Social Media Analytics Project

Mohamed Mustaf Ahmed

In 2024, many things happened in Mogadishu, that shocked everyone and got people talking on social media. This project looks at one of these social media conversations to understand how people felt about the kidnapping and murder of a young businessman, which has become a big topic in Somalia.

Introduction to Dataset

On February 24, 2024, a terrible crime happened in Mogadishu: Kaabax, a young businessman, was kidnapped and then killed. The person behind this awful act was Liban, a former business partner who lives in Dubai. After a disagreement that led to Liban being kicked out of the company and replaced by Kaabax, Liban planned a plot that ended with Kaabax's death.

Qoslaaye, the police captain in Heliwaa district, was a key part of this plot. He was related to Liban and helped plan the kidnapping. Six people were involved in the crime. The crime was discovered when a witness (Victim’s cousin) was accidentally let go by the kidnappers. This led to an investigation and the arrest of everyone involved, including Qoslaaye and his team.

In the first week of April, the case took a significant turn. The military court sat for the case, and decisions were made. This is when the case heated up and received more attention on social media. The court handed down different sentences to the individuals involved. Qoslaaye and Qoone, who played a key role in the plan and killed Kaabax, were sentenced to death. Liban was sentenced to 15 years of military imprisonment, and the rest were sentenced to 10 years.

This incident caused a new wave of discussions on social media, with people expressing their views on the justice system, the severity of the sentences, and the implications for societal values and norms. The project will go deeper into these discussions, analyzing the sentiments expressed and the impact of this verdict on the collective conscience of the Somali community.

This project wants to look at the digital traces of this event on social media, especially YouTube comments. These comments are how the Somali community shared their feelings, thoughts, and reactions to the crime. The project will look at over 1,800 comments from videos about the incident to understand the range of feelings—positive (wanaag), negative (xumaan), and neutral (dhexdhexaad)—expressed by the public. This will help us understand how the community sees justice and morality, and will also show us the societal values, tribal affiliations, and the collective conscience of the Somali people in response to such a tragic event.

How We Do It

The project uses a structured approach to social media analytics. It starts with collecting data through YouTube's API, then cleaning the data, doing exploratory data analysis (EDA), sentiment analysis, and finally, visualizing the findings. We use machine learning techniques to categorize the sentiments of the comments, which helps us understand public opinion in a detailed way. Through this thorough analysis, the project tries to capture the essence of the community's response, highlighting the different perspectives and the underlying feelings that are part of the social conversation about the case.

Why It Matters

By looking at the stories that came out on social media after this crime, the project hopes to provide valuable insights into how public sentiment, societal values, and the impact of social media as a platform for collective expression work in Somalia. This analysis not only helps us understand how the Somali community reacted to the incident but also shows the role of social media analytics in understanding complex societal trends and feelings.

By bringing together data, sentiment, and societal conversation, this project shows the power of social media analytics in connecting individual expressions and collective feelings, offering a detailed picture of a community in the face of tragedy.

Data Collection

We used a step-by-step process to collect social media data for this project.

Using the YouTube API

Setting Up the API: First, we got an API key from the Google Developer Console. This key let us make requests to the API to get data.

Finding Video IDs: The main part of our data collection was using video IDs. Each YouTube video has a unique video ID in its URL. We got 10 videos that had a lot of discussion about the case we were studying.

Getting the Data: After we found the video IDs, we made a Python script to get the data automatically. This script used the YouTube API and the video IDs to get comments. For each video, we got data like the comment text, user ID, number of likes, number of replies, and the length of each comment.

Storing Data

Making CSV Files: As we got data from each video, we stored it in separate CSV files. Each file had comments from one video. These CSV files had columns for all the data points we got from the API.

Combining the Data: After we got data from all the videos, we combined the CSV files into one big dataset. This let us clean up the data and analyze it on a larger scale.

Task 2

Looking at the Dataset

The dataset, which comes from YouTube comments, has 1,837 entries and six columns. These columns are:

- **`author`:** The person who wrote the comment.
- **`date`:** When the comment was written.
- **`Comment`:** The text of the comment.
- **`like_count`:** How many likes the comment got.
- **`reply_count`:** How many replies the comment got.
- **`comment_length`:** How long the comment is in characters.

Summary of Descriptive Statistics

We looked at the dataset's numerical attributes and found:

- **like_count:** The most likes a comment got was 255, but on average, comments got about 2.07 likes. This shows that a few comments got a lot more likes than others.

- **reply_count:** The most replies a comment got was 18, but most comments didn't get any replies.

- **comment_length:** The length of comments varied a lot, from 1 to 1,734 characters. This shows that people comment in many different ways.

Out[ ]:

	like_count	reply_count	comment_length
count	1837.000000	1837.000000	1837.000000
mean	2.073489	0.208492	98.122482
std	9.442899	0.926242	103.404849
min	0.000000	0.000000	1.000000
25%	0.000000	0.000000	39.000000
50%	0.000000	0.000000	72.000000
75%	1.000000	0.000000	125.000000
max	255.000000	18.000000	1734.000000

Checking the Data's Integrity and Structure

Shape of the Data: The dataset has 1,837 rows and 6 columns, which means we have a lot of data to analyze.

Data Types: The data types are correct for analysis, with text data as `object` and numerical data as `int64`.

Task 3

Let’s clean the data by removing the extra data and outliers, dropping or filling the missing values, etc. Create the final DataFrame for further analysis.

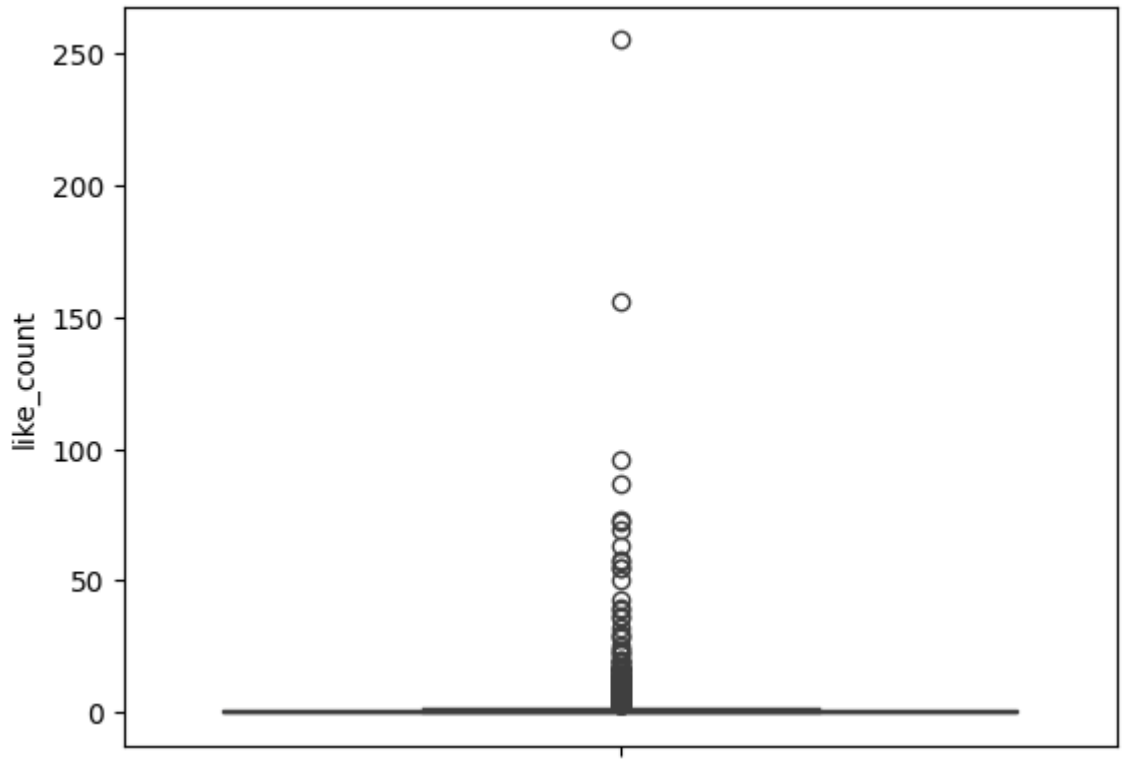
Checking for Missing Values

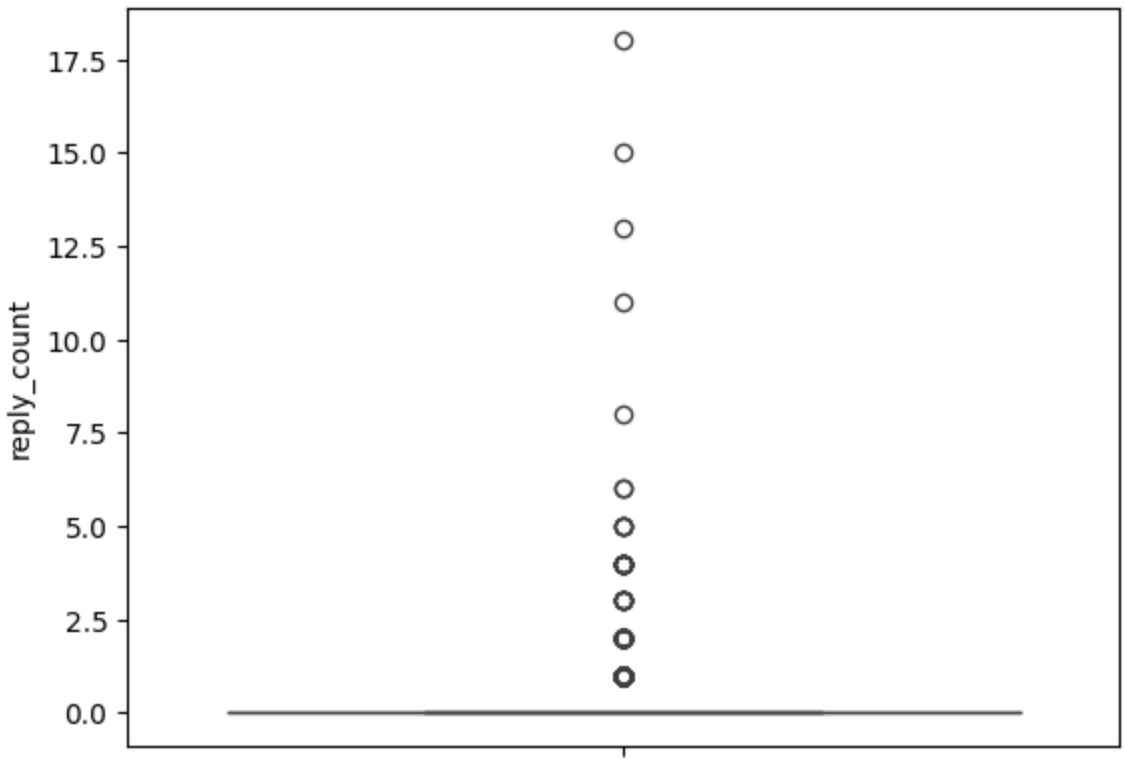
We didn’t find any missing values in the dataset. We checked this by using .isnull().sum() on all the columns, and it returned zeros for each column.

Checking for Duplicates

We checked for duplicate entries in the dataset to make sure each data point was different. We did this by using the `.duplicated().sum()` method in Pandas, which found one duplicate entry. We got rid of the duplicate we found using the `.drop_duplicates()` method, making sure the dataset was unique and relevant.

Looking at Outliers





We found outliers, especially in the `like_count` and `reply_count` columns, by creating box plots.

Because this is social media data, where more likes and replies can mean the comment is more important or engaging, we decided not to get rid of these outliers. We thought these data points could give us valuable insights into how people engage and feel.

Checking Data Types

We made sure each column had the right data type for analysis. We checked the data types using the `.dtypes` attribute. We confirmed that all columns had the right data types (`object` for text data and `int64` for number data), making the analysis straightforward.

Task 4

Detailed Look at Analyzing and Showing Comment Length and Word Counts

Counting Words

We counted the words in each comment by splitting the text on spaces and counting the parts.

We used this code: `df['word_count'] = df['Comment'].apply(lambda x: len(x.split()))`. We added a new column, `word_count`, to the dataset to show how wordy the comments are.

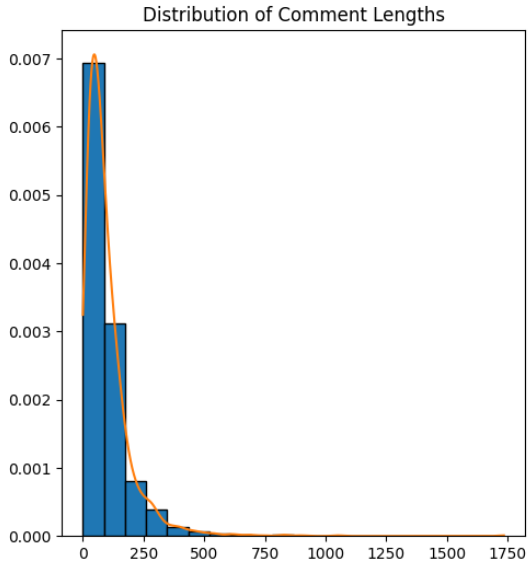
Analyzing Comment Lengths

We looked at how long the comments were in characters to see how wordy they were. We used statistical tools to understand how this attribute was distributed. The descriptive statistics showed a wide range in comment lengths, from short to long contributions.

	comment_length	word_count
count	1836.000000	1836.000000
mean	98.155773	15.517429
std	103.423174	16.455248
min	1.000000	1.000000
25%	39.000000	6.000000
50%	72.000000	11.000000
75%	125.000000	20.000000
max	1734.000000	282.000000

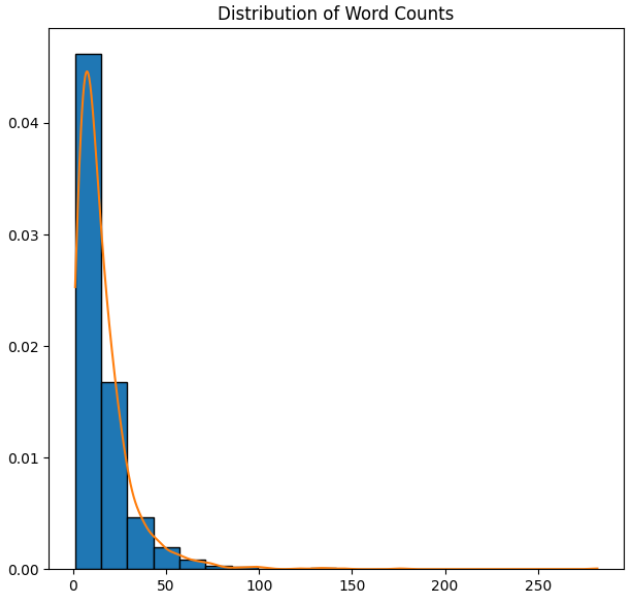
Showing Comment Length Distribution

We made a histogram to show the distribution of comment lengths. The distribution is skewed right, with most comments being short, but with a tail of longer comments showing that some users provide more detailed discussion.



Showing Word Count Distribution

We also made a histogram for word counts. The pattern looks like the distribution of comment lengths, with most comments having fewer words, but with some comments being quite wordy.



Task 5

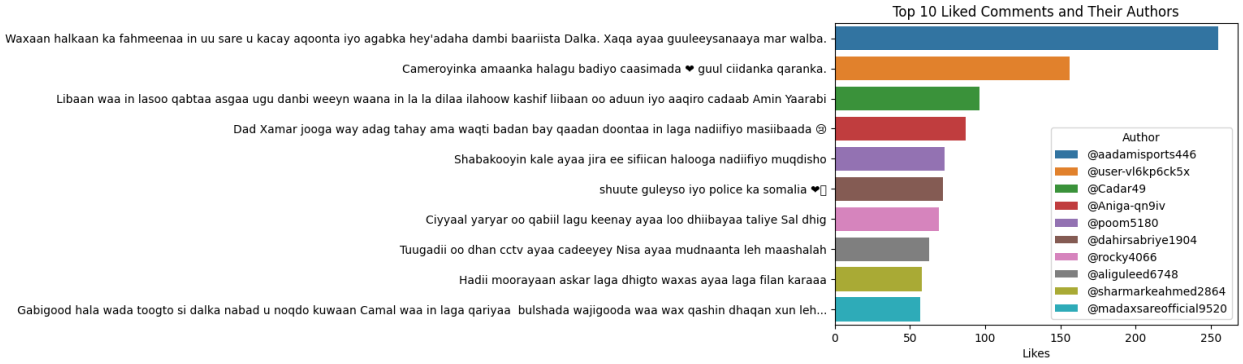
Detailed Look at Analyzing and Showing the Most Popular comments

Measuring Engagement

We measured the engagement of comments by sorting them based on the number of likes they got. We sorted the dataset by ‘like_count’ using the .sort_values() method, which showed us the most popular comments. We found the comments that were the most influential in the online discussion.

Showing the Top 10 Liked Comments

We made a horizontal bar chart to show the top 10 comments with the most likes. The length of each bar shows the number of likes, giving a clear comparison of engagement among the top comments.



The chart clearly shows the different levels of popularity of the comments, with some comments getting a lot more likes than others.

We looked at the content of the top-liked comments to find any common themes or feelings that resonated with the community. We found that certain topics or expressions likely led to higher engagement, such as calls for justice, expressions of grief, or thoughtful reflections on the incident.

We made another horizontal bar chart to show the top comments with the most replies, using the length of the bars to show the number of replies.

Task 6

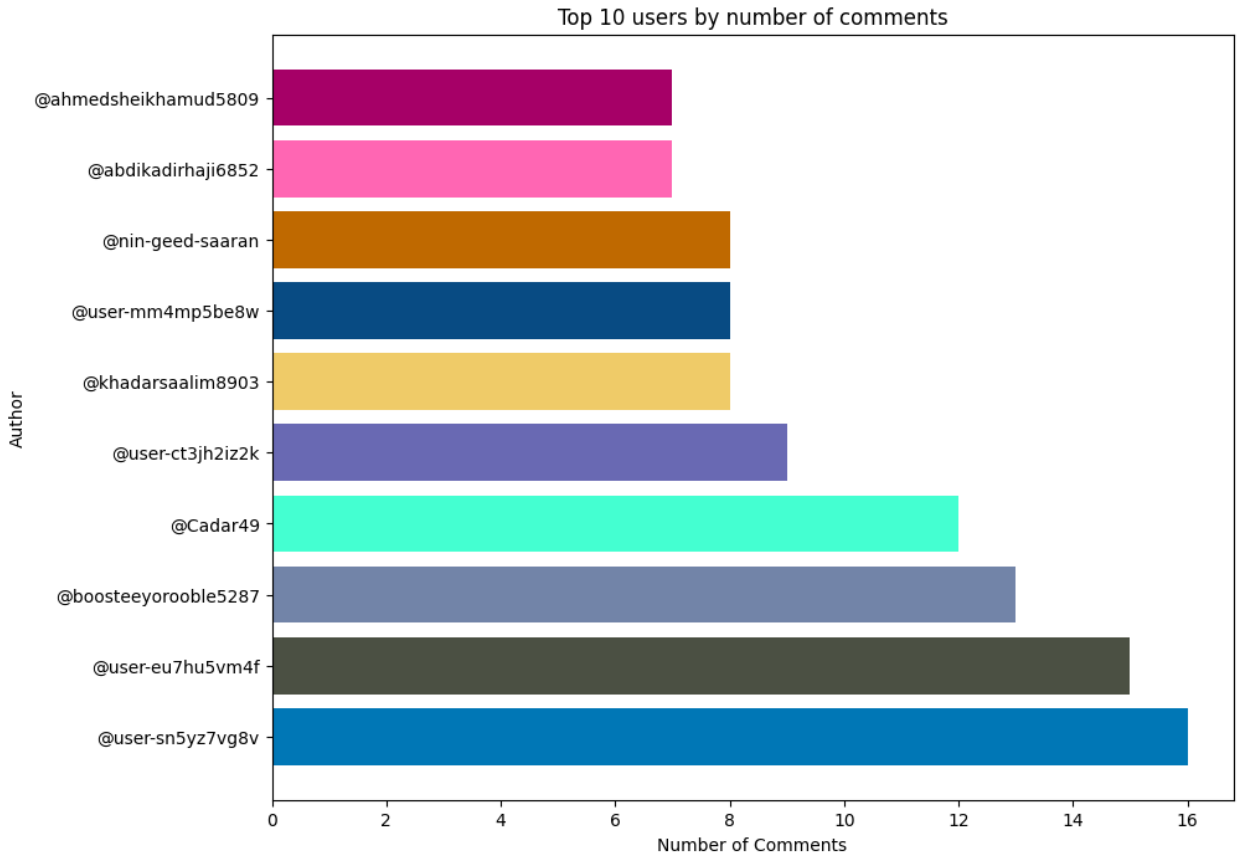
The goal of Task 6 is to find and show the users who commented the most. This can help us understand who was the most active in the discussion.

Measuring User Activity

We measured how active each user was by counting the number of comments they posted. We grouped the comments by user using the .groupby() method in Pandas and then counted the number of comments for each user. We found the users who contributed the most to the discussion.

Showing the Top 10 Users

We made a horizontal bar chart to show the top 10 users by the number of comments they posted.



The chart (picture provided) clearly shows the most active users, with the length of each bar showing their relative activity levels in terms of comment count.

Task 7

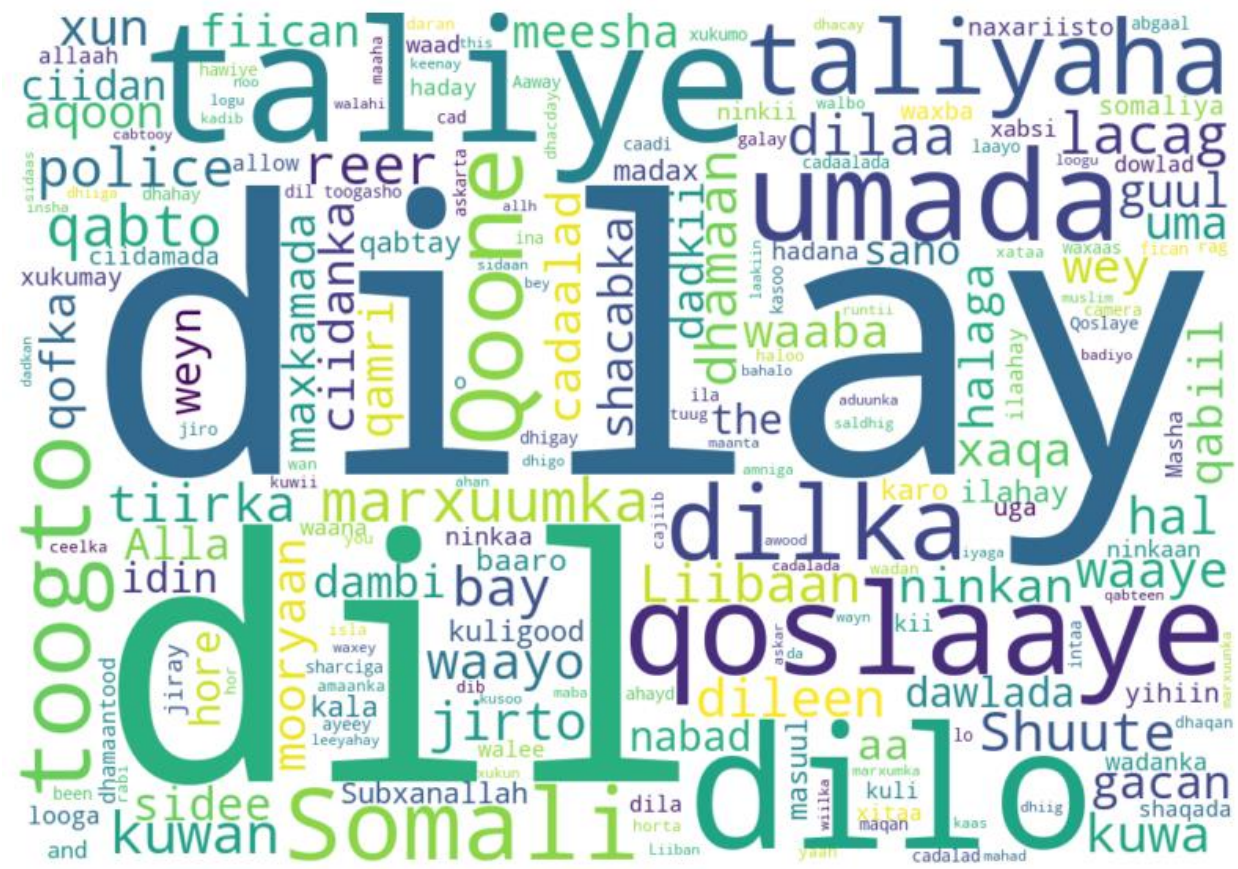
Detailed Look at Analyzing and Showing the Most Common Words

The goal of Task 7 is to make a WordCloud to show the most common words in the comments. This can help us understand the main themes and feelings within the community's discussions.

Preparing Data and Mentioning Stopword

We put all comments into one text group and processed it to leave out common stopwords. Stopwords are usually the most common words in a language that don't have unique meaning (like "aa," "ayey," etc.). Leaving out stopwords let us focus the WordCloud on words that are more important to the specific context of the discussions, giving a clearer picture of the community's main points.

Making and Interpreting WordCloud



The provided WordCloud shows words in different sizes, based on how often they appear in the dataset.

Key Words:

- **Dilay' (Killed):** The most common word, showing the community's focus on the outcome of the case.

- **Dil' (Kill):** The second most common word, showing discussions centered around the act of killing.

- **Taliye' (Captain):** The third most common word, likely referring to the police captain involved.

- **'Dilo' (Sentence him to death):** A word related to the community's feeling towards justice.

- **‘Qoslaaye’:** The name of the police captain, pointing to the main person in the incident.
- **‘Dilka’ (The killing):** Relating to the event itself, which is the topic of discussion.

The WordCloud picture shows the important issues and words the community talks about the most, especially those related to the incident and the quest for justice.

The importance of words related to ‘killing’ and ‘justice’ shows the community’s serious concerns and the demand for accountability in the face of the tragedy. It’s a good picture of the collective sentiment and the specific parts of the incident that are most important to the commenters.

Task 8

Let’s visualize in what STATE/COUNTRY were the top posts posted that mention your TOPICT.

This task can't be done because YouTube data does not contain users Location.

Task 9

The goal of Task 9 is to find and show the days in the month of April (The month which the case heated up as the captain and his team appeared in court). This is when the topic got the most attention on social media. This can help us understand when and how much the public interacted.

Changing Date Data

We changed the ‘date’ column to the datetime format to make it easier to analyze over time using df['date'] = pd.to_datetime(df['date']). We made sure we changed it correctly by checking the DataFrame information with df.info().

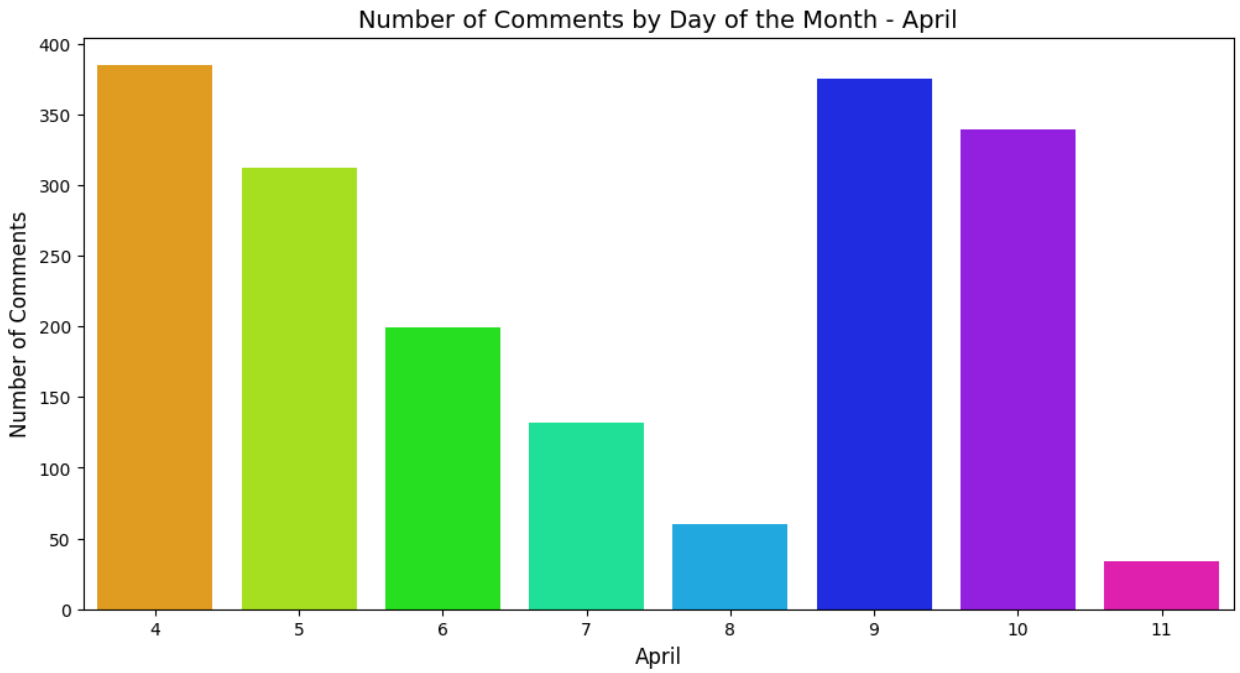
Taking Out Date Parts

We took out the year, month, and day from the ‘date’ column to analyze how often comments were made. We added new columns for the year, month, and day using the .dt accessor.

Grouping Comments by Day

We grouped the comments by day to count the number of comments for each day of the month. We used df.groupby('day')['Comment'].count() to make a series of comment counts by day. We got a series with the count of comments per day, showing the days with the most activity.

Showing Comment Distribution



The provided bar chart shows the distribution of comments across different days of April. The picture shows that the 4th and 9th of April were the days with the most comments, suggesting these days were the days the court hearing and the sentences were made.

The 4th and 9th of April stand out as the busiest days, which could coincide with significant developments in the case. The case was heard in court and decision made on the 4th and 5th, and a documentary about the crime was released on the 9th and 10th of April. This pattern of engagement over time provides valuable insights into how the community's attention shifted over the days.

Task 10

The goal of Task 10 is to collect a lot of statements from social media and put them into sentiment groups. Then, a sentiment analysis model was trained which was then used to see what the public's sentiment is about the 'Qoslaaye' case.

Collecting Sentiment Statements

We aimed to collect at least 3,000 statements from social media, which express sentiments that are positive, negative, or neutral. The statements we collected made a diverse dataset, showing a wide range of public sentiment.

Labeling Sentiment and Preparing Dataset

We labeled the collected statements as ‘Wanaag’ (Positive), ‘Dhexdhexaad’ (Neutral), or ‘Xumaan’ (Negative), based on the sentiments they expressed.

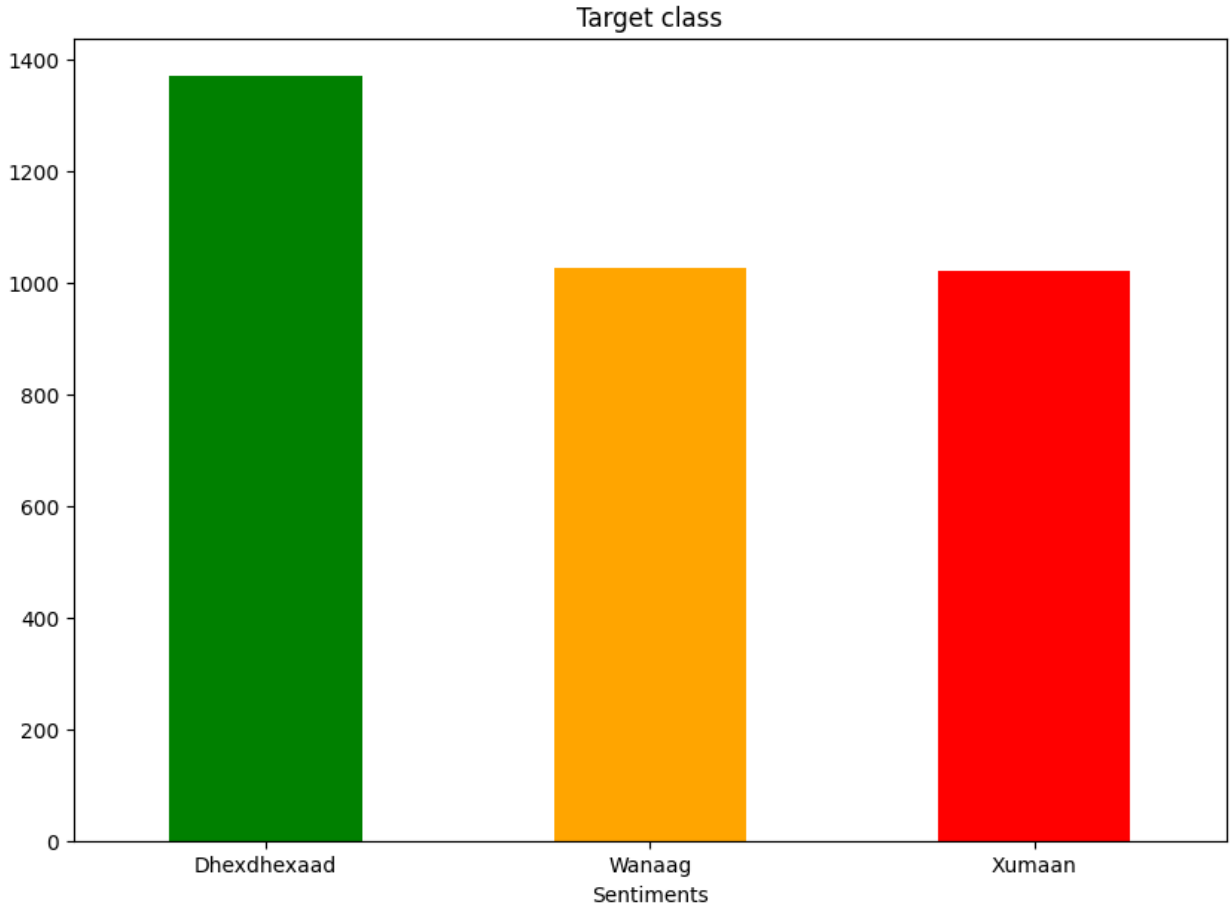
This labeled dataset served as the foundation for training a machine learning model that can classify sentiment.

Training the Sentiment Analysis Model

We trained a machine learning model on the labeled dataset to classify sentiments. The training of the model aimed at giving it the ability to see and categorize sentiment in new, unlabeled data.

Using the Model on the ‘Qoslaaye’ Case

With the model trained, it was then used to classify sentiments specifically about social media data that relates to the ‘Qoslaaye’ case. This step gave us quantitative insights into the community’s sentiment about the case, showing whether the public reaction leaned towards positive, negative, or neutral sentiments.



The chart shows that neutral sentiments were the most common, suggesting a balanced or varied public response, followed by negative and then positive sentiments.

The sentiment analysis model's application to the 'Qoslaaye' case gives significant insights into the public sentiment, shown by the distribution of sentiments observed in the data. This task not only highlights the public's perception but also shows the potential of sentiment analysis in monitoring and understanding public opinion on sensitive societal issues.

Task 11

The goal of Task 11 is to make a machine learning model that can understand the feelings expressed in social media statements.

Getting Data Ready for Model Training

We split the dataset into feature and target sets, with 'ProcessedStatement' as the input x, and 'Sentiments' as the target y. We split the data into training and testing sets, using 80% of the data for training and the remaining 20% for testing to check the model.

Defining Machine Learning Steps

We set up a pipeline that combines a TfidfVectorizer for getting text features and a LogisticRegression classifier for understanding sentiment. The pipeline made the process of getting text features and understanding sentiment more efficient.

Training the Model

What We trained the machine learning model on the training set to learn the connection between the text data and the matching sentiment labels. The training made a model ready for testing and checking on new data.

Checking the Model

We checked how well the model works on the test set using accuracy as the measure. The model got an accuracy of 81.29%, showing a high level of performance on the test data.

Even though the accuracy is high, it's important to note that the complexity and variability of the Somali language, which doesn't follow a strict pattern, might limit how well the model works. This complexity could lead to potential inaccuracies in understanding sentiment for more nuanced or context-dependent expressions.

Task 12

The goal of Task 12 is to use the trained model to understand the sentiments of comments related to the 'Qoslaaye' case. This helps us see what the public thinks in a structured way.

Model Use Process and Checking

We used the model on the processed comments from the dataset about the ‘Qoslaaye’ case. We used the predict function of the pipeline to understand the sentiment of each comment.

The model gave sentiment predictions for each comment, putting them into ‘Wanaag’ (Positive), ‘Dhexdhexaad’ (Neutral), or ‘Xumaan’ (Negative).

Adding to the Dataset

We added the predicted sentiments to the original dataset, making a new column, ‘Predicted Sentiment’, next to the processed comments. This addition let us directly compare the comments and the model’s sentiment predictions, making it easier to review and analyze.

Checking Predicted Sentiments

We showed a subset of the original comments along with their predicted sentiments to check the model’s performance in a qualitative way.

```
                                ProcessedComment  \
0                tareenka qoslaaye aqoon shaqadiisa
1               qareenka qoslaaye shaqadiisa yaqaan
2  midib dal muslim dhulka katagaayo subxanallah ...
3  kiiskaan madmadow jira gacanta cida gaysatay c...
4           iimaanka qaad qoslaaye mahi dhaho hartay

                                ProcessedComment
0                tareenka qoslaaye aqoon shaqadiisa
1               qareenka qoslaaye shaqadiisa yaqaan
2  midib dal muslim dhulka katagaayo subxanallah ...
3  kiiskaan madmadow jira gacanta cida gaysatay c...
4           iimaanka qaad qoslaaye mahi dhaho hartay
```

The check gave a snapshot of how the model understands various comments and the sentiments it assigns, giving a sense of how well the model works.

Task 13

The goal of Task 13 is to count and show the distribution of predicted feelings for comments related to the 'Qoslaaye' case, shown as percentages of the total.

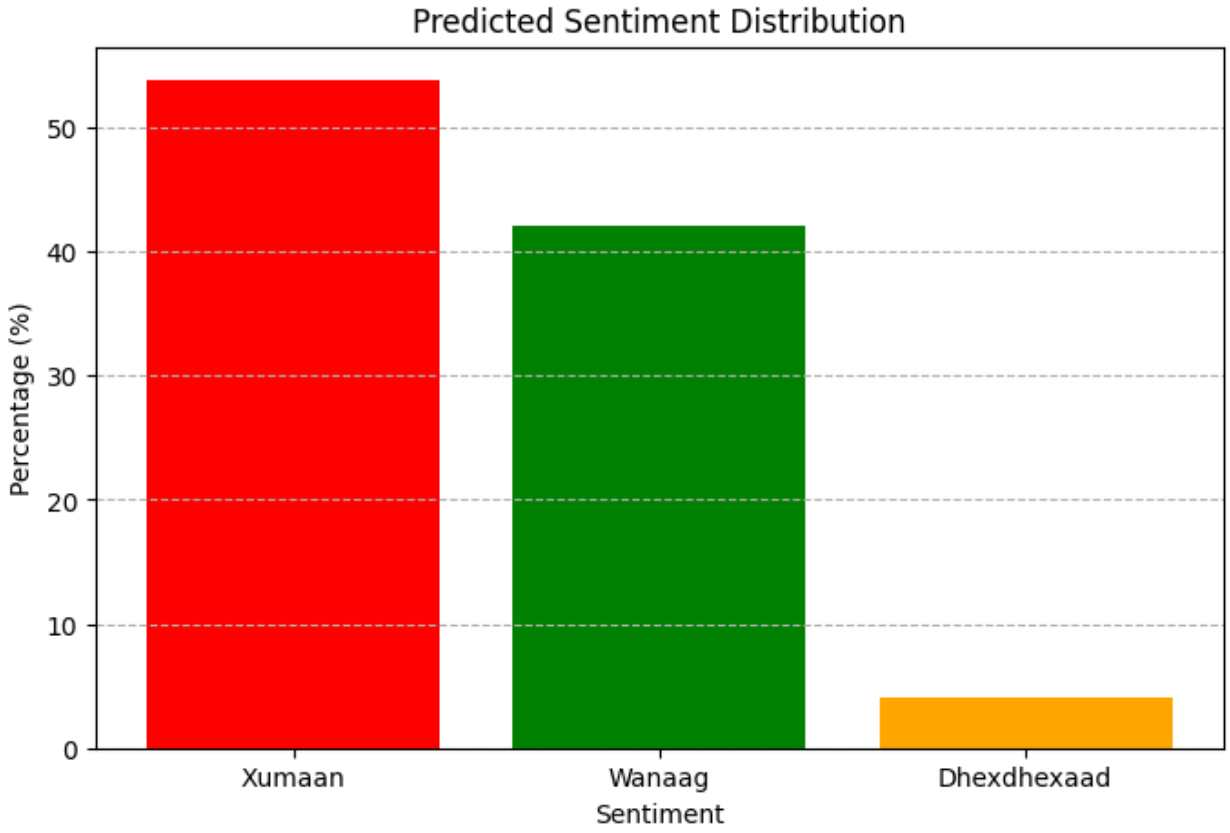
Calculating Sentiment Distribution

We calculated the distribution of predicted sentiments within the dataset to understand the proportion of each feeling category. We used the `value_counts(normalize=True)` method on the 'Predicted Sentiment' column to find the percentage distribution. We got the feeling distribution

showing 'Xumaan' (Negative) at 53.81%, 'Wanaag' (Positive) at 42.05%, and 'Dhexdhexaad' (Neutral) at 4.14%.

Creating Bar Chart

We made a bar chart to visually show the feeling distribution, using color coding to differentiate between feeling categories. We used `matplotlib` to plot the bar chart with color mapping: 'Wanaag' (green), 'Xumaan' (red), and 'Dhexdhexaad' (orange).



The chart shows each feeling as a percentage of the total, with different colors representing each sentiment category. The picture shows that negative feelings are the most common in the dataset, followed by positive sentiments, with neutral feelings being the least common.

The feeling distribution chart from Task 13 gives an informative snapshot of community sentiments, offering a visual summary that can quickly communicate the overall emotional tone of the discussion. It's an important part of the analysis, adding depth to our understanding of public reactions and feelings towards the 'Qoslaaye' case.

Task 14

The goal of Task 14 is to find and show the keywords most often associated with positive sentiments in the discussions about the ‘Qoslaaye’ case.

Choosing Positive Comments

We chose comments that the sentiment analysis model said were positive (‘Wanaag’).We got a group of comments predicted to express positive feelings ready for more analysis.

TF-IDF Vectorization

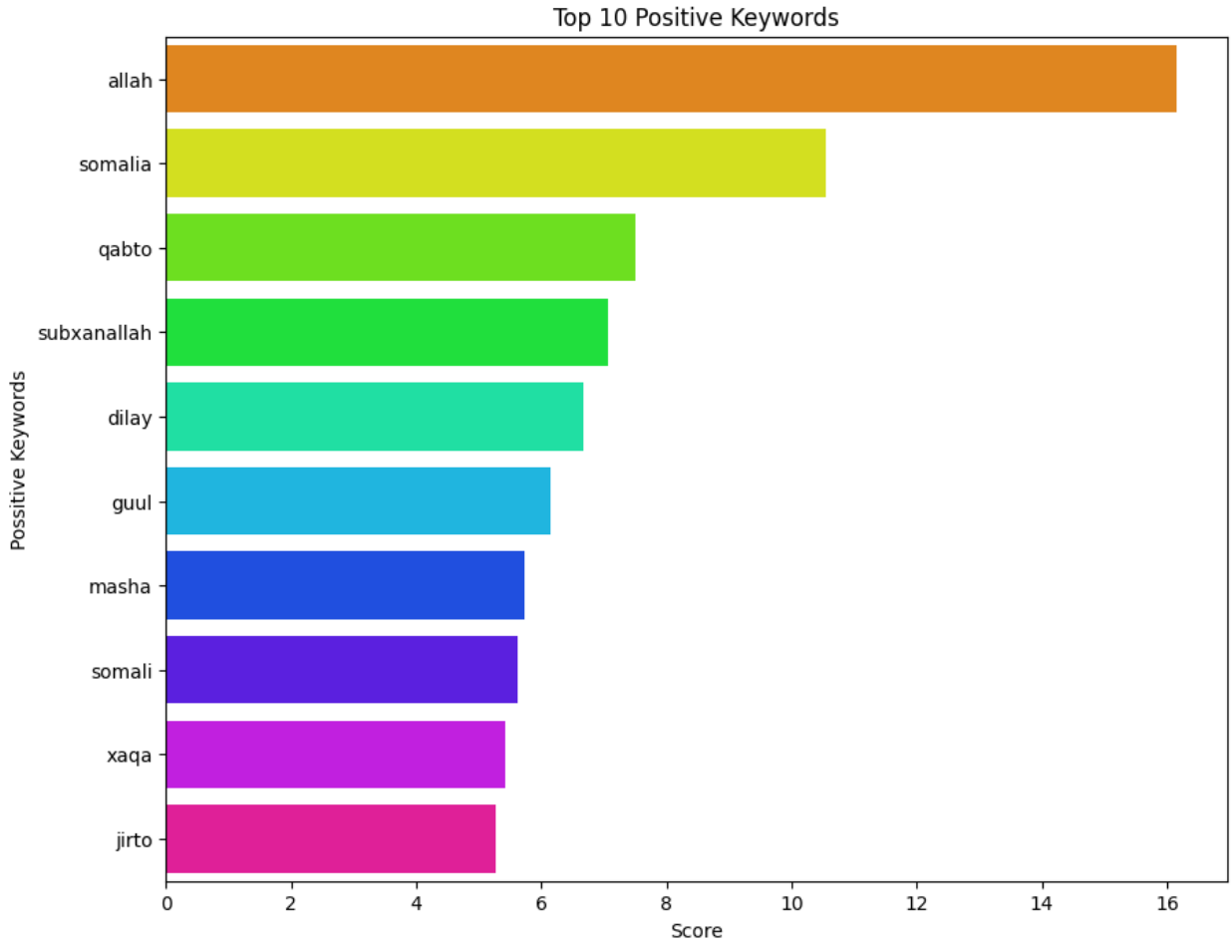
We used TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to change the text data into a number format, focusing on the most unique words in positive comments. We used TfidfVectorizer from sklearn, and the resulting TF-IDF matrix had scores showing the importance of each word in the group of positive comments.

Getting and Scoring Keywords

We got the feature names (words) and their matching TF-IDF scores to find the most relevant words in the positive feeling context. We gave scores to each word, showing its importance within the positive comments.

Sorting and Choosing Top Keywords

We sorted the words by their TF-IDF scores in descending order to choose the most important keywords. We made a list of the top 10 positive keywords based on their scores.



The picture (provided) highlights the words ‘allah’, ‘somalia’, ‘qabto’, ‘subxanallah’, ‘dilay’, ‘guul’, ‘mashallah’, ‘somali’, ‘xaqa’, and ‘jiirto’ as the most common within positively classified comments, suggesting themes of faith, nationalism, affirmation, and justice.

The bar chart gives a visual summary of the keywords that contribute most to positive feelings in the data related to the ‘Qoslaaye’ case.

The presence of these keywords in the discussion can guide further engagement strategies and provide insight into the elements of the case that foster a positive public response.

Task 15

The goal of Task 15 is to find and show the words most often associated with negative sentiments in the discussions about the 'Qoslaaye' case.

Choosing Negative Comments

We chose comments that were labeled as negative ('Xumaan') from the dataset to focus on the text that has this sentiment. We got a group of text ready for more analysis.

TF-IDF Vectorization for Negative Words

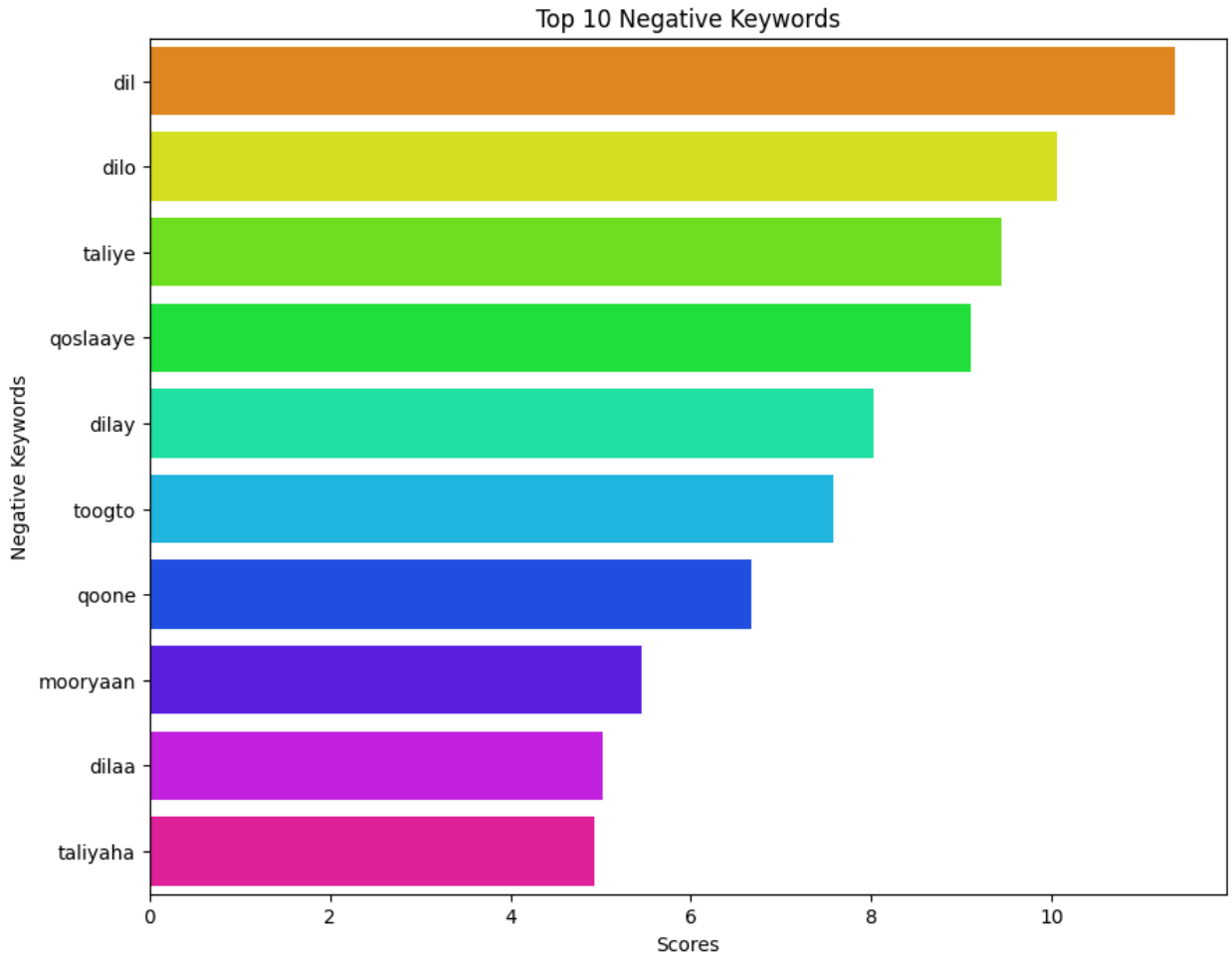
We used TF-IDF vectorization to change negative comments into a weighted number format, focusing on the most unique words. We used the `TfidfVectorizer` and got TF-IDF scores for words within the negative comments.

Scoring and Ranking Words

We got TF-IDF scores for each word and ranked them to find the most important words within the negative feeling context. We identified a list of words based on their TF-IDF scores, showing their importance.

Choosing Top Negative Words

We chose the top 10 words with the highest TF-IDF scores as the most important negative words. The choosing process highlighted the words most central to negative feelings expressed about the case.



The chart identifies words such as 'dil' (kill), 'dilo' (kill him), 'taliye' (captain), and 'qoslaaye' as the most important within negative statements, reflecting themes of violence, authority, and the specific subject of the case—Qoslaaye.

The picture effectively shows the words that dominate negative conversations, giving insight into the underlying themes that fuel unfavorable feelings. Understanding these negative words is essential for stakeholders to address the community's concerns and for any communicative efforts aimed at providing information or clarifying misunderstandings about the incident.

Conclusion

We collected, clean, and explore data, and we ended up with over 3,400 social media comments. We put these comments into feeling groups, which let us train a model to understand feelings in the Somali language.

The model worked well, with an accuracy of 81.29%, which is pretty good considering how complex language can be. But we know that there are challenges in understanding language, especially languages with lots of dialects and slang like Somali. Even though the model was accurate, we know that it might not fully understand all the subtle feelings because the Somali language is so rich and varied.

Our sentiment analysis showed a mix of reactions from the community. While there were a lot of neutral feelings ('Dhexdhexaad'), there were also a lot of negative feelings ('Xumaan'), which were more than the positive ones ('Wanaag'). This shows that the community has different feelings and thoughts about the events of the case.

Charts played a big role in telling the story of the data, from the sentiment distribution to the words associated with both positive and negative feelings. We found that discussions got more intense on certain days, which might line up with important events related to the case. This suggests that the public is responsive and involved with each development.

Negative words were mostly about harm and authority, showing the community's focus on the seriousness of the case and what it means. Positive words, though less common, showed a range of hope, justice, and national pride, showing some optimism and support in the discussion.

This project has not only given valuable insights into the public sentiment about the 'Qoslaaye' case but has also shown the potential of social media analytics as a powerful tool for understanding public opinion. Being able to go through lots of unstructured data and find meaningful patterns is really important for understanding how society reacts and guiding informed responses.

In summary, our findings give a detailed understanding of the Somali community’s sentiment in response to a big and troubling event. They show a society that is actively involved, emotionally invested, and vocal in its pursuit of justice and truth, as shown through social media interactions.

As social media continues to play a big role in public discussion, the methods and insights from this project will be a reference for future research. They highlight the importance of considering cultural and language contexts in feeling analysis and the need for continued development in machine learning models to better handle the complexities of human language and feeling.