Classifying Reddit Comments on the Israel - Palestine Conflict: A Model for the Identification of Political Affiliation

Shahar Oded
Ben Gurion University of The Negev
Be'er Sheva, Israel
odedshah@post.bgu.ac.il

Amitai Kellerman
Ben Gurion University of The Negev
Be'er Sheva, Israel
amitaike@post.bgu.ac.il

Omri Haller
Ben Gurion University of The Negev
Be'er Sheva, Israel
haller@post.bgu.ac.il

Lior Broide
Ben Gurion University of The Negev
Be'er Sheva, Israel
broidel@post.bgu.ac.il

Anton Dzega
Ben Gurion University of The Negev
Be'er Sheva, Israel
dzega@post.bgu.ac.il

ABSTRACT

Classifying political affiliations in conflict-driven social media poses challenges due to emotionally charged language, evolving linguistic trends, and context-specific nuances. This study addresses the need for tools to analyze polarized discourse by developing a machine learning classifier to predict political affiliations—Pro-Israel, Pro-Palestinian, and Undefined—in standalone Reddit comments, without relying on external context like threads or user data. Motivated by the potential to advance research on misinformation, polarization, and group dynamics, we propose a novel methodology combining manual annotation, GPT-4-mini-based automated tagging, WordNet [8] semantic augmentation, and contextual embeddings from a fine-tuned DistilBERT [19] model. We evaluate feature extraction and embedding techniques using classifiers such as Logistic Regression, SVM [5], XGBoost [3], and Deep Neural Networks [12]. Results show that contextual embeddings significantly enhance accuracy and robustness of traditional machine Learning models, providing scalable tools for analyzing polarized discussions and enabling further research in computational social science.

KEYWORDS

Israel-Palestine Conflict, Political Affiliation, Social Media, Natural Language Processing (NLP), Text Classification, Machine Learning, Deep Learning, Social Pattern Analysis

ACM Reference Format:
Shahar Oded, Amitai Kellerman, Omri Haller, Lior Broide, and Anton Dzega. 2025. Classifying Reddit Comments on the Israel - Palestine Conflict: A Model for the Identification of Political Affiliation. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 INTRODUCTION

Analyzing political discourse on social media presents unique challenges, especially in contexts of ongoing conflicts like the Israel-Palestine war. Platforms like Reddit host discussions characterized by emotionally charged language, evolving linguistic patterns, and polarized narratives. Understanding these dynamics requires advanced computational tools capable of handling the nuances of standalone comments, without relying on external context such as user information or hosting threads.

This study addresses these challenges by developing a hybrid framework that integrates large language models (LLMs) with traditional machine learning techniques to predict political affiliations—Pro-Israel, Pro-Palestinian, and Undefined—within Reddit comments. Unlike prior research focused on sentiment analysis or toxicity detection, our approach targets the linguistic markers and rhetorical strategies specific to political affiliations. This task is further complicated by the evolving nature of conflict-driven discourse, which necessitates models capable of adapting to dynamic language trends and capturing subtle rhetorical cues.

To tackle this task, we propose a novel methodology that integrates multiple techniques:

- **Data Annotation:** Combining manual annotation with automated tagging using GPT-4-mini, ensuring scalability while maintaining consistency with human-labeled data.
- **Semantic Augmentation:** Leveraging WordNet to enrich linguistic diversity and mitigate class imbalance, improving the representation of minority affiliations.
- **Contextual Embeddings:** Fine-tuning DistilBERT to capture the nuanced language patterns in conflict-related text and using its embeddings for downstream classification tasks.
- **Hyper-parameters Optimization:** Thorough grid search-based optimization to maximize the potential of each classification head.

The enriched dataset is evaluated using several classifiers, including Logistic Regression, SVM, XGBoost, and Deep Neural Networks. Through rigorous empirical evaluation, we compare the impact of augmentation and contextual embeddings on classification performance. Our results demonstrate that contextual embeddings significantly improve accuracy and robustness, providing insights into the linguistic and rhetorical patterns underlying political discourse.

Beyond classification, the proposed tool has broader implications for social research. By identifying political affiliations in conflict-driven discussions, this tool opens avenues for investigating key social issues. These include the spread of false or polarized information within specific groups, tendencies toward violence or extremism, and the evolution of sentiments and toxicity in politically charged contexts. For example, researchers could analyze how misinformation strategies differ between Pro-Israel and Pro-Palestinian groups or explore whether shifts in group sentiment

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

correspond to major geopolitical events. Such insights can contribute to understanding the mechanisms of online polarization and inform interventions to foster healthier online discourse.

This paper is organized as follows: Section 2 reviews related work, highlighting the limitations of existing methods and their relevance to this study. Section 3 presents the technical background of foundational tools like WordNet and BERT. Section 4 details our methodology, including data collection, annotation, augmentation, and model training. Section 5 provides an empirical evaluation of our approach, followed by a discussion of the results and their implications. Finally, Section 6 concludes with key findings and suggestions for future research directions.

By addressing the challenges of classifying politically polarized discourse, this research contributes to computational social science and provides scalable tools for analyzing ideological patterns in conflict-driven social media discussions. With minimal effort, this pipeline can be adapted to similar classification problems in other geopolitical conflicts or polarized contexts, enabling broader applications for understanding online discourse and its societal implications. By leveraging a small fraction of manually tagged comments, this pipeline significantly reduces the cost of classifying large conflict datasets while ensuring consistent and reliable tagging across the entire dataset.

2 RELATED WORK

Our study addresses a classification task that, to the best of our knowledge, has not been directly tackled before. While prior machine learning research has explored the Israeli-Palestinian topic, no efforts appear to have focused on developing a political affiliation classifier for social media content. A classifier such as this could facilitate comparative analyses of political group dynamics within this conflict.

Although this specific problem remains unexplored, related studies offer methods and insights that inform our approach. These works provide valuable context for the tools and techniques planned in our study, which we outline below.

The study by Wei et al. [20] explores narrative origin classification of Israeli-Palestinian conflict texts, distinguishing between Israeli and Palestinian narratives based on linguistic patterns. It introduces two datasets—historical excerpts (SBS) and news articles (IP-News)—preprocessed to eliminate stylistic biases. Techniques such as synonym replacement and sliding window augmentation were employed to enhance data utility, with models like CNN and LSTM leveraging pretrained word embeddings. The CNN model achieved the highest F1 scores (85. 1% for SBS, 91. 9% for IP-News), highlighting the effectiveness of data augmentation. Although limited by the size of the dataset, its focus on English texts and reliance on formal sources, the study offers valuable insights into classifying texts in socio-political contexts. Its findings are particularly relevant to our work, as the demonstrated ability to classify origins using linguistic features parallels our goal of detecting political affiliation from social media texts. Additionally, the models and data augmentation techniques employed in the study could inform our approach, especially as we also work with a relatively small dataset on the same topic.

The study by Lee et al. [13] focuses on predicting political party affiliations by classifying U.S. politicians' tweets as Democratic or Republican using recurrent neural network architectures, particularly a bidirectional GRU model. Achieving a high accuracy of 91.6%, the study demonstrates the effectiveness of pre-processing techniques, such as tokenization and placeholder replacement for hashtags and mentions, combined with GloVe embeddings for feature representation. Evaluation metrics, including precision and recall, highlighted robust classification capabilities, while detailed error analysis revealed challenges in handling ambiguous or non-partisan tweets. However, the study is limited to politicians' tweets and single-tweet analysis, which may restrict its applicability to non-politicians or evolving linguistic trends. While the classification task differs in topic, the methods and findings in this study underscore the potential of linguistic features for predicting political affiliations in social media texts, offering valuable insights for our own classification task on a different political subject.

The study by Ramdhani et al. [18] examines sentiment on the Palestine-Israel conflict in Indonesian-language tweets, classifying them as positive, negative, or neutral using a convolutional neural network (CNN) algorithm. With a dataset of 3,632 tweets, preprocessed through case-folding, stopwords removal, stemming, and tokenization, and leveraging GloVe embeddings for feature extraction, the CNN model achieved 70% accuracy with a 90:10 train-test split. This demonstrates the feasibility of sentiment classification in this context. However, the study is limited by dataset size, the informal nature of tweets, and selection bias from hashtag-based data collection. While this study narrows its focus to sentiment classification of Indonesian tweets, its pre-processing techniques and CNN architecture offer valuable insights for our classification of political affiliations. The linguistic characteristics of conflict-driven text, as demonstrated in this study, share similarities with our domain, making its methods applicable to our goals.

The study by Miner et al. [14] examines how social media interactions evolve in response to geopolitical events, focusing on toxicity levels during the Ukraine–Russia and Hamas–Israel conflicts. Employing LDA for topic modeling and supervised models such as Linear Regression (LR) and BERT for toxicity prediction, the study identified significant increases in toxicity post-conflict. The LR model achieved a low MAE of 0.0461 for Hamas–Israel data, while LDA uncovered divisive topics like "war crime" and "human shield." Despite its effectiveness, the study's reliance on toxicity oversimplifies conflict discourse, and challenges such as imbalanced datasets and linguistic biases persist. Although this study focuses on toxicity trends rather than political affiliations, its use of topic modeling to uncover discourse patterns offers valuable insights for our work. Linguistic styles and concepts likely vary between political groups in similar conflict-driven discussions, and as our dataset evolves mid-conflict, we anticipate elevated toxicity levels that could introduce prediction challenges and require careful model tuning to mitigate biases.

The research by Pittaras et al. [17] investigates semantic augmentation to improve text classification by enriching word embeddings with hierarchical and frequency-based semantic features from WordNet. By combining semantic vectors with neural embeddings and refining them using threshold-based dimensionality filtering and hypernym propagation, the study adds contextual

depth to embeddings. Evaluated on benchmark datasets such as 20-Newsgroups and Reuters, the approach demonstrated significant accuracy gains compared to baseline methods, showcasing the benefits of integrating lexical and semantic features. However, the method is computationally intensive, depends on WordNet's quality and coverage, and does not explore multilingual or domain-specific datasets. While this study addresses a different classification task, its semantic augmentation methods could inform our approach. WordNet-based techniques, in particular, may uncover nuanced relationships and implicit connections in conflict-related Reddit discussions, enriching context-sensitive classification within our dataset.

The research by Gopi et al. [10] presents an approach using traditional machine learning to classify sentiment polarity in tweets. Several popular algorithms, including Decision Tree, Random Forest, Naive Bayes, and Support Vector Machine (SVM), were evaluated, with the SVM method ultimately selected for its superior performance. The study focused on classifying movie review tweets as positive, negative, or neutral by improving the RBF kernel by modifying the value of the gamma parameter and using a soft margin instead of a strict margin. This approach achieved an impressive accuracy of 98.8%, outperforming other RBF kernels and models across three datasets. While the study does not compare its method to deep learning approaches, which are prevalent in similar text classification tasks, its techniques can inform our research. Specifically, adapting their SVM model improvements to our classification task could provide valuable insights, and our study intends to address the gap by incorporating comparisons with deep learning methods.

3 BACKGROUND

To provide context for the tools and methods planned in this study, we examined research from two key domains: text classification tasks (with a focus on social media text) and automated data tagging. These fields offer insights into strategies, models, and techniques that directly inform our methodological framework. In this chapter, we review foundational studies that highlight approaches to data labeling, classification accuracy, and the integration of linguistic and contextual features. This background establishes the theoretical and practical basis for the tools and methodologies employed in our research.

The two works on automated annotation with LLM-generated training labels [16] and the "Fabricator" [9] both tackle the challenge of creating labeled datasets for NLP tasks, offering innovative, LLM-driven solutions. FABRICATOR streamlines dataset generation with workflows for label-conditioned data, annotating unlabeled datasets, and few-shot learning. Its modular design and integration with HuggingFace [7] libraries make it versatile, though it faces limitations in handling complex tasks and addressing biases in LLM-generated data.

The first work [16] demonstrates the effectiveness of GPT-4-generated labels as a cost-efficient alternative to human labeling, achieving a median F1 gap of only 0.039 across multiple tasks. By fine-tuning models like BERT with these surrogate labels, the study highlights the scalability of LLM-assisted labeling while emphasizing the importance of human validation to ensure reliability.

These methods are highly relevant to our study, where we aim to leverage LLMs for tagging conflict-related Reddit comments. Fabricator's workflows can streamline the initial tagging process, while the work by Pangakis and Wolken [16] provides a validated framework for combining LLM-generated tags with human supervision to enhance accuracy and mitigate biases. Together, these approaches illustrate the potential of LLMs to improve efficiency and reliability in dataset creation.

WordNet, introduced in the seminal paper "WordNet" [8], has become a cornerstone of natural language processing. It provides a semantic network that organizes English words into synonym sets (synsets) and captures structured relationships such as hypernymy (general-specific) and meronymy (part-whole). Widely applied in NLP, WordNet supports tasks like text classification, semantic similarity, and information retrieval. Its hierarchical structure also enriches ontologies and knowledge graphs, serving as a foundation for advanced semantic analysis and machine learning applications. As highlighted in the Related Work, we aim to leverage WordNet to enrich input features and explore its potential for data augmentation [17, 20].

The foundational work of the "BERT" model [11] introduces BERT (Bidirectional Encoder Representations from Transformers), a pre-trained language model that leverages a bidirectional Transformer architecture to process context from both directions. By using masked language modeling (MLM) and next sentence prediction (NSP) during pre-training, BERT excels in capturing nuanced language patterns and relationships between sentences. Fine-tuned for tasks such as text classification, question answering, and named entity recognition, BERT achieves state-of-the-art results on benchmarks like GLUE and SQuAD. Its versatility across NLP tasks makes it a powerful tool for enriching text features and understanding complex linguistic structures. In our study, BERT's contextual embeddings can enhance the analysis of conflict-related social media text, improving classification performance and capturing ideological nuances. Due to resource limitations we settled for a 'DistilBERT' [19] model, which is a significantly lighter yet very accurate version of 'BERT'.

Finally, to address the task at hand, we will compare the performance of three foundational models on the contextual embeddings of each comment: Support Vector Machines (SVM) [5], XGBoost [3], and Deep Neural Networks (DNN) [12]. SVMs, introduced by Cortes and Vapnik in 1995, are supervised learning models designed to identify the optimal hyperplane for separating data into distinct classes, effectively handling high-dimensional spaces and supporting various kernel functions. XGBoost, developed by Chen and Guestrin in 2016, is an efficient implementation of gradient boosting for decision trees, known for its scalability and outstanding performance in machine learning competitions. DNNs, inspired by the structure of the human brain, consist of multiple interconnected layers capable of learning hierarchical data representations, excelling at capturing complex patterns across diverse tasks. By leveraging our custom embeddings with these models, we aim to assess their ability to capture semantic nuances and enhance classification outcomes based on the political affiliation.

4 METHODOLOGY

This study aims to develop and evaluate a classifier capable of predicting political affiliations—Pro-Israel, Pro-Palestinian, and Undefined—in conflict-related Reddit comments. Our model seeks to capture fine linguistic nuances and sensitivities to the distinct language patterns and rhetorical strategies developed around the Israel-Palestine conflict.

Example 4.1. The following comments illustrate how the classification process works:

- A comment labeled **Pro-Israel**: *"Are you forgetting that Hamas is the one to blame for those?"*
- A comment labeled **Pro-Palestinian**: *"The IOF used an ambulance and costumes to try to rescue him, and ended up getting themselves all killed in the effort. If he had just stayed a hostage, he would have most likely still been alive today."*
- A comment labeled **Undefined**: *"Hang on. Who are the good guys?"* (lacking a clear affiliation without additional context).

The classification process is demonstrated in Figure 1

To achieve this classification, we combine manual annotation, automated tagging, semantic augmentation, bidirectional embedding, and advanced machine learning techniques. The manual annotation process is guided by specific linguistic markers for each affiliation. Pro-Palestinian comments often center on themes such as perceived mistreatment, "apartheid," and "genocide," while Pro-Israeli comments emphasize the IDF's defensive actions and characterize Hamas's behavior as terrorism. Comments that mock, discredit, or critique opposing narratives, or employ sarcasm and indirect language, are carefully evaluated for tone and intent. To enhance the accuracy of tagging, we conducted a keyword study to identify distinctive phrases commonly used by users of each political affiliation. This study leveraged TF-IDF analysis to isolate frequently occurring words and phrases unique to each group, refining the prompts used in automated tagging. By incorporating both explicit linguistic features and nuanced rhetorical cues, this methodology addresses a previously unexplored classification task, ensuring a robust and context-sensitive model.

The process begins with the integration of a dataset unrelated to the Israeli-Palestinian conflict, such as comments from Russia-Ukraine conflict [4]. These comments are tagged as "Undefined," as they are assumed to lack relevance to the targeted conflict, and are intended to diverse the data. Simultaneously, we manually annotate a subset of our primary dataset [2] into three categories: Pro-Israel, Pro-Palestine, and Undefined. This manually labeled subset provides the ground truth for evaluating and refining the performance of automated tagging models.

To scale the annotation process, we use GPT-4o-mini to tag a larger subset of comments. The model's performance is validated against the manually labeled subset, allowing for iterative refinement until satisfactory accuracy is achieved. This step significantly increases the dataset's size while maintaining consistency with the ground truth through human supervision.

After assembling the complete dataset, we performed basic text normalization, including replacing URLs and user mentions with placeholders, cleaning hashtags, and removing irrelevant characters and spaces. To enrich the dataset and address label imbalance, we applied semantic augmentation techniques inspired by WordNet. By leveraging hierarchical and frequency-based semantic relationships, this approach introduces variability and enhances the dataset's linguistic representation. In the augmentation process, we replaced adjectives, verbs, nouns, and adverbs, as these carry the primary meaning and context of a sentence, ensuring meaningful variability while preserving intent. Stop words and Other POS, such as pronouns, prepositions, and conjunctions, were excluded as they contribute minimally to semantic diversity. The augmented examples diversify the dataset, improve the representation of minority labels (Pro-Israel, Pro-Palestine), and strengthen the model's ability to generalize to unseen data.

Class	Samples
Pro-Israel	5561
Pro-Palestine	7757
Undefined	29893
Total	**43211**

Table 1: Number of samples per class in the final dataset (showing unbalanced class representation).

The enriched dataset is utilized to fine-tune a DistilBERT model, which produces high-quality contextualized embeddings for each comment. The [CLS] token, representing the entire comment, is extracted to serve as input to downstream classifiers. To ensure robust embeddings, DistilBERT is trained on a subset 70% of the data. Augmentation is applied independently to each subset after the dataset is split, ensuring no overlap or data leakage between the two.

Finally, the embeddings from the fine-tuned model are used as feature vectors in the training and testing of three machine learning models as classification heads: Support Vector Machines (SVM), XGBoost, and Deep Neural Networks (DNN). Hyper-parameters optimization is performed using Optuna [1], an optimization framework, which we used to fine-tune each model to it's best performance. The remaining 30% are used for testing and ablation studies of the best configurations. By comparing the results, we assess each model's ability to capture the semantic and contextual nuances inherent in the dataset.

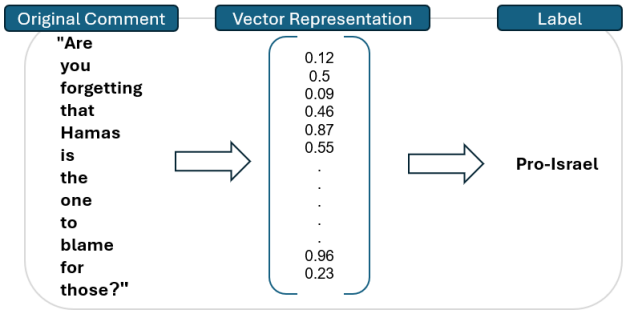

Figure 1: Illustration of the classification process. This process involves data pre processing of comment text to a representative vector, followed by classification to one of the three labels. For instance, as shown in Example 4.1, the comment "Are you forgetting that Hamas is the one to blame for those?" is preprocessed and classified as *Pro-Israel*.

Our methodology combines several established tools and techniques in a novel application to classify political affiliations in sociopolitical discourse. The use of DistilBERT for embeddings, GPT-4o-mini for scalable annotation, and semantic augmentation enriches the dataset and ensures a robust evaluation framework. This approach provides a scalable and adaptable solution for addressing the challenges in classifying politically charged social media text.

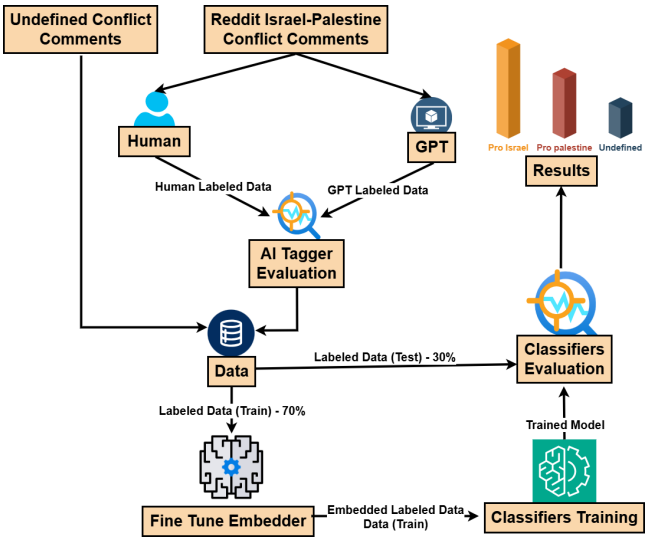

Figure 2: Illustration of the methodology, including the acquisition of tagged data from untagged original datasets, data partitions, processing with a fine-tuned embedding model, classifier training for the specific task, and the presentation of results.

5 EMPIRICAL EVALUATION

5.1 Experimental Setup

The primary goal of these experiments is to evaluate the performance of the proposed classification pipeline and validate the contribution of each step. The research questions are:

- How effective are large language models in automated tagging compared to manual annotation?
- How does semantic augmentation impact the model's performance?
- Which machine learning model achieves the highest accuracy using DistilBERT's embeddings?

Data selection begins with the acquisition of a large dataset of Israel-Palestine conflict-related comments from Reddit [2]. The original dataset contains approximately 2 million text comments, from which 1.75% were sampled due to computational and budgetary constraints, resulting in a research dataset of 38,800 comments. This dataset is further supplemented with an additional dataset from a different geopolitical conflict, Russia-Ukraine [4], which provides 5,000 randomly selected comments to serve as examples of "Undefined" political affiliations. To ensure compatibility with BERT-based embeddings, comments are truncated to a maximum length of 512 tokens, the model's context window limit. This truncation retains meaningful content while adhering to input size constraints.

The primary dataset is first divided into a manually labeled subset for validation and an unlabeled subset for automated tagging using the LLM. Following label generation, the data is split into training (70%) and testing (30%) subsets. Details of the data pipeline illustrated in Figure 2.

Fine-tuning DistilBERT is performed on Google Colab using NVIDIA Tesla T4 GPUs (16 GB VRAM), while other processes, such as semantic augmentation (WordNet) and hyper-parameters optimization (Optuna), are executed locally on an Intel Core i7 CPU (16 GB RAM, 512 GB SSD). Open-source libraries including HuggingFace, sci-kit learn, XGBoost, Torch, and Optuna are utilized. Tagging is implemented using the GPT-4o-mini API. Experiments are conducted with a fixed random state of 42 for reproducibility.

Performance is evaluated using micro-level metrics such as accuracy and F1-score, to assess overall effectiveness. Automatic tagging is compared to manual annotations, while hyper-parameters optimization, data handling method and classifier head selection are guided by the F1-score using 5-fold cross-validation. Benchmarking is performed against a baseline model (e.g., Logistic Regression).

Ablation studies are conducted to evaluate the contribution of key components to the overall methodology. Performance differences will be evaluated using the classifier's F1-score on the test set, with significance determined using a paired T-test on the binomial distribution of the correct and incorrect classifications. A p-value threshold of 5% will be used to validate the statistical significance of the results. Specifically:

- We will compare our embedding method against a standard TF-IDF representation of the comments.
- Performance will be assessed both before and after applying semantic augmentation.

- The best-performing classifier head will be compared to alternative classifiers.
- Results will be organized into a table, with columns representing different configurations (e.g., TF-IDF, TF-IDF + Augmentation, Embedding, Embedding + Augmentation) and rows for each tested model (e.g., Logistic Regression, SVM, XGBoost, DNN).

This approach provides a clear and systematic analysis of the impact of each component on the overall performance.

For models not inherently supporting multi-class classification, the one-vs-rest extension in sci-kit learn is applied. Hyper-parameters tuning via Optuna ensures optimal configurations. This experimental design provides a comprehensive evaluation of the methodology and its potential for political affiliation classification in conflict-related social media text.

5.2 Results

In this section, we present the evaluation results of our methodology, focusing on two key components: the Auto Tagger, used to scale up the data, and the classification models, used for the classification of the comment.

5.2.1 Auto-Tagger Evaluation. This subsection evaluates the performance of the automated tagging process using GPT-4o-mini. The accuracy of automated tags was compared against manually annotated labels to assess the reliability and consistency of the Auto Tagger. The evaluation metrics, including accuracy, F1-score, and the confusion matrix, are presented below. These metrics demonstrate the Auto Tagger's ability to replicate human annotations effectively.

Metric	Value
Micro Accuracy	90.48%
Micro F1 Score	90.38%

Table 2: Performance metrics for Auto Tagger evaluation.

The confusion matrix, shown in Table 5, provides detailed performance across the three classes. The model performed strongly on Pro-Israel and Pro-Palestinian labels, with minimal confusion between classes. However, some challenges were noted in differentiating Undefined comments from Pro-Israel and Pro-Palestinian comments.

Actual \ Predicted	Pro-Israel	Pro-Palestinian	Undefined
Pro-Israel	41	1	3
Pro-Palestinian	3	80	11
Undefined	2	0	69

Table 3: Confusion matrix for Auto Tagger on manually tagged test batch.

The confusion matrix highlights that the automated tagger achieves high precision for Pro-Israel and Pro-Palestinian classifications but shows a tendency to over classify comments as Undefined. This occurs particularly in cases where manually labeled politically affiliated comments lack clarity or reference highly specific events. Despite this limitation, these results provide a solid benchmark for training the classifier model. A deeper analysis of these misclassifications is presented in Section 5.3.

5.2.2 Classifier Selection and Evaluation. This subsection evaluates the performance of classification models trained using two vectorization methods: contextual embeddings (DistilBERT) and aggregated TF-IDF vectors. We compare the F1-scores of Logistic Regression, SVM, XGBoost, and Deep Neural Networks, while also examining the impact of semantic data augmentation. Performance differences between models are analyzed based on the average evaluation metrics across folds, accounting for their distribution.

Hyper-parameters optimization was conducted using Optuna with 5-fold cross-validation. The best training configurations were selected based on F1-scores, prioritizing configurations that demonstrated statistically significant improvements with minimal variance across folds. To determine the optimal augmentation ratio, F1-scores were compared across datasets augmented at ratios of X2, X3, and X4, with X3 yielding the best results. These model selection findings are presented in Table 4.

Table 4: Model Selection: Cross-validation F1 scores for TF-IDF and embedding-based models with and without data augmentation.

Model	No Augmentation	X3 Augmentation
Logistic Regression	0.52 ± 0.00	0.34 ± 0.00
SVM	0.82 ± 0.00	0.77 ± 0.01
XGBoost	0.84 ± 0.01	0.82 ± 0.02
Deep Neural Network	0.60 ± 0.00	0.43 ± 0.00

Training Results using Contextual Embedding, showing the dominance of the XGBoost model without augmentation.

Model	No Augmentation	X3 Augmentation
Logistic Regression	0.68 ± 0.00	0.37 ± 0.00
SVM	0.75 ± 0.01	0.78 ± 0.00
XGBoost	0.76 ± 0.02	0.78 ± 0.01
Deep Neural Network	0.67 ± 0.00	0.33 ± 0.00

Training Results using TF-IDF Representative Vector, showing dominance of the XGBoost and SVM models, slightly benefiting from augmentation.

Our experiments revealed that the optimal configuration is the XGBoost model utilizing contextual embedding vectors, without augmentation, but instead employing under-sampling to achieve label balance. Leveraging the best-performing configurations identified during cross-validation our model reached an Accuracy score of **83.1%** and an F1-score of **83.64%**.

The results of the ablation studies on the test set are as follows:

Table 5: Confusion matrix of the best classifier, highlighting its test set performance, with slightly higher misclassification in the Pro-Israel class.

Actual \ Predicted	Pro-Israel	Pro-Palestinian	Undefined
Pro-Israel	824	226	580
Pro-Palestinian	152	1666	564
Undefined	306	282	8272

Table 6: Test set F1-scores for TF-IDF and embedding-based vectorization approaches, with and without semantic augmentation, highlighting the superior performance of the XGBoost model with contextual embeddings over the benchmark logistic regression and TF-IDF vectors.

Model	F1 Score
TF-IDF Results	
Logistic Regression	0.589
Logistic Regression + Augmentation	0.533
XGBoost	0.676
XGBoost + Augmentation	0.676
Embedding Results	
Logistic Regression	0.828
Logistic Regression + Augmentation	0.818
XGBoost	0.836
XGBoost + Augmentation	0.825

- Hypothesis 1: An optimized XGBoost model significantly outperforms an optimized benchmark Logistic Regression model when using the same feature set ($p\text{-}value = 0.009$).
- Hypothesis 2: Semantic augmentation enhances classification performance. This hypothesis was dis-proven in the context of this problem.
- Hypothesis 3: Contextual embeddings significantly outperform TF-IDF feature vectors as input for the XGBoost classifier in this task ($p\text{-}value = 0.00$).

A detailed analysis is provided in Section 5.3.

5.3 Discussion

In this section, we analyze the performance of the Auto Tagger and the classification models, addressing their strengths, limitations, and broader implications. We also discuss the observed patterns and challenges in misclassifications, as well as the practical implications of our results.

5.3.1 Auto Tagger Analysis. The Auto Tagger demonstrated a cautious approach in assigning politically affiliated tags (Pro-Israel / Pro-Palestinian). While it effectively recognized nuances, sarcasm, and criticism within comments, it occasionally failed to leverage polarized keywords, such as "IOF" (Israel Offense Force), "Setanyahu" (derogatory for Israel's current prime minister), or "Pallywood" (dismissing Palestinian testimonies as fake propaganda). Despite being informed of these terms during training, the model sometimes overlooked their significance, which could have strengthened its ability to detect political affiliation.

Furthermore, the Auto Tagger struggled with comments referring to specific events and their criticisms, often misclassified them as Undefined due to uncertainty over which side was being criticized. This limitation aligns with the model's tendency to prioritize precision over recall, particularly when assigning politically affiliated labels. However, this trade-off was acceptable for our purposes, as reducing false positives is critical in such a sensitive task. Overall, the Auto Tagger performed remarkably well given the complexity of the task, providing a robust benchmark for the downstream classifier training.

5.3.2 Classifier Analysis. The classification heads were analyzed for their strengths and limitations. Replacing the transformer's single classification layer with traditional ML models (e.g., XGBoost, SVM) consistently outperformed neural network (NN)-based heads. This improvement is likely due to the ability of classic ML models to effectively leverage dense representations, creating robust decision boundaries while mitigating over-fitting on smaller datasets. In contrast, NN-based heads typically require extensive parameter tuning and larger datasets, which were beyond the scope of our experimental setup.

Semantic augmentation, while theoretically sound, did not improve classification performance in most cases. Closer inspection revealed that some augmented texts introduced significant grammatical errors, introducing noise and biases that hindered model effectiveness. Refining the augmentation process or exploring alternative techniques could address these issues and improve results in future work.

Our proposed method outperformed the benchmark across both feature extraction and classification techniques, validating its robustness. However, classifiers trained using this methodology inherit biases from the auto-tagger benchmark, which may influence predictions. Interestingly, misclassifications analysis revealed that some comments labeled as 'Undefined' by the benchmark were correctly classified as politically biased by our model, suggesting it may outperform the benchmark in certain cases. Despite these strengths, the model struggled with highly contextualized terms (e.g., 'genocide,' typically linked to Pro-Palestinian affiliation but appearing in Pro-Israeli contexts) and performed worse on Pro-Israel comments overall, achieving only 0.64 precision for this group, highlighting challenges in handling highly nuanced language.

Our evaluation used F1-score and accuracy, which measure overall performance but may overlook the importance of prioritizing politically affiliated comments over Undefined ones. Future work could explore weighted F1-scores or class-specific precision to better address critical classifications in sensitive or imbalanced datasets.

In summary, the combination of XGBoost classifier with contextual embeddings demonstrated clear superiority over benchmark models. While the approach struggled with contextually complex comments, it proved robust and adaptable, aligning with findings from similar studies (§2). These results validate the potential of our pipeline for text classification tasks and highlight areas for further refinement.

6 CONCLUSIONS AND FUTURE WORK

This study set out to address the challenges of classifying politically polarized discourse in conflict-driven social media discussions. By leveraging a combination of manual annotation, automated tagging, semantic augmentation, contextual embeddings and machine learning algorithms, we developed a scalable pipeline capable of predicting political affiliations—Pro-Israel, Pro-Palestinian, and Undefined—in standalone Reddit comments. The results highlight the effectiveness of embedding-based vectorization in capturing the nuanced and polarized language characteristic of this domain.

One key takeaway is the strength of the Auto Tagger in providing a relatively reliable benchmark for classification, prioritizing precision over recall to minimize false positives in politically sensitive tasks. Despite its limitations and occasional errors, the benchmark proved sufficient as a baseline for training several effective classification models, particularly SVM and XGBoost. These models demonstrated robust performance, further highlighting the capabilities of modern large language models (LLMs) in generating rich contextual representations of text, especially within the nuanced domain of social media discourse.

Despite the promising results, this study faced several limitations:

- **Misclassifications by the Auto Tagger:** The Auto Tagger occasionally struggled with comments containing highly polarized keywords or references to specific events, resulting in misclassifications. Furthermore, labeling political comments is inherently subjective, and some classifications may not achieve unanimous agreement among annotators.
- **Evolving Nature of the Conflict:** The Israeli-Palestinian conflict, like other ongoing conflicts (e.g., the Russia-Ukraine war, Syrian civil war, Sudanese civil war), evolves over time. Consequently, new events and emerging linguistic patterns may fall outside the model's recognition capabilities, reducing its relevance and effectiveness as the conflict progresses. However, the sourced dataset for this study is daily updated, and the research pipeline can be used to integrate these updates into the model. With minimal effort in updating the dataset and performing additional manual tagging to recalibrate the Auto Tagger, the methodology remains adaptable and capable of reflecting the conflict's progression over time.
- **Platform-Specific Dataset:** The dataset used in this study is sourced exclusively from Reddit due to its availability and the platform's rich repository of political discourse. Different social media platforms (e.g., TikTok, Instagram, Facebook, Twitter) may exhibit distinct linguistic trends, cultural dynamics, and rhetorical styles, which could yield varying classification outcomes.

These limitations underscore the need for continued refinement and expansion of the study's approach.

Future work could address these limitations through:

- **Integrating contextual information**: Including thread-level discussions, user metadata, or news sources to provide richer context and improve classification accuracy. These external sources could offer additional context to better disambiguate comments that reference specific events or polarized rhetoric.
- **Expanding the dataset**: Incorporating comments from various geopolitical conflicts to validate the pipeline's generalization across sociopolitical contexts.
- **Exploring advanced embeddings**: Testing alternative methods like Sentence-BERT or GPT-derived embeddings to better capture nuanced and polarized language.
- **Refining augmentation techniques**: Enhancing semantic augmentation to account for linguistic subtleties and improve model's performance, as the current augmentation technique was not sufficient.

Another promising area for future exploration lies in the analysis of social patterns. Using the classifier as a tool for examining patterns of misinformation, group-specific tendencies toward toxicity, and sentiment dynamics in conflict-driven discussions offers valuable insights into the ideological and social behaviors underlying such discourse. A previous analysis [15] provides three illustrative examples that demonstrate the application of these computational methods:

- Figure 3 presents a comparative analysis of sentiment distributions across various subtopics relevant to the conflict. Sentiment scores were calculated using a pre-trained BERT-based polarity sentiment model available from Hugging Face [7]. The visualization highlights how sentiment varies between different discussion themes, with higher values indicating a more positive sentiment on average for the group members when discussing specific topics. Notably, the Pro-Palestinian group exhibits a consistently more positive sentiment across most subtopics compared to the Pro-Israeli group.
- Figure 4 explores the comparative toxicity distribution across subtopics related to the conflict. Using a BERT-based toxicity detection model from Hugging Face [7], the analysis calculates average toxicity probabilities for each theme. Higher values signify a greater tendency toward toxic speech within comments on the respective topics. The results show that the Pro-Palestinian group, on average, leans more toward a toxic speech style, with the gap being particularly pronounced in discussions about the state of Israel and the conflict itself.
- Figure 5 examines the speech styles of the two groups, focusing on their tendency toward factual versus emotional expressions when discussing the topic of hostages. A Word2Vec-based method was used to analyze linguistic patterns and infer tendencies. The findings reveal that the Pro-Israeli group tends to adopt a more factual speech style on this topic, while the Pro-Palestinian group demonstrates a more emotional approach.

These visualizations are thoroughly explained in the referenced project and serve as examples of how computational methods can uncover nuanced social dynamics in polarized discourse.

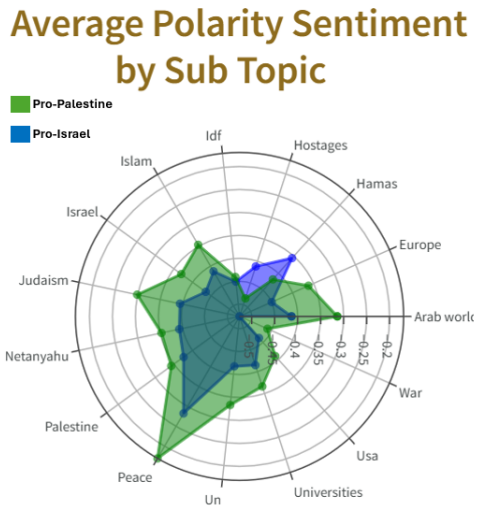

Figure 3: A comparative visualization of each group's sentiment distribution across subtopics relevant to the conflict, calculated using a BERT-based polarity sentiment model. The avg score for a sub-topic can range between [-1,1] with 1 being the most positive sentiment. The Pro-Palestinian group consistently shows a more positive sentiment across most subtopics.

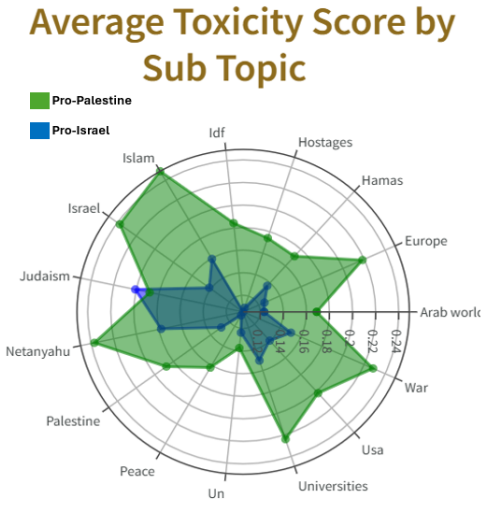

Figure 4: A comparative visualization of each group's tendency toward toxic speech across subtopics related to the conflict, calculated using a BERT-based toxicity detection model. The avg score for a sub-topic can range between [0,1] with 1 representing a high level of toxicity. The Pro-Palestinian group exhibits a higher tendency toward toxic speech, especially in discussions about the state of Israel and the war itself.

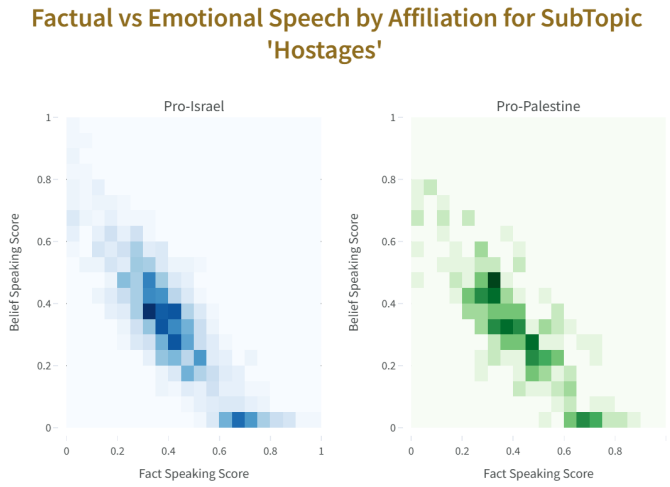

Figure 5: A comparative visualization of each group's speech style (factual vs. emotional) when discussing the hostages in this conflict's comments. The plot demonstrates a heat map for the frequency of factual and emotional speech resemblance, which can range between [0,1] where 1 is the highest resemblance. The Pro-Israeli affiliated users demonstrate a more factual speech style, while the Pro-Palestinian group tends toward more emotional expressions.

In conclusion, this study demonstrates the potential of machine learning models in analyzing ideological patterns within conflict-driven discussions. By tackling the complexities of politically polarized discourse, this research lays the groundwork for future investigations into related phenomena, such as the spread of misinformation, evolving sentiments, and toxicity trends within specific ideological groups. These insights could serve as valuable tools for researchers and policymakers aiming to understand and mitigate the impact of online polarization.

CODE AVAILABILITY

The code supporting this study is publicly available on GitHub in two repositories:

- **Israel-Palestine Political Affiliation Text Classification:** https://github.com/shaharoded/Israel-Palestine-Political-Affiliation-Text-Classification [6], which contains the complete classification pipeline, including automated tagging methods, contextual embedding generation, and scripts for model training, optimization, and evaluation.
- **Israel-Palestine War Reddit Analysis:** https://github.com/shaharoded/Israel-Palestine-War-Reddit-Analysis [15], featuring tools for language-based feature extraction, sentiment and toxicity analysis, and visualization generation, including Figures 3, 4, and 5.

These repositories include datasets, preprocessing scripts, and modular tools designed for easy replication and adaptation. They serve as resources for researchers looking to extend this methodology to new datasets, explore ideological discourse, or refine classification and visualization techniques.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[2] Asaniczka. 2024. Daily Public Opinion on Israel-Palestine War. https://doi.org/10.34740/KAGGLE/DSV/10319079

[3] 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*. 785–794.

[4] Kaggle Community. n.d.. Ukraine-Russia War Reddit Data. https://www.kaggle.com/datasets/diyacharya/ukraine-russia-war-reddit-data. Accessed: December, 2024.

[5] Corinna Cortes. 1995. Support-Vector Networks. *Machine Learning* (1995).

[6] Shahar Oded et al. 2025. Israel-Palestine Political Affiliation Text Classification. https://github.com/shaharoded/Israel-Palestine-Political-Affiliation-Text-Classification. GitHub repository.

[7] Hugging Face. 2024. Hugging Face: State-of-the-Art Machine Learning Models. https://huggingface.co/. Accessed: December, 2024.

[8] Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*. Springer, 231–243.

[9] Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch, and Alan Akbik. 2023. Fabricator: An Open Source Toolkit for Generating Labeled Training Data with Teacher LLMs. *arXiv preprint arXiv:2309.09582* (2023).

[10] Arepalli Peda Gopi, R Naga Sravana Jyothi, V Lakshman Narayana, and K Satya Sandeep. 2023. Classification of tweets data based on polarity using improved RBF kernel of SVM. *International Journal of Information Technology* 15, 2 (2023), 965–980.

[11] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. Minneapolis, Minnesota, 2.

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[13] Catherine Lee, Jacob Shiff, and Sridatta Thatipamala. 2018. Predicting US Political Party Affiliation on Twitter. *Training* 60 (2018), 7.

[14] Jordan Miner and John E Ortega. 2024. NLP Case Study on Predicting the Before and After of the Ukraine-Russia and Hamas-Israel Conflicts. *arXiv preprint arXiv:2410.06427* (2024).

[15] Shahar Oded and Nir Rahav. 2024. Israel-Palestine War Reddit Analysis. https://github.com/shaharoded/Israel-Palestine-War-Reddit-Analysis. GitHub repository.

[16] Nicholas Pangakis and Samuel Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. *arXiv preprint arXiv:2406.17633* (2024).

[17] Nikiforos Pittaras, George Giannakopoulos, Georgios Papadakis, and Vangelis Karkaletsis. 2021. Text classification with semantically enriched word embeddings. *Natural Language Engineering* 27, 4 (2021), 391–425.

[18] Muhammad Ali Ramdhani, Dian Sa'adillah Maylawati, Undang Syaripudin, Eva Nurlatifah, Rifqi Syamsul Fuadi, et al. 2023. Sentiment Analysis on the Issue of the Palestine-Israel Conflict on Twitter Using the Convolutional Neural Network Algorithm. In *2023 9th International Conference on Wireless and Telematics (ICWT)*. IEEE, 1–6.

[19] V Sanh. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[20] Jason Wei and Eugene Santos Jr. 2020. Narrative Origin Classification of Israeli-Palestinian Conflict Texts. In *The Thirty-Third International FLAIRS Conference*.