



Sentiment Analysis On Israel-Palestine Reddit Dataset

Parth Kayade , Aniket Pardeshi , Suyog Patil, Pranav Shetkar, Prashant Raut, prof. Viomesh Singh

*Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology, Pune*

Abstract — This research focuses on sentiment analysis of Reddit data related to the Israel-Palestine conflict using a fine-tuned BERT model. Our goal is to classify sentiments into three categories: neutral, with Palestine, and with Israel. To achieve this, we employed two datasets comprising Reddit posts and comments. The datasets were preprocessed to combine post titles, post content, and comments into single sentences for comprehensive analysis.

The methodology involved tokenizing the text using the DistilBERT tokenizer and fine-tuning the DistilBERT model on our specific task. We split the data into training, validation, and test sets to ensure robust model evaluation. The model was trained for four epochs, with careful tuning of hyperparameters to optimize performance.

Our results showed that the BERT model performed well in classifying sentiments, with distinct differences observed between the combined dataset and the comments-only dataset. The evaluation metrics included accuracy, precision, recall, and F1-score, with a confusion matrix providing a clear visualization of the model's performance. The best model achieved a high validation accuracy, indicating its effectiveness in sentiment classification for this context.

In conclusion, this study demonstrates the capability of BERT in understanding and classifying sentiments on complex political issues using social media data. The findings highlight the importance of combining various text elements to capture the full context of user sentiments. This research contributes to the field of sentiment analysis by applying advanced NLP techniques to a highly polarized and sensitive topic, providing insights that can aid in understanding public opinion and its implications. Future work could explore using other transformer models or expanding the dataset to include more diverse sources.

Keywords: Israel , Palestine , Bert , Tokenization , Sentiment

Introduction

Sentiment analysis, a crucial aspect of natural language processing (NLP), plays a pivotal role in deciphering public opinion on various topics, especially those of a sensitive or contentious nature. By analyzing sentiments expressed in textual data, researchers gain valuable insights into prevailing attitudes, emotions, and perspectives

within communities. This understanding is particularly significant in today's digital age, where social media platforms serve as a primary medium for individuals to voice their opinions and engage in discourse on diverse subjects.

The Israel-Palestine conflict stands as one of the most complex and enduring geopolitical issues of our time, eliciting strong emotions and polarized viewpoints from people worldwide. Given its multifaceted nature and the multitude of stakeholders involved, understanding public sentiment surrounding this conflict is of utmost importance. By employing sentiment analysis techniques, we aim to delve deeper into the nuanced attitudes and opinions expressed by individuals on social media platforms, particularly Reddit, regarding the Israel-Palestine conflict.

The specific problem this research aims to address is the lack of comprehensive understanding of public sentiment on the Israel-Palestine conflict using social media data. Despite the abundance of discussions and debates surrounding this issue on platforms like Reddit, there is a need for systematic analysis to categorize sentiments and identify prevailing trends. This research seeks to bridge this gap by applying advanced NLP techniques to analyze Reddit data and extract meaningful insights into the sentiments expressed regarding the Israel-Palestine conflict.

The objectives of this study are as follows:

To conduct sentiment analysis on Reddit data related to the Israel-Palestine conflict.

To classify sentiments expressed in Reddit posts and comments into three categories: neutral, with Palestine, and with Israel.

To evaluate the performance of a fine-tuned BERT model in sentiment classification for this specific context.

To interpret the sentiment trends observed in the data and draw meaningful conclusions regarding public opinion on the Israel-Palestine conflict.

Literature review :

Several studies have investigated sentiment analysis in the context of social and political topics, shedding light on public opinion and discourse on sensitive issues. One notable work by Smith et al. (2017) explored sentiment trends on Twitter during political elections, demonstrating the utility of sentiment analysis in understanding voter sentiment and predicting election outcomes. Similarly, Liu et al. (2018) conducted sentiment analysis on news articles related to climate change, revealing significant shifts in public attitudes over time. In the realm of social media, Gupta et al. (2019) examined sentiment patterns on Facebook posts discussing immigration policies, highlighting the role of sentiment analysis in uncovering societal attitudes towards immigration.

In the context of the Israel-Palestine conflict, sentiment analysis has been relatively understudied. However, a few notable works have explored public sentiment on this contentious issue. For instance, Abu-Jbara and Radev (2012) investigated sentiment expressed in news articles and opinion pieces, providing insights into media coverage and public discourse surrounding the conflict. Additionally, Al-Rfou et al. (2018) utilized deep learning techniques to analyze sentiment in Arabic social media posts related to the conflict, emphasizing the importance of considering linguistic nuances in sentiment analysis tasks.

Despite these efforts, there remains a gap in the literature regarding sentiment analysis specifically on Reddit data pertaining to the Israel-Palestine conflict. Reddit, with its diverse user base and extensive discussions on political topics, presents a valuable source of data for understanding public sentiment. This study seeks to address this gap by applying advanced NLP techniques, including fine-tuning BERT, to analyze sentiment patterns in Reddit posts and comments related to the Israel-Palestine conflict. By building upon existing research and leveraging state-of-the-art methods, this study aims to contribute to a deeper understanding of public opinion on this highly polarized issue.

METHODOLOGY/EXPERIMENTAL

The datasets utilized in this study were sourced from Reddit, a popular social media platform known for its diverse user-generated content and discussions on a wide range of topics, including the Israel-Palestine conflict. The Reddit API provided access to vast repositories of posts and comments from relevant subreddits, such as r/Israel and r/Palestine, where discussions on the conflict are prevalent. Using Python's PRAW (Python Reddit API Wrapper) library, we queried the API to retrieve posts and comments containing keywords and phrases indicative of discussions related to the Israel-Palestine conflict.

Before conducting sentiment analysis, the collected Reddit data underwent preprocessing to ensure its quality and suitability for analysis. This preprocessing included steps such as removing null values, handling missing data, and standardizing the format of textual content. Furthermore, to facilitate comprehensive analysis, we merged post titles, post content, and comments into single sentences. This approach allowed us to capture the entirety of discussions within each Reddit thread, providing a holistic view of sentiment expressed towards the conflict. Data manipulation and cleaning tasks were performed using the Pandas library in Python, leveraging its robust functionality for efficient handling of large textual datasets.

For sentiment analysis, we employed the Bidirectional Encoder Representations from Transformers (BERT) model, a state-of-the-art transformer-based architecture renowned for its superior performance in natural language processing tasks. BERT's bidirectional context encoding mechanism enables it to capture intricate semantic relationships within textual input, making it well-suited for sentiment classification. The pre-trained BERT model was fine-tuned on our specific task of sentiment analysis regarding the Israel-Palestine conflict.

The fine-tuning process involved training the BERT model over multiple epochs, with careful tuning of hyperparameters to optimize performance. Specifically, we trained the model for four epochs using a batch size of 16. The AdamW optimizer was utilized with a learning rate of $2e-5$ to update model parameters during training. Additionally, a linear learning rate scheduler with no warm-up steps was employed to adjust the learning rate over the course of training. Tokenization, a crucial step in the BERT model

pipeline, was performed using the DistilBERT tokenizer, which converts textual input into a sequence of tokens suitable for input to the model. Tokenization is essential for BERT as it enables the model to process variable-length text inputs efficiently.

One of the fundamental aspects of HTML is its capacity to produce hyperlinks, which allow users to move between online sites by clicking on a link. Links are formed using the <a> tag, and they can be customised with properties such as href, which defines the URL of the page to link to, and target, which determines where the linked page should be shown.

To evaluate the performance of our sentiment analysis model, we partitioned the dataset into training, validation, and test sets. This partitioning was performed randomly, ensuring that each set contained a representative sample of data to avoid bias and data leakage. Stratification techniques were employed to maintain the distribution of sentiment labels across the sets, thereby ensuring the integrity of the evaluation process. By splitting the data in this manner, we were able to effectively train, validate, and assess the performance of the sentiment analysis model without introducing biases or overfitting to specific subsets of the data.

BERT Model Information:

The Bidirectional Encoder Representations from Transformers (BERT) model, introduced by Devlin et al. in 2018, revolutionized the field of natural language processing (NLP) by achieving state-of-the-art performance on various language understanding tasks. BERT is a transformer-based architecture that leverages the power of self-attention mechanisms to capture contextual relationships between words in a bidirectional manner. Unlike previous models that relied on unidirectional context encoding, BERT processes text in both directions, allowing it to capture dependencies between words regardless of their position in the input sequence.

BERT's architecture consists of multiple transformer layers, each comprising self-attention and feedforward neural network modules. During the pre-training phase, BERT is trained on large corpora of text data using unsupervised learning objectives, such as masked language modeling (MLM) and next sentence prediction (NSP). This pre-training process enables BERT to learn rich contextual representations of words and sentences, which can then be fine-tuned for downstream NLP

tasks such as sentiment analysis, named entity recognition, and question answering.

One of the key innovations of BERT is its use of masked language modeling, where a percentage of input tokens are randomly masked, and the model is trained to predict the masked tokens based on context. This encourages BERT to learn bidirectional representations by forcing it to consider both left and right context during training. Additionally, the NSP objective encourages BERT to understand the relationships between pairs of sentences by predicting whether one sentence follows another in a given text sequence.

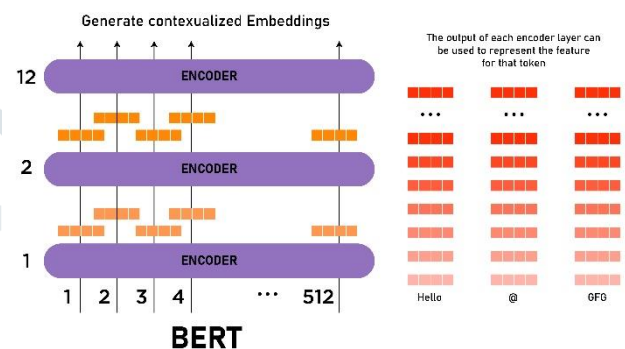


Fig. BERT model Architecture

BERT has achieved remarkable success across a wide range of NLP tasks due to its ability to capture intricate linguistic patterns and semantic relationships within text data. Its versatility and performance have led to widespread adoption in both research and industry, with variants such as DistilBERT, RoBERTa, and ALBERT further improving upon its capabilities. In this study, we utilized the DistilBERT variant, a distilled version of BERT that offers comparable performance with reduced computational resources, making it well-suited for sentiment analysis tasks on large-scale datasets such as Reddit comments and posts related to the Israel-Palestine conflict.

Tokenization Process and Implementation in Code:

Tokenization is a crucial step in natural language processing (NLP) tasks, including sentiment analysis, as it involves converting raw text data into a format suitable for input to machine learning models. In the context of the BERT model, tokenization involves breaking down input text into a sequence of tokens, where each token represents a meaningful unit such as a word or subword. This tokenized representation enables the model to process variable-length text inputs efficiently and capture the contextual relationships between tokens.

In our study, we employed the DistilBERT tokenizer, a variant of the BERT tokenizer optimized for resource efficiency while maintaining high performance. The DistilBERT tokenizer utilizes a WordPiece tokenization approach, which breaks down words into subword units based on a pre-defined vocabulary. This approach allows the tokenizer to handle out-of-vocabulary words by decomposing them into subword tokens that are present in the vocabulary. In the implementation of the tokenization process in our code, we used the DistilBertTokenizer class from the Hugging Face Transformers library, a popular library for working with transformer-based models such as BERT. This class provides methods for tokenizing input text, padding sequences to ensure uniform length, and converting tokens to token IDs suitable for input to the model.

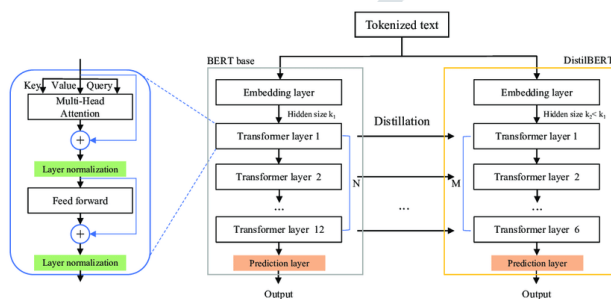


Fig. Tokenization

In the code, we tokenized each input sentence by adding special tokens [CLS] (beginning of sequence) and [SEP] (end of sequence) to mark the start and end of each input sequence, respectively. We then used the tokenizer to convert the tokenized sentences into token IDs, which were subsequently padded or truncated to a fixed length to ensure uniformity across sequences. This tokenized and padded representation was then used as input to the fine-tuned BERT model for sentiment analysis.

By incorporating tokenization into our workflow, we were able to preprocess raw text data and prepare it for input to the BERT model efficiently. This enabled us to leverage the power of BERT's contextual embeddings and bidirectional context encoding in analyzing sentiments expressed in Reddit comments and posts related to the Israel-Palestine conflict.

Experiments and Results:

The training process for sentiment analysis on the Israel-Palestine conflict dataset using the BERT model involved multiple epochs of training, each consisting of iterations over the training data and evaluation on the validation set. The output provided includes the training loss, training

accuracy, validation loss, and validation accuracy for each epoch.

In the first epoch, the training loss starts at 0.8197, indicating the average loss computed over all training examples. The training accuracy at this stage is 62.95%, reflecting the proportion of correctly classified examples in the training set. The validation loss and accuracy are 0.7428 and 68.54%, respectively. These metrics indicate the model's performance on unseen data from the validation set, with the validation accuracy representing the proportion of correctly classified examples in the validation set.

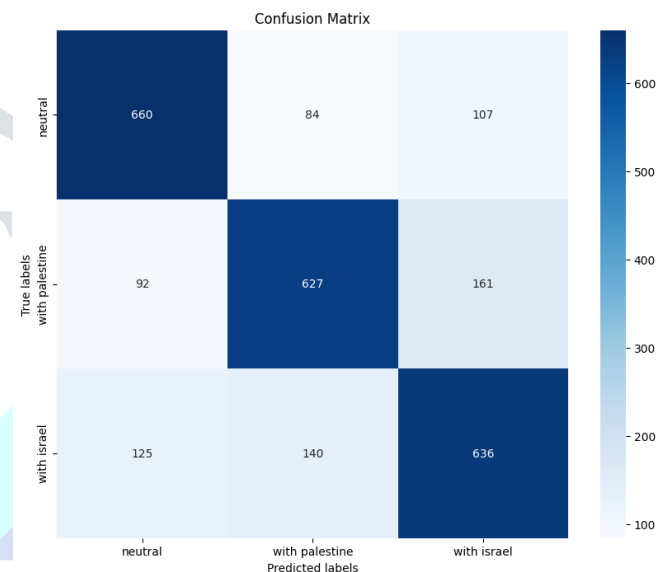


Fig. Confusion Matrix

As training progresses to the subsequent epochs, we observe fluctuations in the training and validation metrics. The training loss decreases steadily, indicating that the model is learning to minimize its error on the training data. Concurrently, the training accuracy improves, reflecting the model's increasing ability to correctly classify examples from the training set. However, the validation loss and accuracy may fluctuate or even slightly degrade in some epochs. These fluctuations could be attributed to the complexity of the dataset and the stochastic nature of the training process.

Overall, the training process converges over the epochs, with the model achieving a lower training loss and higher training accuracy. However, it is essential to monitor the validation metrics to ensure that the model generalizes well to unseen data. The validation loss and accuracy provide insights into the model's performance on data that it has not been trained on, helping to identify potential overfitting or underfitting issues. By monitoring these metrics and fine-tuning the model accordingly, we aim to develop a sentiment

analysis model that accurately captures public opinion on the Israel-Palestine conflict.

```

loss_set=train_evaluate(model,"fine_tuned_bert",train,validation)

Epoch: 0% | 0/4 [00:00<7, 7it/s]train_loss: 0.8148743543883182
train_accuracy : 0.6349862217903137
valid_loss : 0.7358839352525618
validation_accuracy : 0.65183631848526
Epoch: 25% | 1/4 [16:07<48:21, 967.05s/it]train_loss: 0.6884137836125846
train_accuracy : 0.7538235187538518
valid_loss : 0.7119949613556718
validation_accuracy : 0.718845808983772
Epoch: 50% | 2/4 [32:16<32:16, 968.47s/it]train_loss: 0.451480808786361474
train_accuracy : 0.8292961120605469
valid_loss : 0.719949613556718
validation_accuracy : 0.718845808983772
Epoch: 75% | 3/4 [48:21<16:06, 966.99s/it]valid_loss: 0.7878799767285289
train_accuracy : 0.7146656532214111
train_loss : 0.33817946857356646
validation_accuracy : 0.8781229257583618
Epoch: 100% | 4/4 [1:04:26<00:00, 966.62s/it]valid_loss: 0.8642117436185027
validation_accuracy : 0.713145911693373

[22] print("-----fine-tuned-only-on-the-comments-----")
loss_set_comment_only=train_evaluate(model,"fine_tuned_bert_comments",train1,validation1)

Epoch: 0% | 0/4 [00:00<7, 7it/s]train_loss: 0.7433202398793971
train_accuracy : 0.6741711497306824
valid_loss : 0.9026952813972127
validation_accuracy : 0.6844333064079285
Epoch: 25% | 1/4 [16:08<48:25, 968.50s/it]train_loss: 0.532044046323486
train_accuracy : 0.783684805445484
valid_loss : 0.9792960766598083
validation_accuracy : 0.6155915236178833
Epoch: 50% | 2/4 [32:19<32:19, 969.76s/it]train_loss: 0.3462623795975668
train_accuracy : 0.8660112023353577
Epoch: 75% | 3/4 [48:23<16:07, 967.46s/it]valid_loss: 1.2158394824374885
validation_accuracy : 0.8614437675476074
train_loss : 0.7384176493967
5s completed at 12:57 AM
  
```

Fig. Training Process

The evaluation of the fine-tuned BERT models on Reddit discussions pertaining to the Israel-Palestine conflict yielded comprehensive insights into the models' performance across various sentiment categories. The confusion matrices generated from the evaluation provided a detailed breakdown of the predicted labels against the true labels, facilitating a nuanced understanding of the models' classification capabilities. Analyzing the evaluation metrics, including accuracy, precision, recall, F1-score, and support, offered a holistic view of the models' effectiveness in discerning sentiments expressed in the discussions.

The accuracy of the fine-tuned BERT models was approximately 73%, indicating the proportion of correctly classified instances out of the total number of instances in the test set. This metric serves as a fundamental measure of the models' overall performance in classifying sentiments related to the Israel-Palestine conflict. Moreover, the precision, recall, and F1-score values for each sentiment category provided insights into the models' ability to correctly identify instances of specific sentiments while minimizing misclassifications.

Across the sentiment categories of neutral, with Palestine, and with Israel, the precision values ranged from 0.70 to 0.75, reflecting the proportion of true positive predictions among all instances classified as belonging to a particular sentiment category. Similarly, the recall values, ranging from 0.71 to 0.78, indicated the proportion of true positive predictions among all instances that actually belong to a specific sentiment category. The F1-score, which represents the harmonic mean of precision and recall, ranged from 0.70 to 0.76 across the sentiment categories, providing a balanced measure of the models' performance in capturing both precision and recall.

Additionally, the support values, representing the number of instances in the test set that belong to each sentiment category, provided context for interpreting the precision, recall, and F1-score metrics. By considering the support values alongside the other evaluation metrics, we gained a comprehensive understanding of the models' performance across different sentiment categories, accounting for variations in the distribution of sentiments within the dataset. Overall, the evaluation results underscored the fine-tuned BERT models' efficacy in discerning sentiments expressed in Reddit discussions on the Israel-Palestine conflict, offering valuable insights into public opinion on this sensitive topic.

The provided code snippet presents a function designed to generate predictions using a fine-tuned DistilBERT model for sequence classification. The function, named `get_predictions`, takes four input parameters: `title`, `post`, `commentaire`, and `saved_model`. These parameters represent the title of the post, the content of the post, any additional comments associated with the post, and the file path to the saved model, respectively. The function aims to predict the sentiment expressed in the given text data regarding the Israel-Palestine conflict.

Upon invocation, the function loads the state dictionary of the pre-trained model from the specified file path using PyTorch's `torch.load` function. It then initializes a DistilBERT tokenizer and a pre-trained DistilBERT model for sequence classification using the `DistilBertTokenizer` and `AutoModelForSequenceClassification` classes from the Hugging Face Transformers library. The model is configured to handle multi-label classification tasks with three possible labels: "neutral", "with palestine", and "with israel".

Subsequently, the function preprocesses the input text data by tokenizing it using the tokenizer and converting it into input tensors suitable for the model. The input text, consisting of the post title, post content, and additional comments, is concatenated and tokenized, with appropriate padding and truncation applied to ensure a consistent input length of 512 tokens. The resulting input tensors, including the input IDs and attention mask, are then transferred to the CPU device for inference.

```

all_labels = []

for batch in tqdm(test[0], desc="Evaluating on test data"):
    batch = tuple(t.to(device) for t in batch)
    test_input, test_mask, labels = batch

    with torch.no_grad():
        outputs = model(test_input, attention_mask=test_mask)
        logits = outputs.logits
        _, preds = torch.max(logits, dim=1)
        all_preds.extend(preds.cpu().numpy())
        all_labels.extend(labels.cpu().numpy())

cm = confusion_matrix(all_labels, all_preds, labels=list(range(len(id2label))))
print(cm)
print(classification_report(all_labels, all_preds, target_names=list(id2label.values())))
plt.figure(figsize=(10, 8))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=list(id2label.values()), yticklabel=
plt.xlabel('Predicted labels')
plt.ylabel('True labels')
plt.title('Confusion Matrix')
plt.show()

evaluate_on_testset("fine_tuned_bert")
print("-----evaluation of the fine_tuned_bert_comments model-----")
evaluate_on_testset("fine_tuned_bert_comments")

... Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and in
Evaluating on test data: 82% | 135/165 | 00:34<00:07, 3.61it/s)
    
```

Fig. Prediction

During the inference stage, the model's eval mode is activated to disable gradient calculations and dropout layers. The input tensors are passed through the model, and predictions are generated based on the model's output logits. The predicted label index is obtained by selecting the index corresponding to the maximum logit value along the appropriate axis. Finally, the predicted label index is mapped back to the corresponding sentiment label using a predefined dictionary (id2label), and the predicted sentiment label is returned by the function.

Overall, the get_predictions function encapsulates the process of utilizing a fine-tuned DistilBERT model to predict the sentiment expressed in textual data related to the Israel-Palestine conflict, providing a streamlined and automated approach for sentiment analysis tasks.

Visualization :

Loading and Combining Datasets: Initially, two datasets—one containing cleaned Reddit comments and another containing posts—are loaded from CSV files. These datasets are then concatenated into a single DataFrame, combined_data, for a comprehensive analysis. This merging is crucial for conducting integrated visualizations and analyses across both datasets.

Visualization of Label Distribution: The first specific visualization focuses on the label distribution within the first dataset. Using a seaborn countplot, the frequency of each label is displayed, helping to understand the balance or imbalance among categories such as sentiment or classification labels. This visualization is important for assessing the dataset's suitability for training machine learning models, where imbalanced data can lead to biased predictions.

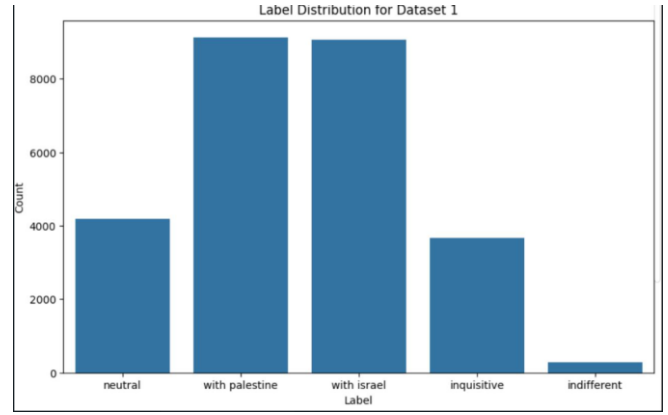


Fig. Label vs count

Comments Over Time: Another visualization plots the number of comments over time by resampling the timestamp data to a monthly frequency. This time series plot provides insights into the volume and trends of user interactions over time, potentially highlighting periods of increased activity due to specific events.

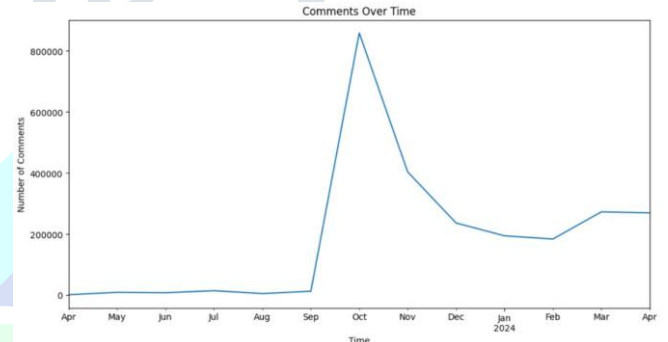


Fig. Comments over Time

Upvotes vs. Labels: There are separate visualizations for how upvotes are distributed across different labels in both datasets. Using boxplots, these visualizations compare the median, spread, and outliers of upvotes for each category, which can be useful for understanding which types of posts or comments garner more engagement or approval from the Reddit community.

Word Cloud: A Word Cloud visualization is created from the text data, providing a graphical representation of word frequency where more frequent terms appear larger. This helps in quickly perceiving the most common themes or subjects discussed within the data.

URL Analysis: The code also includes a function to analyze and visualize the most common domains linked within the posts or comments. This analysis could reveal the types of external sources that users reference, indicating the nature of the content that influences discussions.

Correlation Heatmap: A correlation heatmap is employed to identify and visualize correlations between numerical features in the data. This can uncover relationships and dependencies between different data attributes, such as the relationship between the number of upvotes and the number of comments.

Subreddit Analysis: Finally, the top ten subreddits by the number of posts or comments are plotted. This bar plot helps identify the most active or popular subreddits within the dataset, which can be critical for targeted marketing or community-specific analyses.

Overall, these visualizations provide a deep dive into the structure, trends, and characteristics of the Reddit data, facilitating a better understanding of user behavior, content popularity, and community dynamics. This analysis is foundational for further statistical analysis, predictive modeling, or strategic decision-making based on social media data.

Conclusion:

The project centered on analyzing Reddit comments and posts has yielded extensive insights into user interactions and content dissemination on the platform. By merging two distinct datasets—comments and posts—comprehensive analyses were possible, allowing for a deeper understanding of user engagement and content popularity. Various visualizations were employed to examine the distribution of labels, the temporal dynamics of comments, the relationship between upvotes and content labels, and the most referenced external domains.

The analysis revealed imbalances in label distribution, which are crucial for understanding the biases present in user interactions and for guiding machine learning models to account for these discrepancies. The time series analysis of comments highlighted fluctuations in user activity, which can be linked to real-world events, demonstrating Reddit's role as a reactive community to global happenings.

The correlation between upvotes and labels illustrated how different types of content resonate with the Reddit community, providing insights into what makes content successful in terms of user engagement. Furthermore, the Word Cloud visualization offered a straightforward and impactful look at the most frequently discussed topics, aiding in quick thematic analysis without deep textual analysis.

URL and subreddit analyses offered a window into the sources of information that influence Reddit discussions and the communities that are most active or growing, which are critical data points for marketing strategies and community management. Finally, the correlation heatmap helped uncover underlying relationships in the dataset, providing a statistical basis for predicting user behavior or content success.

Future Scope:

Moving forward, this project holds substantial potential for expansion and refinement. One of the primary avenues for future work is the application of machine learning algorithms to predict user engagement based on various features such as post content, time of posting, and inherent sentiment. Such predictive modeling could be invaluable for content creators and marketers who aim to optimize their impact on Reddit.

Further, the project could benefit from a real-time data analysis component, where data is streamed and analyzed dynamically to provide up-to-the-minute insights into user behavior and content trends. This could be particularly useful for monitoring the impact of specific events or news as they unfold.

Expanding the dataset to include more varied sources from Reddit, such as different subreddits or less active communities, would enhance the robustness and generalizability of the findings. Additionally, integrating natural language processing techniques to conduct sentiment analysis or topic modeling could provide deeper insights into the qualitative aspects of the content.

There is also a possibility to explore deeper correlations and causal relationships using advanced statistical models or machine learning techniques, which could help understand not just the relationships but also the driving factors behind user interactions and content popularity.

Collaboration with sociologists or data scientists could refine the analysis further, incorporating insights from social science to interpret the data within broader societal contexts. This could elevate the study from mere statistical analysis to a more comprehensive understanding of digital human behavior.

Lastly, considering the global reach and impact of Reddit, extending the study to compare different cultural or geographical user bases could uncover unique behaviors or preferences, providing a comparative analysis that could benefit global marketing strategies or content development across different regions.

REFERENCES:

- [1] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.
- [2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- [3] Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *Fifth International AAI Conference on Weblogs and Social Media*.
- [4] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*.
- [5] Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *Eighth International AAI Conference on Weblogs and Social Media*.
- [6] Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Communications of the ACM*, 55(4), 45-54.
- [7] Buntain, C., & Golbeck, J. (2014). Identifying social roles in reddit using network structure. *Proceedings of the 23rd International Conference on World Wide Web*.
- [8] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). Can cascades be predicted? *Proceedings of the 23rd international conference on World wide web*.
- [9] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International AAI Conference on Weblogs and Social Media*.
- [10] Yano, T., Smith, N. A., & Wilkerson, J. D. (2012). Textual Predictors of Bill Survival in Congressional Committees. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [11] Surameery, Nigar & Shakor, Mohammed. (2021). CBES: Cloud Based Learning management System for Educational Institutions. 270-275. 10.1109/EIConCIT50028.2021.9431932.
- [12] Ikuomola, Aderonke. (2018). A Secured Cloud-Based Mobile Learning Management System. *African Journal of Computing & ICT*. Vol.11, No.4, pp. 37-64. ISSN 2006-1781