2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)
02-04 May 2024, Military Institute of Science and Technology (MIST), Dhaka-1216, Bangladesh

2024 6th International Conference on Electrical Engineering and Information &amp; Communication Technology (ICEEICT) | 979-8-3503-8577-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICEEICT62016.2024.10534429

Analyzing Sentiment and Unveiling Geopolitical Perspectives: A Comprehensive Study of Reddit Comments on the Contemporary Israel-Palestine Conflict

Kazi Nubila Nushin
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
kazinubilanushintajin@gmail.com

Md. Shahid Uz Zaman
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
szaman22.ruet@gmail.com

Mohiuddin Ahmed
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
mohiuddin.nirob.mn@gmail.com

***Abstract*—The ongoing Israel-Palestine conflict has triggered intense discussions on various social media platforms, reflecting the diverse perspectives and sentiments of users worldwide. In this study, we present a comprehensive analysis of Reddit comments related to the conflict. The aim is to understand the prevailing sentiments and geopolitical stances expressed by users in this dynamic and sensitive context. Our research comprised three key phases. In the initial stage, we conducted data labeling on a dataset containing 436,725 Reddit comments using the VADER tool, labeling them as Positive, Negative, and Neutral sentiments. Subsequently, our work progressed to the second stage, wherein Latent Dirichlet Allocation (LDA) was employed for topic modeling, shedding light on the geopolitical perspectives expressed by users. Transitioning to the third stage, we utilized Logistic Regression (LR), Stochastic Gradient Descent (SGD), Support Vector Classifier (SVC), and Naïve Bayes (NB) as learning algorithms to classify sentiments, with a focus on Accuracy, F1 Score, Precision, and Recall for performance evaluation. Notably, our experimental findings highlighted that the Logistic Regression model demonstrated the highest accuracy, achieving 96.96%.**

***Index Terms*—Israel-Palestine, Reddit comments, sentiment, VADER, LDA, Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, Naïve Bayes**

I. Introduction

Ongoing hostilities between Israel and Hamas-affiliated Palestinian militants, mainly concentrated in the Gaza Strip since October 7, 2023, have also resulted in clashes in the West Bank and along the border of Israel and Lebanon. Diverse perspectives exist within Israeli and Palestinian communities. Some view Israeli actions against Palestinians as a form of genocide, while others argue that Palestinians employ global terrorism to advance their interests [1]. Social media has introduced new viewpoints but has simultaneously intensified polarization and the formation of echo chambers.

In the last decade, the significant rise in social media engagement has empowered millions to share their perspectives on diverse subjects [2]. This surge has led to a substantial expansion of data accessible for analysis on social media platforms like Reddit, offering valuable insights into the thoughts and sentiments of individuals. The ability to discern public opinion on political matters is crucial for informing decisions related to international policies, alliances, and positions. Reddit is a well-established social media platform, reported to have 70 million daily active users [3]. According to information, about 36% of Reddit users are aged 18 to 29, while another 22% fall in the age group of 30 to 49. Reports also indicate that only 10% of Reddit users are older than 50, with just 4% of teenagers reporting usage of the platform [3].

Statistical studies have indicated that Reddit functions as a valuable reservoir of data for examining global events, including political matters. Our dataset, titled 'Daily Public Opinion on Israel-Palestine War' [4], is derived from comments on Reddit posts. This dataset encompasses comments from Reddit posts related to the current situation in Israel and Gaza. It is regularly updated, and for this paper, we analyzed discussions up to November 16, 2023, offering insights into this ongoing conflict.

Sentiment analysis, a machine learning tool, assesses the polarity of texts, ranging from positive to negative sentiments. Through training on examples of emotional expressions in text, machines acquire the ability to detect sentiment autonomously, without direct human intervention. Machine learning enables computers to learn tasks without explicit programming, allow-

979-8-3503-8577-9/24/$31.00 ©2024 IEEE

ing sentiment analysis models to comprehend nuances such as context, sarcasm, and the use of words in unintended ways [5]. This paper aims to introduce a classification system for Reddit comments utilizing a rule-based data labeling tool explicitly designed for social media text, namely VADER (Valence Aware Dictionary and Sentiment Reasoner) [6]. The process involves extracting features through a set of preprocessing components, with sentiments categorized into three distinct labels: Positive, Negative, and Neutral. We used LDA (Latent Dirichlet Allocation), a frequently employed probabilistic generative model in natural language processing [7], for the purpose of topic modeling to understand users' geopolitical stance.

Then, we employed three machine learning techniques and an optimization technique to classify sentiment: Naïve Bayes (Multinomial), Logistic Regression, Support Vector Classifier, and Stochastic Gradient Descent. The experimental results indicated that Logistic Regression exhibited the best performance.

Here, our core contributions can be summarized as follows:

- We used VADER, a rule-based sentiment labeling tool, to label the sentiments in our dataset.
- We applied LDA model for graphical representation of four kinds of user stance— (i) Neutral, (ii) Against both Israel-Palestine, (iii) Supporting Israel and (iv) Supporting Palestine.
- We applied multiple machine learning models and compared results of the models for classifying sentiments.

Subsequent sections of this paper include various components, starting with a review of related works in sentiment analysis covered in Section II. Section III details the methodology employed in our study, followed by the experimental setup and implementation outlined in Section IV. Section V presents the experimental outcomes and discussions. Limitations are covered in Section VI. Section VII presents final conclusions and outlines avenues for future research.

II. Related Works

Given the substantial user presence on social media platforms, incorporating public opinions into decision-making has become crucial. These opinions offer valuable insights into how individuals respond to specific topics. Numerous studies on sentiment analysis of textual data have been conducted in recent years.

Charitha et al. [8], delved into sentiment analysis related to public opinion mining. The researchers employed feature extraction methods, namely Bag of Words (BOW) and TF-IDF, on a dataset comprising 50,000 reviews from the IMDB dataset. The machine learning classifiers utilized in the experiment included Logistic Regression, Support Vector Machines (SVM), Multinomial Naïve Bayes, Decision Tree, K-Nearest Neighbors (KNN), and XGBoost. The results of their study revealed that the highest accuracy, reaching 90%, was achieved by employing the BOW model separately with both Logistic Regression and SVM.

Gangwar et al. [9], gathered English tweets containing hashtags such as IsraelUnderAttack, IStandWithIsrael, WeStandWithIsrael, and IsraelPalestineconflict. The researchers utilized a Bag of Words (BOW) model for vectorization. Subsequently, they employed three distinct machine learning techniques—Support Vector Classifier (SVC), Decision Tree, and Naïve Bayes—to identify sentiment classes. Their experimental findings indicated that SVC achieved an accuracy rate of 89%, Decision Tree exhibited an accuracy of 92%, and Naïve Bayes yielded the highest accuracy at 93%.

Bengesi et al. [10], performed sentiment analysis using machine learning on a dataset comprising 500,000 multilingual tweets discussing the Monkeypox outbreak shared on the Twitter platform. They utilized VADER and TextBlob to categorize tweets based on their sentiments, distinguishing between positive, negative, and neutral tones. Following that, a range of machine learning algorithms such as Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Logistic Regression (LR) were applied for sentiment classification. Their investigation revealed that the most effective model incorporated TextBlob annotation, lemmatization, CountVectorizer, and Support Vector Machine (SVM), achieving an impressive accuracy of 93%.

Sabba et al. [11], utilized a Convolutional Neural Network (CNN) to analyze 50,000 reviews from the IMDB dataset. Regrettably, the paper did not address the overfitting concern. The experimental outcomes indicated a training accuracy of 99% and a testing accuracy of 89%.

Al-Agha et al. [12], introduced an innovative multilevel data analysis model, incorporating dual levels of examination: country-level and individual-level analyses, utilizing a dataset comprising 178,524 manually collected tweets consisting of tweets about Israel and Palestine. The implementation involved the application of the Logistic Regression model by LongPipe, and a 10-fold cross-validation. Their classifier exhibited an accuracy rate of 80.63%.

Arora et al. [13], employed the Word2Vec model, specifically the Continuous Bag-Of-Words and Skip-Gram approaches, to transform text into vectors within the IMDB movie review dataset. For language understanding, they adopted the BERT (Bidirectional Encoder Transformers) method. Their model included extra layers, specifically, a Dense layer with a tanh activation function, two Dropout layers with a dropout rate of 0.5 each, and an output layer that employed the softmax activation function. The chosen loss function was cross-entropy loss. Their experimental results demonstrated an accuracy of 92.40%.

III. Methodology

In our study, we initiated an experimental framework designed to unveil the latent sentiment within the "Daily Public Opinion on the Israel-Palestine War" dataset. The global framework of our sentiment classification system is illustrated in Fig. 1, containing three principal modules.

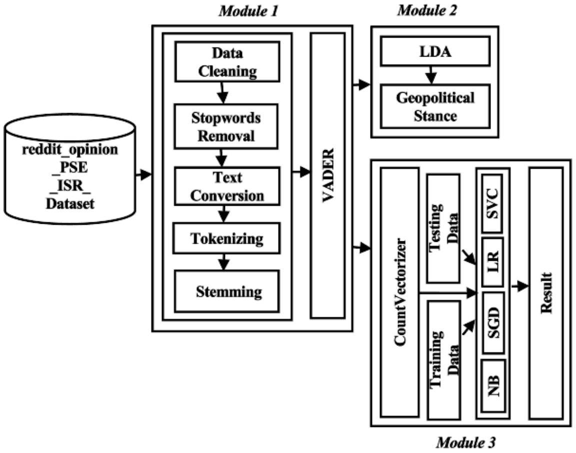
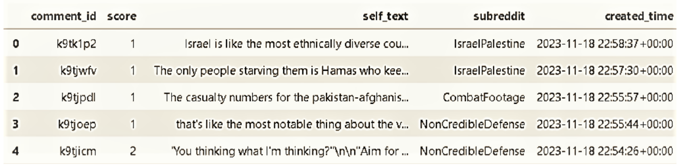


Fig. 1. Global framework of the sentiment classification system.

The initial module involves diverse preprocessing techniques in natural language processing, including the removal of punctuation marks, hashtags, user tags, stopwords, and numbers. Additionally, text conversion to lowercase and stemming procedures were applied. Sentiment scores for the dataset were computed using VADER. The second module entails the application of the Latent Dirichlet Allocation (LDA) model to discern the geopolitical stances expressed by users. Finally, the third module revolves around text vectorization using CountVectorizer and the construction of classification models using machine learning methods such as Naïve Bayes, Support Vector Classifier, Logistic Regression, and an optimization method, Stochastic Gradient Descent.

A. Dataset

Our dataset is 'Daily Public Opinion on the Israel-Palestine War,' derived from comments on Reddit posts. This dataset was collected from Kaggle and is regularly updated [4]. In this paper, we analyzed discussions up to November 16, 2023, providing insights into the ongoing situation in Israel and Gaza. Our dataset consists of 5 unique columns, namely: comment_id, score, self_text, subreddit, and created_time, each having 436725 entries. We found the dataset unlabeled in the case of the sentiment category, showed in Fig 2.



	comment_id	score	self_text	subreddit	created_time
0	k9tk1p2	1	Israel is like the most ethnically diverse cou...	IsraelPalestine	2023-11-18 22:58:37+00:00
1	k9tjwfv	1	The only people starving them is Hamas who kee...	IsraelPalestine	2023-11-18 22:57:30+00:00
2	k9tjpdl	1	The casualty numbers for the pakistan-afghanis...	CombatFootage	2023-11-18 22:55:57+00:00
3	k9tjoep	1	that's like the most notable thing about the v...	NonCredibleDefense	2023-11-18 22:55:44+00:00
4	k9tjicm	2	'You thinking what I'm thinking?"\n\n"Aim for ...	NonCredibleDefense	2023-11-18 22:54:26+00:00

Fig. 2. Unlabeled raw dataset.

B. Module 1

1) Data Preprocessing:

We removed punctuation, numbers, special characters, URLs, and other undesired elements for text cleaning purposes. Regular expressions were also utilized to define patterns for replacement, contributing to reduced memory consumption and enhanced efficiency in the learning process.

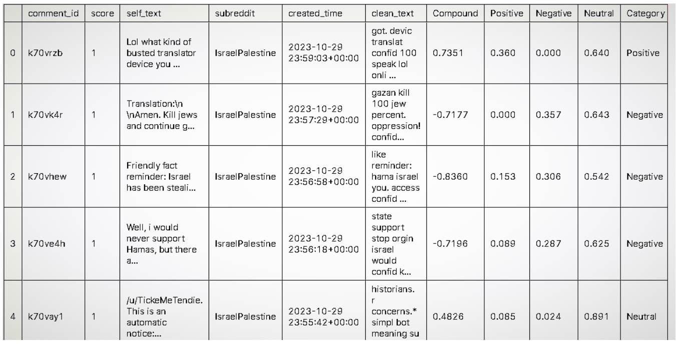
We removed stopwords (e.g., 'and', 'the', 'is') from the text because they don't carry much meaning and can be considered noise. We converted the text of the comments to lowercase. This was done to ensure that the comparison with stopwords remained case-insensitive. We performed tokenization to dissect a piece of text into smaller units called tokens. These tokens typically represent words, although they can also be subwords or characters, depending on the specific tokenization approach. Words can look different but mean the same in text because there's no set way of writing them. To address this, we used stemming, which shortens words to their core, ensuring that we capture their essential meaning.

2) Text Labeling Using VADER:

VADER, which stands for Valence Aware Dictionary and sEntiment Reasoner, is a sentiment analysis tool that has been pre-trained to specifically analyze text data for sentiment. It is particularly useful for social media text and short sentences. It calculates four sentiment scores: positive, neutral, negative, and compound. The compound score is a combination of the three individual scores and is often used to determine the overall sentiment of the text. The labeling process in VADER, as outlined by Hutto et al. [6] can be summarized by Equation 1:

$$\text{Label} = \begin{cases} \text{Positive} & \text{if score} \geq 0.5 \\ \text{Negative} & \text{if score} \leq -0.5 \\ \text{Neutral} & \text{if } -0.5 < \text{score} < 0.5 \end{cases} \quad (1)$$

We used VADER for data labeling shown in Fig 3, then conducted sentiment analysis. Comments were categorized based on compound scores: $\geq$ 0.5 as 'Positive,' $\leq$ -0.5 as 'Negative,' and the rest as 'Neutral.' Our dataset showed an imbalance, with 'Neutral' being the largest class, potentially biasing performance. To address this, we applied stratified sampling during cross-validation.



	comment_id	score	self_text	subreddit	created_time	clean_text	Compound	Positive	Negative	Neutral	Category
0	k70vrzb	1	Lol what kind of busted translator device you ...	IsraelPalestine	2023-10-29 23:59:03+00:00	got. devic translat confid 100 speak lol onli ...	0.7351	0.360	0.000	0.640	Positive
1	k70vk4r	1	Translation:\n \nAmen. Kill jews and continue g...	IsraelPalestine	2023-10-29 23:57:29+00:00	gazan kill 100 jew percent. oppression! confid...	-0.7177	0.000	0.357	0.643	Negative
2	k70vhew	1	Friendly fact reminder: Israel has been steali...	IsraelPalestine	2023-10-29 23:56:58+00:00	like reminder: hama israel you. access confid ...	-0.8360	0.153	0.306	0.542	Negative
3	k70ve4h	1	Well, i would never support Hamas, but there a...	IsraelPalestine	2023-10-29 23:56:18+00:00	state support stop orgin israel would confid k...	-0.7196	0.089	0.287	0.625	Negative
4	k70vay1	1	/u/TickeMeTendie. This is an automatic notice:...	IsraelPalestine	2023-10-29 23:55:42+00:00	historians. r concerns.* simpl bot meaning su	0.4826	0.085	0.024	0.891	Neutral

Fig. 3. Data labeling using VADER.

C. Module 2

Latent Dirichlet Allocation (LDA) is a statistical model that is employed for topic modeling and operates based on probabilistic generative principles. In Module 2, the geopolitical stance is detected using LDA. LDA posits that every document is composed of multiple topics in varying proportions, with each topic being characterized by a distribution of words. By analyzing the word patterns across documents, LDA can uncover these latent topics. The section below summarizes the topic modeling in LDA according to Blei et al. [14]. LDA assumes the following generative process for each document w in a corpus D:

1) Choose $N \sim \text{Poisson}(\xi)$.
2) Choose $\theta \sim \text{Dir}(\alpha)$.
3) For each of the N words w_n:
    a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
    b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

D. Module 3

1) Vectorization:

We used CountVectorizer [15] to convert text data into a bag-of-words [16] representation for machine learning. We applied the fit-transform operation to the training data, which served the dual purpose of fitting the vectorizer and converting the text into a bag-of-words representation. Subsequently, we utilized the transform method on the test data to convert it according to the vocabulary acquired from the training data.

2) Learning Algorithms:

In this study, we implemented and assessed three machine learning algorithms and an optimization algorithm. The subsequent discussion provides an analysis of the implemented algorithms.

- Naïve Bayes (NB):
  Naïve Bayes is a probabilistic method used to build data classification models. It involves assessing the probability, or 'likelihood,' of data belonging to a particular class [17]. We used the multinomial Naïve Bayes model to classify data. The classification is performed by estimating the probability $p(W_i \mid C)$ *i = 1..n*, that the occurrences of terms in set *S* are observed in documents *D* belonging to class *C*. The multinomial model is based entirely on the assessment of a decision function derived from the Bayes theorem and centered on probabilities. The Bayesian probability $p(C_k|W)$ is computed as follows:
  $$p(C_k|W) = \frac{p(C_k) \cdot p(W|C_k)}{p(W)} \tag{2}$$
- Stochastic Gradient Descent (SGD):
  Stochastic Gradient Descent proves to be a highly effective method for optimizing linear classifiers and regressors when dealing with convex loss functions, such as those found in Support Vector Machines and Logistic Regression [18]. We described here the mathematical details of the SGD procedure using equation 3, Given a set of training examples $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in \mathbf{R}^m$ and $y_i \in \mathcal{R} y_i \in -1, 1$ for classification. Our goal is to learn a linear scoring function $f(x) = w^T x + b$ with model parameters $w \in \mathbf{R}^m$ and intercept $b \in \mathbf{R}$. In order to make predictions for binary classification, we simply look at the sign of $f(x)$. To find the model parameters, we minimize the regularized training error given by,
  $$E(w, b) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w) \tag{3}$$
  where L is a loss function that measures model (m is)fit and *R* is a regularization term that penalizes model complexity; $\alpha > 0$ is a non-negative hyperparameter that controls the regularization strength.
- Logistic Regression (LR):
  Logistic regression is a supervised machine learning algorithm that accomplishes binary classification by predicting the probability of an outcome, or event. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false. For multiclass classification with $y_i \in 1, 2, \ldots, K$, we can extend the logistic regression to the softmax regression. The labels for K different classes can be other real values, but for simplicity they can always be converted or relabeled to values from 1 to K. Softmax regression is also called multinomial logistic regression. The softmax regression model for probability $P(y = k|\tilde{x})$ for $k = 1, 2, \ldots, K$ takes the following form:
  $$P(\tilde{x}) = \frac{e^{w^T \tilde{x}}}{\sum_{j=1}^{K} e^{w^T \tilde{x}}} \tag{4}$$
  where $\mathbf{w} = \begin{bmatrix} w_0 & w_1 & \ldots & w_m \end{bmatrix}^T$ are the model parameters, and w_0 is the bias. The m independent variables or attributes are written as a vector $\mathbf{x}^{\sim} = \begin{bmatrix} 1 & x_1 & \ldots & x_m \end{bmatrix}^T$. The denominator $\sum_{i=1}^{K} \exp(w^T \tilde{x})$ normalizes the probabilities over all classes ensuring the sum of the probabilities to be 1 [19].
- Support Vector Classifier (SVC):
  Support Vector Classifier segregates the data by determining the optimal decision boundary known as the hyperplane. Support vectors, which are the extreme points, are selected to assist in constructing this hyperplane. The positive hyperplane intersects one or more of the nearest positive attributes, while the negative hyperplane intersects one or more of the nearest negative points. The optimal hyperplane is determined by maximizing the margin, which represents the distance between the positive and negative hyperplanes. Equation 6 gives the mathematical expression of SVC model,
  $$Y_i(\vec{w} \cdot \vec{X}_i + b) = 1 - C_i \tag{5}$$

The larger the C is, the more mislabeled sample are allowed. It increases the robustness of the model, but it will decrease the precision [20].

IV. EXPERIMENTAL SETUP AND IMPLEMENTATION

In our sentiment analysis, training was conducted using 80% of the dataset, while the remaining 20% was reserved for testing. We employed the SGD classifier with a hinge loss, L2 penalty, and a fixed random state for reproducibility. This configuration, chosen for scalability and to prevent overfitting, aligned with our experiment's goals, offering an efficient solution for sentiment analysis. Performance evaluation metrics, including accuracy, precision, recall, and F1 score, were employed to measure the effectiveness of each model. 5-fold cross-validation was applied to each model.

V. EXPERIMENTAL RESULTS

This section outlines the findings obtained from our three modules.

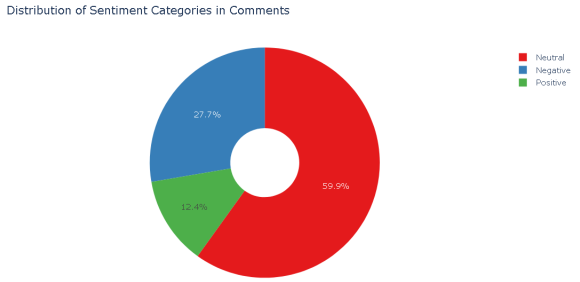


Fig. 4. Distribution of Sentiment Categories.

Fig. 4 demonstrates the distribution of sentiments using VADER tool. As the majority of the comments had a compound score between -0.5 and 0.5, almost 60% of the data was labeled as "Neutral".

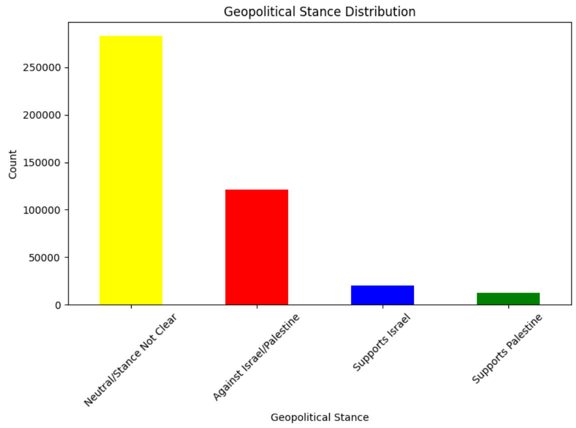


Fig. 5. Geopolitical Stance Distribution.

The dataset in this paper shows bias towards "Neutral" class. Thus, the largest geopolitical stance leans towards "Neutral /Stance Not Clear". Fig. 5 and Table I depict the Geopolitical stances of reddit users in our dataset using LDA model. Most of the comments had neutral words used in "Israel" and "Palestine" topics resulting in the biggest stance being "Neutral/Stance not clear". Negative words usage in both topics, positive words usage in "Israel" topic only and positive words usage in "Palestine" topic only resulted in rest of the three stances.

TABLE I
GEOPOLITICAL STANCE COUNTS

Stance	Count
Neutral/Stance Not Clear	283191
Against Israel/Palestine	120859
Supports Israel	19842
Supports Palestine	12822

TABLE II
COMPARATIVE ANALYSIS OF LEARNING MODELS

Model	Accuracy (%)	Precision	Recall	F1-score
LR	96.96	0.97	0.97	0.97
SVC	94.56	0.95	0.95	0.95
SGD	88.94	0.90	0.89	0.89
NB	77.66	0.84	0.78	0.80

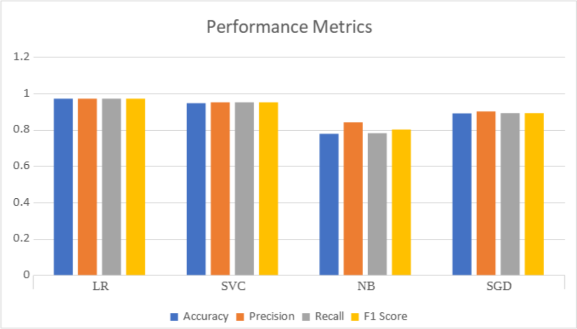


Fig. 6. Graphical representation depicting the performance metrics of each algorithm, including accuracy, precision, recall, and F1 score.

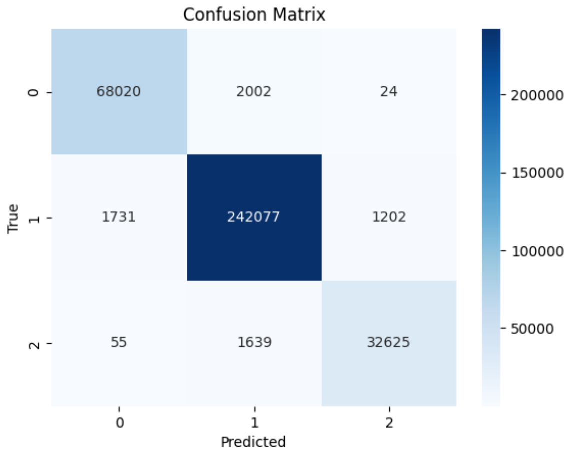


Fig. 7. Confusion Matrix of Logistic Regression.

All the models' performance is evaluated based on metrics such as accuracy, precision, recall, and F1 score. These results are tabulated in Table II. The graph depicting the performance metrics of all algorithms is demonstrated in Fig. 6.

Logistic regression is known for its robustness in handling imbalanced datasets. Since our dataset has imbalanced class distributions, logistic regression performed well in this scenario. As Logistic regression model showed the best performance, we considered it as the winner model. The confusion matrix of the winner model is shown in Fig. 7. Due to the significant class imbalance in our dataset, the confusion matrix displayed such a result.

VI. Limitations

The inclusion of lemmatization within the text data preprocessing phase and the adoption of enhanced preprocessing techniques would enable potential avenues for improvement in the research. Additionally, the implementation of oversampling or undersampling methodologies could enhance the robustness and reliability of the results obtained.

VII. Conclusion

The ongoing conflict between Israel and Palestine has sparked extensive discussions on various social media platforms. In this paper, we aimed to extract valuable insights from such social media platform, Reddit. Data labeling was performed using VADER, while LDA was employed to identify geopolitical stances. Various machine learning techniques, including Support Vector Classifier (SVC), Logistic Regression (LR), Stochastic Gradient Descent (SGD), and Naïve Bayes (NB), were utilized to train the models. The accuracy of these models ranged from 77.66% to 96.96%. Our findings show that the majority of people are 'Neutral' in their stance regarding this conflict. The findings underscore the significance of understanding sentiments for monitoring public opinions. We believe that our analysis will contribute to making informed decisions and taking proactive measures regarding political events in the future.

For our upcoming projects, we plan on utilizing even more extensive dataset. Our strategy includes implementing the TextBlob text labeling technique, various oversampling techniques to enhance the model's performance and incorporating advanced deep learning and transformer algorithms for more precise sentiment analysis.

References

[1] W. contributors, "2023 Israel–Hamas war," 1 2024.

[2] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[3] R. Shewale, "14+ Reddit Statistics for 2024 (Eye-Opening Facts Data)," 12 2023.

[4] Asaniczka, "Daily public opinion on israel-palestine war," 2024.

[5] "Sentiment analysis Machine learning," 4 2020.

[6] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, pp. 216–225, 2014.

[7] M. Bakrey, "All about Latent Dirichlet Allocation (LDA) in NLP - Mohamed Bakrey - Medium," 3 2023.

[8] N. S. L. S. Charitha, K. Yasaswi, V. Rakesh, M. Varun, M. Yeswanth, and J. S. Kiran, "Comparative study of algorithms for sentiment analysis on imdb movie reviews," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 824–828, IEEE, 2023.

[9] A. Gangwar and T. Mehta, "Sentiment analysis of political tweets for israel using machine learning," in *International Conference on Machine Learning and Big Data Analytics*, pp. 191–201, Springer, 2022.

[10] S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, "A machine learning-sentiment analysis on monkeypox outbreak: An extensive dataset to show the polarity of public opinion from twitter tweets," *IEEE Access*, vol. 11, pp. 11811–11826, 2023.

[11] S. Sabba, N. Chekired, H. Katab, N. Chekkai, and M. Chalbi, "Sentiment analysis for imdb reviews using deep learning classifier," in *2022 7th international conference on image and signal processing and their applications (ISPA)*, pp. 1–6, IEEE, 2022.

[12] I. Al-Agha and O. Abu-Dahrooj, "Multi-level analysis of political sentiments using twitter data: A case study of the palestinian-israeli conflict," *Jordanian Journal of Computers and Information Technology*, vol. 5, no. 3, 2019.

[13] K. Arora, N. Gupta, and S. Pathak, "Sentimental analysis on imdb movies review using bert," in *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 866–871, IEEE, 2023.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[15] N. Van Otten, "CountVectorizer Tutorial: How to easily turn text into features for any NLP task," 10 2023.

[16] G. L. Team, "An Introduction to Bag of Words in NLP using Python — What is BoW?," 10 2022.

[17] A. V. Ratz, "Multinomial Nave Bayes' for Documents Classification and Natural Language Processing (NLP)," 4 2022.

[18] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the twenty-first international conference on Machine learning*, p. 116, 2004.

[19] X. Yang, *Logistic regression, PCA, LDA, and ICA*. 1 2019.

[20] Generanger, "The Math behind Linear SVC Classifier," 8 2018.