedding and classifier models, experimenting with KNN, SVC, and XGBoost, with KNN emerging as the superior classifier. The innovative use of Retrieval RAG further enhanced this approach, involving dynamic prompts tailored for each news article. By retrieving and including examples and definitions based on the most similar annotated documents, focusing on unique labels among these documents, the team refined their prompts, leading to improved classification outcomes. Experiments showed that using 5 similar documents and including one example per unique label in the prompts yielded satisfactory results. This methodology was facilitated by the **gpt-3.5-turbo** model, which efficiently generated text and classified news articles, underscoring the advantage of integrating examples and definitions in prompt-based classification compared to related works relying on static or less dynamic prompting techniques. The results of Propaganda and Bias are shown in the tables 5 and 6.

In Table 7, the results of the FIGNEWS shared task are shown. These results were generated by the shared task committee.

4.1 KNN Method

We use the KNN algorithm with the following hyperparameters: n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, and metric='minkowski'. These settings provided the best cross-validation F1-score for both the propaganda and bias datasets.

Pre-train Model	KNN	SVC	XGBoost	RAG
Multilingual E5 large (Wang et al., 2024)	0.60	0.59	0.58	0.56
BGE M3 (Chen et al., 2024)	0.61	0.59	0.59	0.55
E5 Mistral 7B (Wang et al., 2023)	0.58	0.61	0.61	-
Nomic-Embed (Nussbaum et al., 2024)	0.55	0.57	0.57	-
XLM-Roberta (Conneau et al., 2019)	0.57	0.58	**0.63**	-

Table 5: Propaganda results with different embedding and classification methods (regarding weighted F1 score).

Pre-train Model	KNN	SVC	XGBoost	RAG
Multilingual E5 large	**0.78**	0.75	0.75	0.76
BGE M3	**0.78**	0.77	0.77	0.71
E5 Mistral 7B	0.74	0.76	0.76	-
Nomic-Embed	0.74	0.75	0.72	-
XLM-Roberta	0.75	0.74	0.72	-

Table 6: Bias results with different embedding and classification methods (in terms of weighted F1 score).

4.2 SVC Method

We employ the SVC with the following hyperparameters: C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, verbose=False, and decision_function_shape='ovr'. These settings were determined to yield the optimal performance based on cross-validation for both the propaganda and bias datasets.

4.3 XGBoost Method

We utilize XGBoost with loss = 'log_loss', learning_rate = 0.1, n_estimators = 100, subsample = 1.0, criterion = 'friedman_mse', min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_depth = 3, min_impurity_decrease = 0.0, verbose = 0, warm_start = False, validation_fraction = 0.1, tol = 0.0001, ccp_alpha = 0.0 hyperparameters. These settings have been found to produce optimal performance based on cross-validation for both the propaganda and bias datasets.

4.4 RAG Method

One potential improvement opportunity in our work is our RAG solution. Due to the emerging

Content	Kappa	Acc	Macro F1 Avg	F1 Bias*
Quality	35.7	75.5	41.0	43.2
Centrality	19.7	41.1	21.9	59.7

Table 7: In this table, metrics such as Kappa, Accuracy (Acc), Macro F1 Average, and F1 Bias score are shown for Quality (IAA Within Team) and Centrality (Main B1+B2 Across Teams). These results were generated by the shared task committee for our team.

capabilities of LLMs and state-of-the-art retrievers, RAG-based solutions have more room for generalization, which helps to reduce the side effects of biases in the data. First, we add the definition of each class to the prompt. Then, for each class, we use the retriever to find the most similar annotated sample from our subset dataset and add them to the prompt. Finally, we add the test sample and ask the LLM to determine the label of this sample. The LLM used for predicting labels from prompts is the **GPT-3.5-turbo** model, and the sentence embedder used to create the vector database for the retriever is **Multilingual E5 large** (Wang et al., 2024).

5 Discussion

The team's findings suggest that increasing the size of the manually annotated subset used to train artificial annotators could enhance performance, indicating that a larger, more comprehensive training dataset can improve the accuracy and reliability of automated classification systems. However, the team also observed that annotator bias and conflicts arising from differing cultural and personal beliefs influenced the labeling process(our annotators are from Iran and the USA), reflecting a significant challenge in creating unbiased datasets. This highlights the importance of developing methods to mitigate annotator bias and ensure consistency across diverse teams. These insights contribute to the field by emphasizing the need for larger and more diverse training datasets and robust strategies for managing annotator bias, thereby advancing the development of more accurate and fair automated annotation systems in the context of detecting propaganda and bias.

6 Conclusion

The key insights from this paper underscore the critical need to address the issue of unbalanced subdata, as the low frequency of some labels adversely affects the performance of artificial annotators. Additionally, the reliance on only 2 human annotators from different countries has introduced biases and inconsistencies in the annotations, highlighting the necessity of expanding the annotator pool to improve the diversity and reliability of the curation process. Moreover, the current lack of thorough curation, with only 2 people annotating the sub-data, points to the need for a more systematic and collaborative approach. These findings emphasize that incorporating thorough examples and a comprehensive understanding of related work is essential for enhancing the accuracy and efficacy of annotation tasks, ultimately contributing to the development of more robust artificial annotators.

References

Kylie Anglin, Arielle Boguslav, and Todd Hall. 2020. Improving the science of annotation for natural language processing. *Grantee Submission*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Costanza Conforti, Stephanie Hirmer, David Morgan, Marco Basaldella, and Yau Ben Or. 2020. Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling. *arXiv preprint arXiv:2004.12935*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Tal Perry. 2021. Lighttag: Text annotation platform. *arXiv preprint arXiv:2109.02320*.

Marek Šuppa, Daniel Skala, Daniela Jašš, Samuel Sučík, Andrej Švec, and Peter Hraška. 2024. Bryndza at climateactivism 2024: Stance, target and hate event detection via retrieval-augmented gpt-4 and llama. *arXiv preprint arXiv:2402.06549*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa R. El-Beltagy, and Muhammed Raed AbuOdeh, editors. 2024. *The FIGNEWS Shared Task on News Media Narratives*. Association for Computational Linguistics, Bangkok, Thailand.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356.

Shikun Zhang, Omid Jafari, and Parth Nagarkar. 2021. A survey on machine learning techniques for auto labeling of video, audio, and text data. *arXiv preprint arXiv:2109.03784*.