Analysis of Social Response after a Natural Calamity using Text Mining and Sentiment Analysis

Diksha Jadhav
Data Science
University of Colorado Boulder
Boulder, Colorado, United States of America
dija4011@colorado.edu

Kalyani Kailas Jaware
Data Science
University of Colorado Boulder
Boulder, Colorado, United States of America
kaja7104@colorado.edu

Sharvari Ghatake
Data Science
University of Colorado Boulder
Boulder, Colorado, United States of America
shgh4210@colorado.edu

ABSTRACT

Social Media and News are the most widely used way of communication and expressing emotions during natural calamities and emergency situations. People share their views, emotions and requests for help in the form of text, images, videos and audios. This data is useful in understanding the situation and planning the disaster response. This study focuses on analyzing the sentiments and emotions from the content that is shared through YouTube comments and News articles during the Turkey-Syria Earthquake which occured in February 2023. Techniques such as sentiment classification, clustering, Association Rule Mining, Emotion Detection are used to analyze the data generated during the earthquake. By employing these techniques, this study aims to understand the predominant emotions and sentiments and find the association and patterns among the sentiments expressed in the public response. This research presents several analysis of the dataset which can be used by the response team in order to have better understanding of the public response and their needs during the natural calamity.

KEYWORDS

Sentiment Analysis, Association Rule Mining, Social Media, News, Emotion detection, Text Mining, Clustering, Principal Component Analysis.

ACM Reference Format:

Diksha Jadhav, Kalyani Kailas Jaware, and Sharvari Ghatake. 2018. Analysis of Social Response after a Natural Calamity using Text Mining and Sentiment Analysis. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

During serious life-threatening natural calamities, the public response is widely shared through news articles and social media. The devastating Turkey-Syria earthquake which occurred in February 2023, is one such natural calamity that captured global attention. During such emergency situations, authorities need information related to the event to understand the crisis and help the affected people. Social media and news are an effective ways to get such information. Decision support systems could benefit from the use of user-generated content as a source of big data for vulnerability assessment, early warning, monitoring, and evaluation of disaster risks. However, this data is very unstructured and huge. To understand and ingest the information available in this data, text mining and sentiment analysis is crucial.

The main objective of this research is to delve into the sentiments and emotions expressed by public through YouTube comments and news articles during the Turkey-Syria earthquake. This study explores polarity of sentiments using TextBlob in python, detects emotions using NRCLex. Additionally, an unsupervised machine learning algorithm, that is, K-Means clustering to group similar texts, providing a deeper understand of inherent patterns in the text. In order to uncover and understand patterns or association between the text, association rule mining is used. To enhance the understanding of the dataset further, exploratory data analysis, including the creation of Wordcloud and word frequency bar plots are used for the ease of visual interpretation of patterns in the textual data.

2 METHODOLOGY

2.1 Data Description

The data used in this research is a set of comments and news articles extracted using API. YouTube Data API[5], version 3 is used to extract around 400 comments from a video related to Turkey-Syria earthquake. In order to include variety in the dataset, two news APIs were used to extract news articles from several news websites. Guardian news API was employed to get around 100 news articles and world news API is used to get another 100 news articles. Guardian news API provides articles from it's own website only, whereas, the world news API extracts articles from several websites. The search query used to collect news articles is "turkey syria earthquake". The date used in the search query is 6th February 2023, which is the date on which the deadly earthquake struck several parts of Turkey and Syria.

2.2 Data Preprocessing

2.2.1 Data Cleaning. : In its untreated stage, if they are not preprocessed, newly published articles or YouTube comments are very disorganized and contain redundant features. To address these issues, such preprocessing removes stop words, hyperlinks, mentions, white spaces, emoticons, and unreadable characters. The main purpose of preprocessing is to eliminate noise from the dataset before

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

conducting additional analysis. To pre-process the dataset, we used the Natural Language Toolkit (NLTK)[6]. NLTK is a popular framework for Python programmers to use when developing applications with human language data. The preprocessing procedures comprised eliminating all unnecessary phrases and emoticon characters. As a consequence, the dataset was more organized and clean, ready for additional study. Table 1 provides an example tweet both before and after pre-processing.

2.2.2 Feature Vector. : Before we can feed categorical data—like text or words—to a machine learning model, it must first be converted into a numerical format. In this study, we employ a practical method for converting words into feature vectors: term frequency-inverse document frequency (tf-idf). The definition of the tf-idf is the product of the inverse document frequency and the term frequency:

$$tf - idf(t, d) = tf(t, d) * idf(t)$$

where tf(t, d) is the term frequency, which indicates the number of times that a term t occurs in a document d.

idf(t, d) is the inverse document frequency and can be calculated as follows:

$$\mathrm{idf}(t, d) = \log \frac{n_d}{1 + \mathrm{df}(t, d)}$$

where n_d is the total number of documents, and df(t, d) is the number of documents d that contain the term t. It should be noted that it is best to add the constant 1 to the denominator in order to give terms that appear in every training sample a non-zero value. Low document frequencies are not given undue weight due to the log.

2.3 Exploratory data analysis

Exploratory Data Analysis is an approach to analyze and summarize complex datasets. It helps in understanding the data and identifying patterns. Visualizations provide clear representation of the data and makes it easier to communicate the results from the analysis.

Wordclouds and Bar plots were created for both datasets to visualize the most common words.

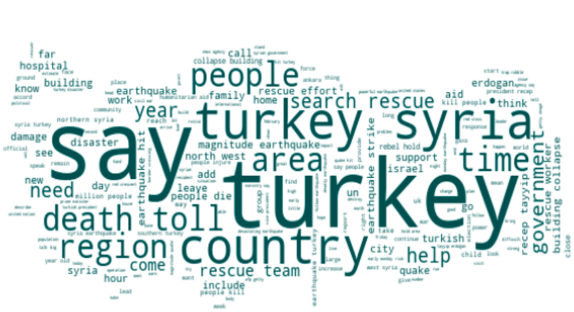

Figure 1: Wordcloud generated from words contained in the news articles

From figure 1, it can be observed that the news articles contain words related to the people in turkey and syria. Few words such as rescue and aid shows that the news covered the rescue operations and help provided to the affected population. The words such as earthquake, magnitude, time and death shows that the news articles also highlight the number of deaths, the magnitude of the earthquake, time when it occured, etc.

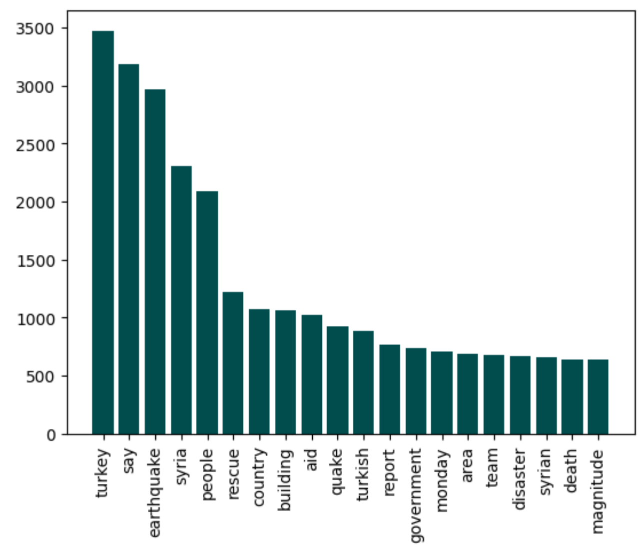

Figure 2: Bar plot of most frequent words from the news articles

Figure 2 shows a bar plot of 20 most frequently occuring words in the news article. It is observed that the words like people, rescue, government, report, country, etc were frequent which indicates that the news articles mentioned things related to resue and aid for public, magnitude and deaths caused by the earthquake and the role of government in these things.

Figure 3: Wordcloud generated from words contained in the YouTube comments

Figure 3 shows wordcloud obtained from YouTube comments. It is seen that words like help, God, pray and similar words in other languages are being used in the comments. These words depict that the affected people need help and they are praying to God and are hopeful about the situation.

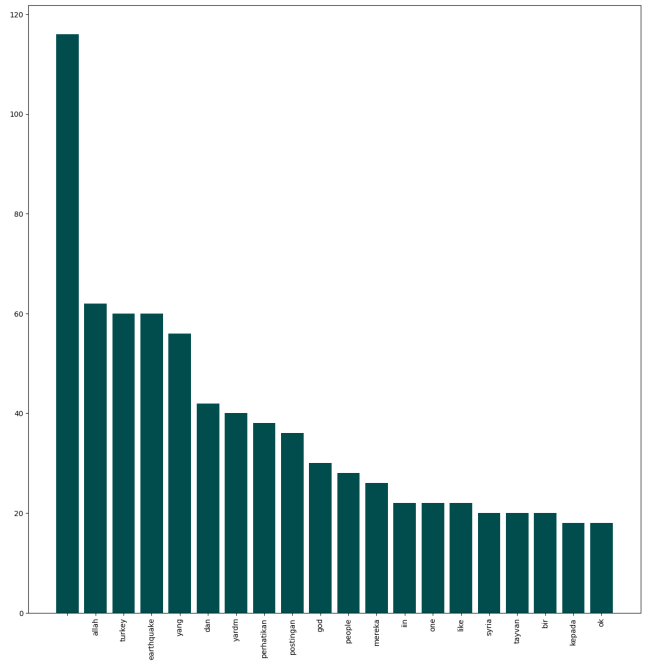

Figure 4: Bar plot of most frequent words from the YouTube comments

Figure 4. shows the 20 most frequent words used in the comments along with their frequency of occurrence. It seems like most of the comments were written in English alphabets but not in English language.

2.4 SENTIMENT ANALYSIS

Sentiment Analysis: Texts are categorized using sentiment analysis into three main groups: positive, negative, and neutral. Next, the extracted data is further categorized as positive, negative, or neutral based on its polarity. It is described as a procedural task to extract opinions' sentiments. While some opinions are sentiment-representative, others are sentimentless. Though subjective, the neutral texts have roughly equal amounts of positive and negative charges. These categories don't meet our needs for managing disasters because negative attitudes play a significant part. In particular, negative emotions like fear, panic, anxiety, and resentment need to be accurately identified as much as possible. For disaster relief, each of these feelings dictates a particular course of action. These sentiment and emotion analysis tools are highly useful for evaluating the effects of disasters and determining what people need in an emergency. The development of analytics-based automated tools for social media is beneficial for disaster relief and mitigation. Sentiment: Opinions, or views in a different sense, are defined as a person's verbal manifestation of their feelings, thoughts, assessments, and so forth. Analysis: To ascertain whether the opinions expressed by a group of users are neutral, positive, or negative [9]. Benefit: provide cost-effective information for decision-making. For this study, we used Textblob for Sentiment Reasoning to identify the emotions that were expressed in the News article and YouTube comments. Positive and negative emotion polarity as well as strong emotion intensity are detectable by a text sentiment analysis model built on textblob[5]. It is included in the NLTK package and can be used directly on unlabeled text data. Textblob's emotional analysis is predicated on a lexicon that translates lexical information into sentiment ratings, which gauge an emotion's intensity. The sentiment score of a text can be calculated by adding the intensities of each word. We discretized the sentiment scores from Textblob into three categories: positive, negative, and neutral in order to further analyze the results. This was done to facilitate the interpretation of the findings by non-technical stakeholders and to give a clear picture of the distribution of sentiments in the data. Scores higher than 0.1 fall into the positive category, and scores lower than -0.1 fall into the negative category. Neutral scores fall between -0.1 and 0.1. With the help of this classification, we are able to ascertain the prevailing opinion expressed in each YouTube and news article comment as well as explore the connections between various humanitarian classes and sentiment.

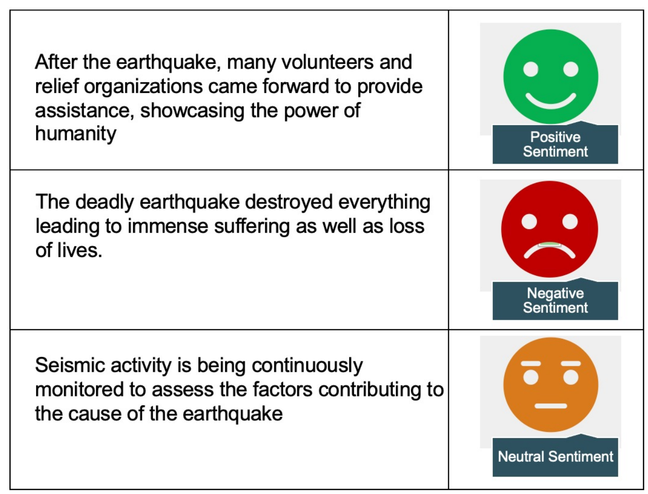

Figure 5: Sentiments Classification

2.5 K-Means Clustering

Among the popular unsupervised machine learning algorithms is the K-means algorithm. Generally, the algorithm determines discrete, non-overlapping clusters where each point is assigned to a group. With the minimum squared distance method, every point is assigned to the closest groups or subgroups. Determining the initial optimal centroids of clusters is one of the primary concerns of the K-means algorithm. Finding the ideal location for the centroids of the original clusters on the first iteration is the hardest task. In order to minimize the number of iterations and execution time, this paper suggests an efficient method for determining the ideal initial centroids. We have conducted experiments using various real-world datasets to evaluate the efficacy of our proposed method. To demonstrate the effectiveness of our suggested approach, we first examined datasets of news articles and YouTube videos. The Elbow method has been employed to ascertain the ideal number of clusters. Later on, we also included Principal component analysis, or PCA. This dimensionality reduction technique reduces the number of variables in a large data set by combining them into a smaller set that still retains the majority of the original data. Additionally, PCA was used in our project to reduce the data dimension to 2. We also employed the Silhouette Score, a quantitative indicator of how clearly defined and distinct the clusters are, to evaluate the suitability of the clustering results. The Silhouette Score measures a data point's uniqueness from other clusters and how well it fits into its designated cluster. It assists in determining whether the clusters are well-separated and internally homogeneous by measuring the cohesion and separation of data points within clusters. The Silhouette score in this case for News articles is 0.52 and YouTube comments data has come out to be 0.62

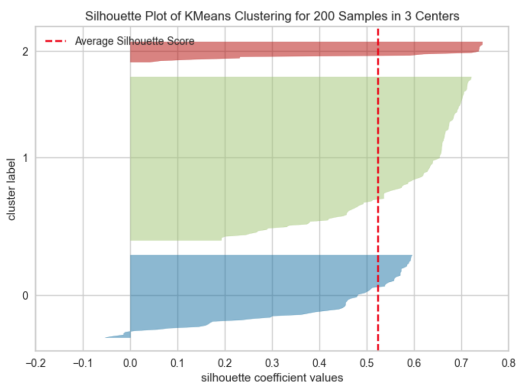

Figure 6: Silhouette Plot for News Articles

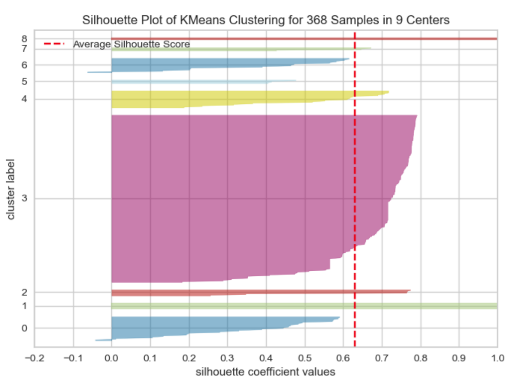

Figure 7: Silhouette Plot for YouTube Comments

In order to determine the optimal number of clusters for performing k-means clustering on our dataset, we used Elbow method. In Elblow method, within cluster sum of squares are calculated for a range of number of clusters. This value is plotted against the number of clusters and the number of clusters where the elbow point occurs, is considered as the optimal number of clusters. We used our vectorized data to calculate within cluster sum of squares values for 1 to 9 clusters for both datasets.

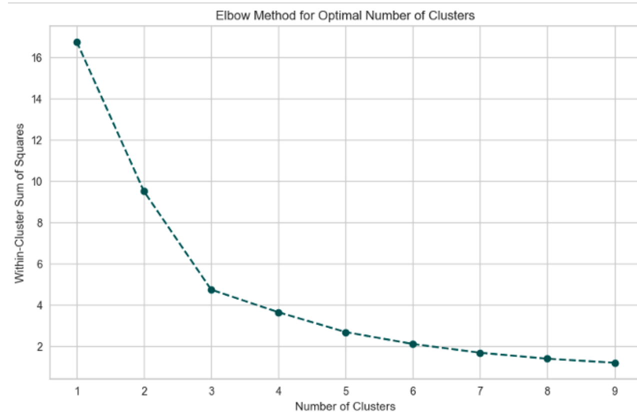

Figure 8: Scree Plot for News Articles

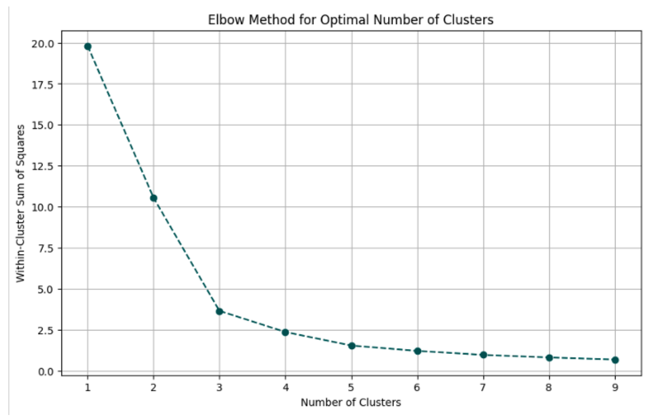

Figure 9: Scree Plot for YouTube Comments

Figure 8 and 9 shows that the elbow point in both dataset occurs at 3. Hence, the optimal number of clusters used in this research for performing k-means clustering is 3.

2.6 Association rule mining

ARM is a data mining technique used to identify interesting patterns, correlations and relationships in the dataset. The process involves identifying frequent itemsets and creating rules to find association between the frequent itemsets using parameters such as support, confidence and lift.

Support - The percentage of transactions containing an itemset.

$$Support = \frac{frequency(itemset)}{Numberoftransactions}$$

Confidence - Suppose, if there are two itemsets then confidence is conditional probability of one itemset given another itemset

$$Confidence = \frac{Sup(A \cup B)}{Sup(A)}$$

Lift - It is a measure of strength of association rule. It is also used to observe randomness by comparing combined and independent probabilities of itemsets.

$$Lift = \frac{Sup(A, B)}{Sup(A) * Sup(B)}$$

Apriori algorithm is one of the most widely used algorithm to perform association rule mining. In this research, apriori algorithm in association rule mining is used to find frequent patterns associated with the textual data. Association rules were derived by setting the minimum support level at 7% and minimum threshold for lift at 1. If the lift value is close to 1, it indicates that the frequent itemsets in the association rules occur together more frequently. Lift value is directly proportional to association. The rules shows association between frequently used words in the news articles and youtube comments. Python library "mlxtend" was used to employ the algorithm on the dataset. A network graph is plotted to visualize and interpret the results obtained from the analysis.

2.7 Emotion Detection

For emotion detection technique, we used NRCLex[1] library from python. The package in python is built on approximately 27,000 words and is also based on the National Research Council Canada (NRC) affect lexicon and NLTK library's WordNet synonym sets. It was used to detect 5 major emotions which were fear, anger, sadness, disgust and joy in our dataset. This aids in understanding the variety of emotiions expressed in the public responses over social media and news.

3 RESULTS

3.1 Sentiment classification

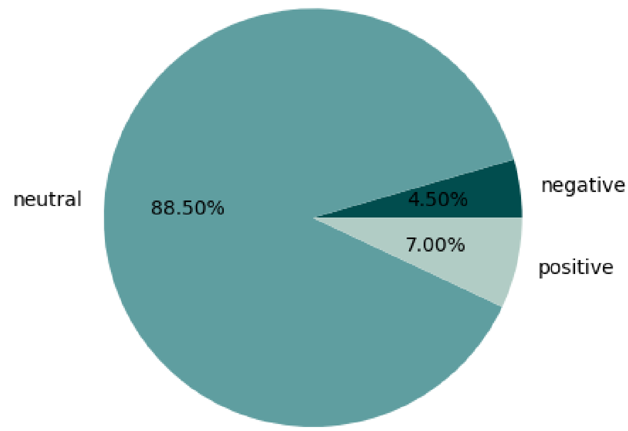

Figure 10: Sentiment classification of news articles

Figure 10 shows the proportion of positive, negative and neutral sentiments associated with the news articles. 88.5% news articles were classified as neutral because the language used in news article is usually formal and doesn't include any words that depict strong sentiments. There are 7% articles depicting positive sentiments. These could be related to the rescue operations or help provided by government or response teams. 4.5% of articles depicted negative sentiment. These could be the articles talking about deaths and sufferings of the affected public.

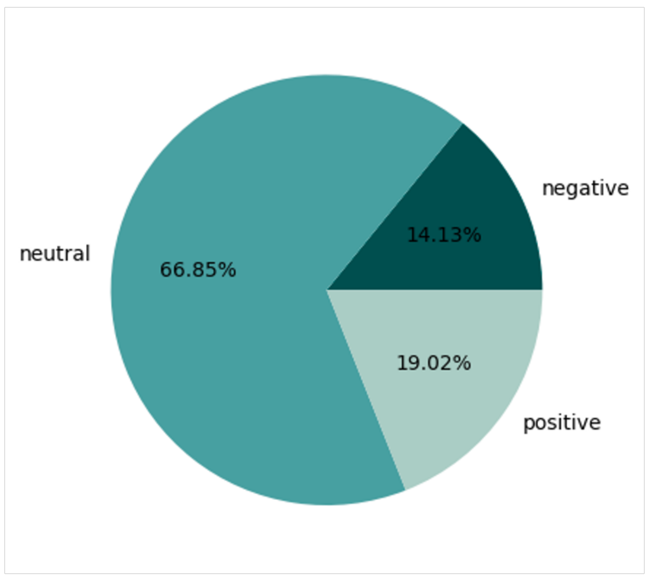

Figure 11: Sentiment classification of YouTube comments

Figure 11 shows the proportion of positive, negative and neutral sentiments associated with the YouTube comments. 68.85% news articles were classified as neutral. There are 19.02% articles depicting positive sentiments. These could be because of the hopeful and praying comments made by the people. 14.13% of articles depicted negative sentiment. These could be the comments talking about loss, damages and deaths.

3.2 K-Means Clustering

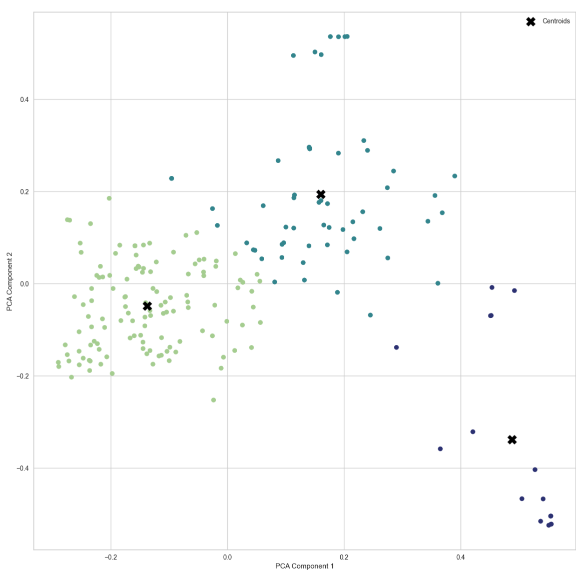

Figure 12: Scatterplot of clusters in news articles

Figure 12 shows the clusters obtained by performing k-means clustering on new articles. Dimensionality reduction using PCA helped to visualize the clusters better on a 2 dimensional axis. There are 3 distinct clusters well separated from each other in the figure. Cluster 1 has highest number of articles and cluster 3 has lowest.

Table 1: 20 most frequent words in each cluster obtained from news articles

Cluster	20 most frequent words
Cluster 0	provide,assistance,emergency,offer,red, personnel,help,military,firefighter,medical, equipment,aid,search,syria,say,dog,team, rescue,turkey,send
Cluster 1	city,collapse,turkish,area,toll,injure, country,report,syrian,death,aid,monday, building,quake,rescue,people, earthquake,syria,say,turkey
Cluster 2	time,2023,city,rubble,rescue,year,area, government,disaster,turkish,quake,country, magnitude,aid,building,people,syria,say, turkey,earthquake

To further analyse the clustering results, top 20 words from each clusters were derived. Table 1 shows those words along with their cluster. Words from cluster 0 represent the articles that were related to rescue teams, aids, providing assisstance during emergency and how military, firefighters and medical personnel are helping. Words from cluster 1 are more inclined towards the damages and loss such as deaths, injuries, building collapses, etc. Words form cluster 3 are about the time and magnitude of the earthquake.

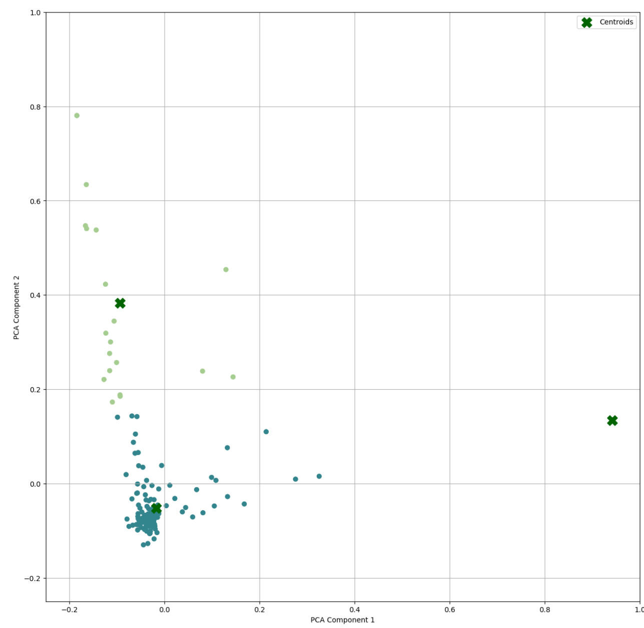

Figure 13: Scatterplot of clusters in YouTube comments

Figure 13 shows the clusters obtained by performing k-means clustering on YouTube comments. 3 distinct clusters are observed but most of the comments are contained in one cluster. One of the clusters has all it's data points at the coordinates of it's centeroid.

Table 2: 20 most frequent words in each cluster obtained from YouTube comments

Cluster	20 most frequent words
Cluster 0	not,people,my,to,one,are,for,it,that,its,earthquake, was,and,god,of,allah,in,is,this,the
Cluster 1	gembira,gelitirilmesini,gelen,geldiini,friday, geici,gaziantep,gave,g7,futuristico,future,further, from,friends,gc,zorluklarna,to,all,family,condolence
Cluster 2	all,al,country,bad,aras,my,fatihah,am,to,prayers, stay,this,is,from,and,for,rip,pray,syria,turkey

For further analysis of clustering results, top 20 words from each cluster were obtained. Words from cluster 0 are about god and people majorly. Cluster 2 contains words that are not in English language but written in English alphabets while commenting. Words from cluster 3 are about people praying for the betterment of the situation in comments.

3.3 Association rule mining

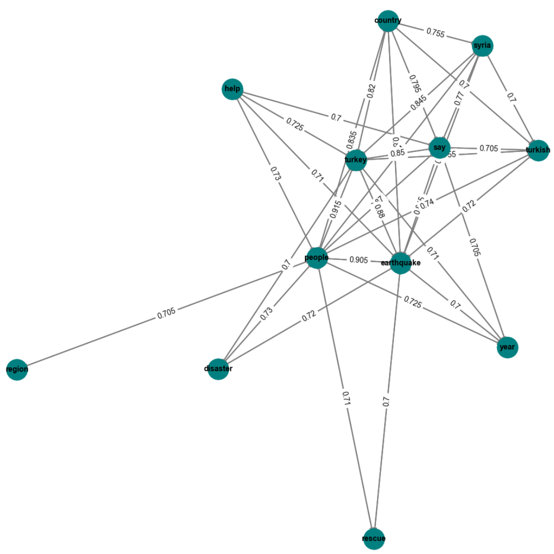

Figure 14: Network graph for association of words in news articles

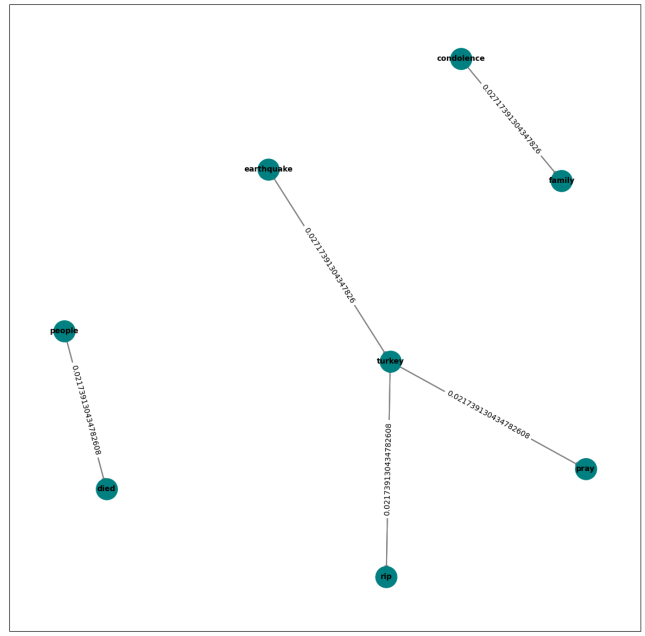

Figure 15: Network graph for association of words in YouTube comments

Top 70 association rules along with their support values are presented in the network graph in figure 14. It shows that highest support is observed between people and turkey, that means in most of the news articles, the words turkey and people occured. There is a strong association between the words shown in the graph.

In case of YouTube comments, the results of association rule mining are not very significant as there are very few association rules obtained. Also, their support values are extremely low. This means that the combinations of words used in every comment were different.

3.4 Emotion detection

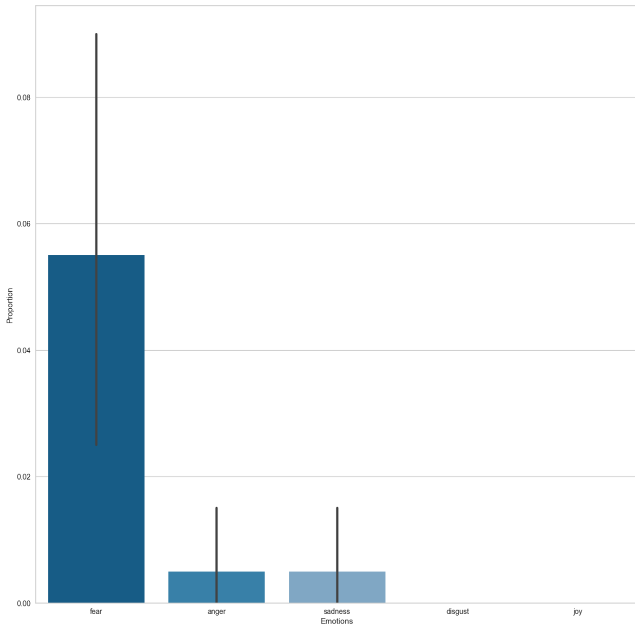

Figure 16: Bar plot of proportion of emotions in news articles

From the figure 16 graph, it is very much evident that news articles had more data which was inclined towards emotions like fear, anger and sadness which are natural in incidents like these.

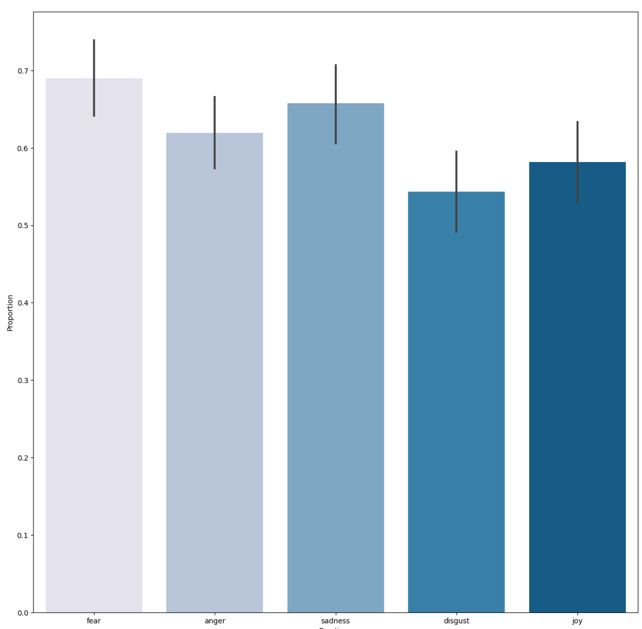

Figure 17: Bar plot of proportion of emotions in news articles

Figure 17 shows emotions of youtube comments data, which represents the comments which has fear, anger and sadness there were comments which also belonged to categories like disgust and joy which is very unlikely in such situations proving the irrelevance of the data.

4 CONCLUSION

There are many potential applications where social media can be useful during a disaster or an emergency situation. In this study, we analyzed data obtained from social media and news related to Turkey-Syria Earthquake.Techniques such as text mining, sentiment analysis, clustering, association rule mining and emotion detection were integrated. This research provides insights such as proportion of positive, negative and neutral news articles as well as YouTube comments, proportion of emotions such as fear, joy, disgust, anger and sadness in the data, grouping of similar comments and articles and finding association within the words in the dataset. These findings can be used by decision support systems to make data-driven decisions as well as to get a thorough understanding of public reactions. By utilizing these methods, we can improve early warning systems, vulnerability assessment, and disaster risk management in general, leading to more adaptable and efficient crisis management plans.

5 LIMITATIONS

One of the major limitation in our research is detection and analysis on languages other than English found in our dataset. These words could not be removed using the python libraries that we used for preprocessing the data because they were written in English alphabets. Another limitation was size of our dataset. Since, this research was focused on a past natural calamity, we were not able to find relevant articles and youtube videos. Due to limitations in news API, we were able to extract only 200 articles.

6 FUTURE SCOPE

The four main goals of this research are to use supervised machine learning techniques to advance the field of sentiment analysis. Initially, we will develop sentiment classification models that are optimized, with benchmarked accuracy objectives across datasets and domains. Second, in order to enhance contextual understanding, we will create multimodal sentiment analysis techniques that incorporate textual data with pictures, videos, and audio. Third, by selecting non-English datasets and developing classifiers in the desired languages, we will improve language support to handle sentiments other than those expressed in English. Fourth, we will make use of transformer-based neural architectures, which have proven to be highly effective at identifying complex semantic patterns in tasks involving text analysis. Overall, this work aims to advance the state-of-the-art accuracy in sentiment analysis by extending language support, utilizing supervised learning strategies integrated into multimodal frameworks, architectures based on transformation that are intended to understand rich emotional expression. Among the many facets of this research, achieving accuracy benchmarks and applying these advancements to practical applications continue to be the main objectives.

ACKNOWLEDGMENTS

We would like to thank Siddharth Tayi, course assisstant for Data mining course, for guiding and helping us in completing our project.

REFERENCES

[1] Muhammad Awais and Sara Durani, 2023. Sentiment and Emotion Analysis on Consumer Review using NRCLex. *International Conference on Engineering, Natural and Social Sciences 1.* (Jan, 2023). DOI: https://as-proceeding.com/index.php/icensos/article/view/51.

[2] Cagri Toraman, Izzet Emre Kucukkaya, Oguzhan Ozcelik and Umitcan Sahin. 2023. Tweets Under the Rubble: Detection of Messages Calling for Help in Earthquake Disaster. *ArXiv, abs/2302.13403.*DOI: https://doi.org/10.48550/arXiv.2302.13403. .

[3] Alam, F., Ofli, F., Imran, M., and Aupetit, M. 2018. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. *ArXiv, abs/1805.05144.*

[4] Suvrat Arora. 2023.Sentiment Analysis Using Python. *Analytics Vidhya Article.*DOI: https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/.

[5] Natassha Selvaraj. 2022. How to Perform Sentiment Analysis with Python?. *365 Data Science Article.* (Aug, 2022). DOI: https://365datascience.com/tutorials/python-tutorials/sentiment-analysis-with-python/.

[6] Abdeladim Fadheli. 2022. How to Extract YouTube Data using YouTube API in Python. *The python code Article.* (Jun, 2022). DOI: https://thepythoncode.com/article/using-youtube-api-in-python/

[7] Pedregosa, F. , Varoquaux, G. , Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. , Prettenhofer, P., Weiss, R. , Dubourg, V. , Vanderplas, J. , Passos, A. , Cournapeau, D. , Brucher, M. , Perrot, M. and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 12. DOI: https://scikit-learn.org/stable/about.html#citing-scikit-learn

[8] Lucas de Sá. 2019. Text Clustering with K-Means. *Medium Article.* DOI: https://medium.com/@lucasdesa/text-clustering-with-k-means-a039d84a941b

[9] Hossain, Md. Murad and Amin, Muhammad Saad and Khairunnasa, Fatema and Hussain, Syed Tahir, Twitter-Based Sentiment Analysis of Emotional Response to the Earthquake in Turkey and Syria. DOI: http://dx.doi.org/10.2139/ssrn.4625551

[10] Myneni Madhu Bala, K. Navya and P. Shruthilaya. Text mining on real time twitter data for disaster response. *International Journal of Civil Engineering and Technology (IJCIET)*Volume 8, Issue 8, August 2017. DOI:http://http://iaeme.com/Home/issue/IJCIET?Volume=8&Issue=8

[11] Mohammadsepehr Karimiziarani and Hamid Moradkhani. Social response and Disaster management: Insights from twitter data Assimilation on Hurricane Ian. *International Journal of Disaster Risk Reduction.* Volume 95. 2023.DOI:https://doi.org/10.1016/j.ijdrr.2023.103865.

[12] Dong, Z.S., Meng, L., Christenson and L. et al. Social media information sharing for natural disaster response. *Nat Hazards 107, 2077–2104.* 2021. https://doi.org/10.1007/s11069-021-04528-9.

[13] Hassan, Syed Zohaib, Kashif Ahmad, Steven Hicks, Pål Halvorsen, Ala Al-Fuqaha, Nicola Conci, and Michael Riegler. 2022. *Visual Sentiment Analysis from Disaster Images in Social Media* Sensors 22, no. 10: 3628. DOI: https://doi.org/10.3390/s22103628.

[14] Beigi, G., Hu, X., Maciejewski, R. and Liu, H. 2016. An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. *Sentiment Analysis and Ontology Engineering. Studies in Computational Intelligence.* vol 639. DOI: https://doi.org/10.1007/978-3-319-30319-2_13

[15] Li, W., Haunert, J-H., Knechtel, J., Zhu, J., Zhu, Q., and Dehbi, Y. 2023. Social media insights on public perception and sentiment during and after disasters: The European floods in 2021 as a case study. *Transactions in GIS.*DOI: https://doi.org/10.1111/tgis.13097.

Presentation Video link: https://www.youtube.com/watch?v=plZsNQDks-o

Github link: https://github.com/kalyani-jaware/Data-Mining-Project