# Sentiment Analysis of Twitter: Turkey Earthquake 2023 Case

Ala Kamal Rashid[1, †], and Oguz Fındık[2, *, †]

[1] *University of Karabuk, Computer Engineering, Karabuk, Turkey*
[2] *University of Karabuk, Computer Engineering, Karabuk, Turkey*

### Abstract

The most devastating earthquake in the past 20 years was February 6, 2023. The earthquake occurred in southern Turkey near the northern Syrian border. Thousands of people died and many more were left homeless, due to the magnitude of the event, it quickly spread all over the world. The earthquake and its damage were discussed and analyzed from all sides.

In this paper, a separate analysis was proposed for tweets posted within 14 days after the earthquake. In this analysis to classify tweets, one type of label did not depend as in previous works that have been done on text classification, but three different types of labels (Manual label, NLTK_VADER label, and Cluster label) are created to classify text tweets by using machine learning algorithms. Then by using the Jaccard similarity coefficient and the cosine similarity measure the two AI labels (NLTK_VADER and Cluster) are compared which result is closer to manual labeling, according to the number of categories (positive, negative, and natural) and accuracy of sentiment in each label.

In the result, we have reached that the accuracy of the VADER labeling is more effective than Cluster labeling because its accuracy is much closer to the Manual labeling.

### Keywords

Turkey Earthquake, Text classification, Machine learning, NLTK VADER, and Cluster.

## 1. Introduction

Sentiment analysis is a technique used to determine the emotional tone or sentiment expressed in a text. It involves analyzing the words and phrases used in the text to identify the underlying sentiment, whether it is positive, negative, or neutral, and has a wide range of applications, such as social media monitoring, customer feedback analysis, and market research [1].

The various research works in sentiment analysis (Özgür Ağralı et al. 2023) presented an article “Twitter Data Analysis: Izmir Earthquake Case”, NLP is used for sentiment analysis and topic modeling[2].

*SMARTINDUSTRY-2024: International Conference on Smart Automation & Robotics for Future Industry, April 18 - 20, 2024, Lviv, Ukraine*

[*] Corresponding author.

[†] These authors contributed equally.

alakamal991@gmail.com (A. K. Rashid); oguzfindik@karabuk.edu.tr (O. Fındık).

0009-0007-7292-5116 (A. K. Rashid), 0000-0001-5069-6470 (O. Fındık).

(Ayşe Berika et al. 2022) In this overview “Comparison of Different Heuristics Integrated with Neural Networks: A Case Study for Earthquake Damage Estimation”, Various Machine Learning (ML) algorithms were compared on a public dataset of earthquakes [3].

(Sean Wilkinson et al. 2022) in the article “Accuracy of a Pre-trained Sentiment Analysis (SA) Classification Model on Tweets Related to Emergency Response and Early Recovery Assessment: The Case of the 2019 Albanian Earthquake” supervised tweets that are classified as either positive, negative, or neutral for comparison with the unsupervised classification [4].

(Asif Malik et al. 2019) This study “Lexicon-Based Sentiment Comparison of iPhone and Android Tweets During the Iran-Iraq Earthquake” quantified the observed sentiment difference between the Android and iPhone tweets using unsupervised classification utilizing a lexicon-based approach [5].

(Cagri Toraman et al. 2023) this paper “Tweets Under the Rubble: Detection of Messages Calling for Help in Earthquake Disaster” Classifies the tweets calling for help or not and visualizes them in an interactive map screen [6]. (Yufei Xie et al. 2023) this study explores the use of CNNs for sentiment analysis on data from Weibo. to investigate this method's effectiveness in the context of NLP tasks and evaluate any possible ramifications [7].

On 6 February 2023, a Mw 7.8 earthquake struck southern and central Turkey, and northern and western Syria. The epicenter was Gaziantep, the largest seismic event in Turkey since 1939 [8]. The devastating earthquake caused heavy damage and many residents were killed and injured under the collapsed buildings.

Social media plays an important role during events, and it is used as a trusted source in many areas, especially Twitter, which is currently the most accurate source among various social networks. Twitter is one of the most vibrant and widespread resources within social media [9], mostly used by academics. Google Scholar lists 27,000 research articles that include the word Twitter in their title [10].

In this research, we used a dataset of 28,000 tweets that express people's feelings during the Turkey earthquake in 2023, available on Kaggle. To analyze the tweet, three different types of labels were created (Manual label, VADER label, and Cluster label), and they were classified by using machine learning algorithms such as (Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree). This study aims to analyze tweets posted during and after the earthquake to indicate which had positive and negative sentiments among citizens and to determine the accuracy of labeling types by indicating which is much closer to manual labeling accuracy.

## 2. Methodology

The best social media dataset for text classification is Twitter [11]. Our dataset consists of 28,000 tweets from Twitter API about the Turkey Earthquake between "2023-02-07 / 2023-02-21". This collected dataset is available from the Kaggle website [12]. It contains 16 columns such as (‘id’, ‘username’, ‘user location’, ‘user description’, ‘date’, ‘text’, ‘hashtags’, ‘source’, ‘retweets’, and so on), and 28,000 rows.

In this study, we used 10% of the dataset, which is 2,800 tweets, and worked on text fields within 3 sections which are Text Pre-processing, Text Labeling, and Text Classifications.

## 2.1. Text Preprocessing

The text of tweets, which are vague data because they are normal people's speech and full of strange words, emojis, hashtags, etc. These texts need to be cleaned up and the meaningless words removed by several processes [13], Here these steps were performed:

Cleaning (Remove Special Characters and Numbers, Convert to Lowercase). Tokenization splits and breaks down the sentences into individual words. Stop Word Removal removes common words like 'the', 'and', 'or' etc. that may not have important meanings and are not considered keywords. Lemmatization returns the words to the original root or the source of the word like "running," and "runs," to the common stem "run".

At the last step in preprocessing, by using the function (get-word) extracts words from text using a regular expression pattern and fills non-values in a specific column ('remove shorts') with an empty string. The text column is cleaned of all unnecessary phrases, emoji, and words, for example, this tweet (Prayers for Türkiye and Syria 🙏 Hope the rescue...) after preprocessing converted to (prayers trkiye syria hope rescue teams from va...).

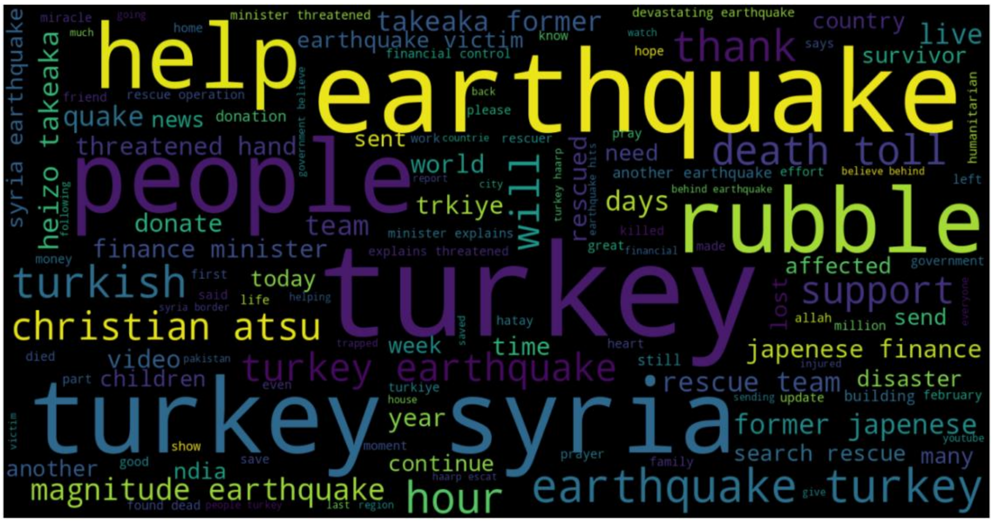

**Figure 1:** Showed The common words used in their tweets by users were examined like (earthquake, Turkey, Syria, people, and help).

## 2.2. Text Labeling

Text labeling is the process of identifying raw texts and adding one or more meaningful labels to provide context so that a machine learning model can learn from it [14]. Labeling is typically done according to several guidelines defined for text labeling. There are several different types of labeling, and the most common types are done manually by human annotators or through automated methods. In this work, 3 types of labeling were created for the text field such as (Manual labeling, VADER labeling, and cluster labeling), for each labeling types have 3 different categories (positive, negative, and natural).

### 2.2.1. Manual labeling

It is assigned by human annotators or experts based on their domain knowledge or specific guidelines and is typically used in supervised learning settings, where the goal is to train a model to predict or classify unseen data based on labeled examples. Manual annotation can provide more accurate and meaningful categorizations compared to other labels, especially when the true underlying structure of the data is known or can be reliably determined [15] but it requires a lot of time and human expertise. Here read and analyzed the tweets carefully according to our experience and with the help of the positive and negative phrases used in the texts we have decided which are positive, negative, and natural.

### 2.2.2. VADER labeling

(Valence Aware Dictionary and Sentiment Reasoner) is one of the greatest options for sentiment analysis in Python, a pre-built library in NLTK that is based on lexicon and rule. This package was created specifically for sentiment analysis on social media [16]. Text sentiment is calculated by VADER, which also provides the probability that a given input sentence is positive, negative, or neutral. The measurement that the library provides is called a compound score, or polarity score. It is the sum of all normalized lexical evaluations between -1 (negative) and +1 (positive). In this study, tweets were categorized according to polarity scores as positive emotion (polarity score > 0), negative emotion (polarity score < 0), and natural emotions (polarity score = 0).

### 2.2.3. Cluster labeling

It is assigned through unsupervised learning techniques typically clustering algorithms such as K-means, hierarchical clustering, or DBSCAN. These labels are derived solely from the data's intrinsic structure without any external guidance or supervision. Each data point is assigned to a cluster based on its similarity or proximity to other data points within the same cluster. Cluster labels are useful for discovering patterns or groupings in the data when the true categories or classes are unknown or not provided [17].

Here (convert texts to numerical format by (TF-IDF, Term Frequency-Inverse Document Frequency), K-Means clustering to group similar documents and used (The elbow Method and Silhouette Score) to determine the optimal number of clusters (K), Applied Principal Component Analysis (PCA) to reduce the dimensionality of the TF-IDF data to 2D components for visualization, Evaluated the quality of clusters using the Silhouette Score and Davies-Bouldin Index) performed, those steps showed in figure 2.

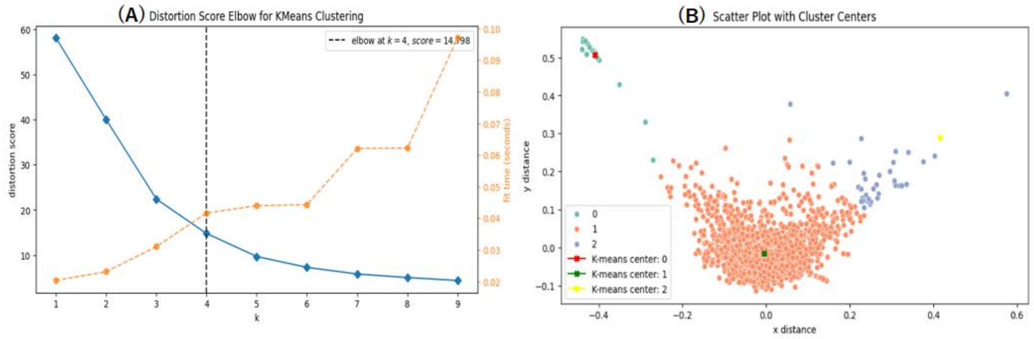

**Figure 2:** part(A) Uses k-means to determine the number of clusters based on the Elbow method and displays the optimal number of clusters was 4 corresponding to the number of text types in the dataset, part(B) re-implementation of K-means clustering based on the optimal number of clusters given.

## 2.3. Text Classification

Classifications are often using three categories to classify sentiment: negative, neutral, and positive. It is still possible that these categories do not reflect the real world [18]. Therefore, several algorithms have been developed to make predictions more accurate in obtaining results such as ML or deep learning etc.

### 2.3.1. Approaches

Machine learning techniques were implemented to classify text tweets with all three labels such as (logistic regression, support vector machine (SVM), K nearest neighbor (KNN), and decision tree) and used accuracy measures to determine which model instances were correctly classified across all classes.

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives

### 2.3.2. Data splitting

Data splitting involves using some of the data for model modification and setting aside the remainder as an assessment set or Training samples [19]. The dataset will be split into an “80:20 ratio” (80% training and 20% testing), using (the stratified random sampling) method. Table_1 shows the splitting dataset.

**Table 1.**
The total data was 2800, after preprocessing the total data will be 2756 divided over 2 parts, one is a train set of 2204, and the other one is a testing set of 552.

| **Train set** | **Total** | **Test set** | **Total** |
|---|---|---|---|
| Total data in the train set: | 2204 | Total data in the test set: | 552 |
| Total text data in the train set: | 2204 | Total text data in the test set: | 552 |
| Total column in train set: | 2 | Total column in test set: | 2 |

## 3. Comparative

In this section, labeling (VADER and cluster) is compared with manual labeling, according to the number of emotions (positive, negative, and natural) in each label, and then the accuracy of the classification models is compared.

### 3.1. Sentiment number comparison

After creating all three labels, the categories are counted according to function (value counts), the results are shown in Table 2 and the Sentiment plot in Figure 3.

**Table 2.**

Category Counts

| **Labeling** | VADER | Manual | Cluster |
|---|---|---|---|
| **Categories** | | | |
| Positive | 1219 | 1462 | 2647 |
| Negative | 888 | 667 | 67 |
| Natural | 649 | 626 | 42 |

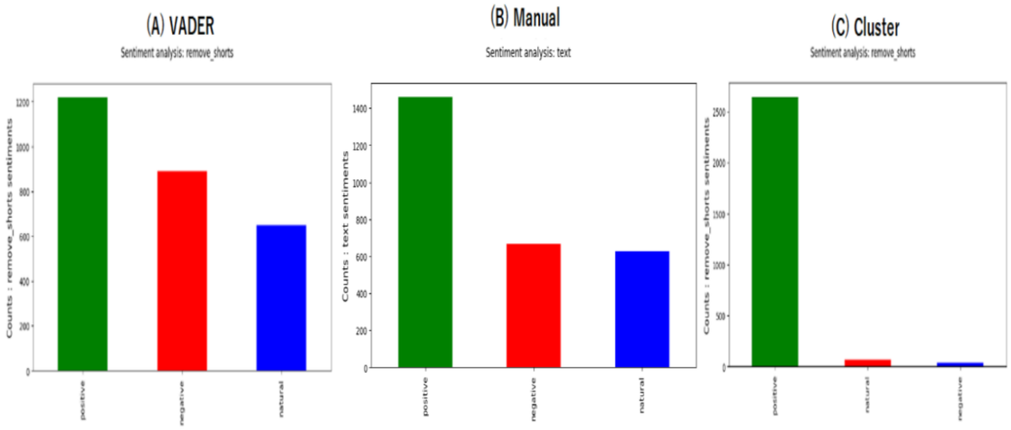

**Figure 3:** Compare Sentiment Labeling Plots

In Table 2. And Figure 3. The distribution of sentiment labels varies across the datasets, indicating potential differences in the labeling criteria. In total data (2756) Cluster labeling has the highest count of positive sentiment (2647), and it has the lowest count of negative sentiment and natural sentiment, the proportion of positive sentiment is generally higher than negative sentiment in all datasets.

### 3.2. Model Accuracy Comparison

machine learning classification algorithms such as (Logistic Regression, Support Vector Machine (SVM), K-nearest neighbor (KNN), and Decision Tree) trained with each label to compare the accuracy results of (VADER and Cluster) labels which one has closer accuracy with the manual label, this comparison is shown in Table_3.

**Table 2.**
Accuracy comparing models and labels.

| Model | Cluster labeling | Manual labeling | VADER labeling |
|---|---|---|---|
| Support Vector Machine | 97.95 | 67.02 | 70.47 |
| Decision Tree | 96.93 | 58.51 | 67.93 |
| K-Nearest Neighbor | 98.97 | 62.68 | 58.87 |
| Logistic Regression | 97.95 | 67.93 | 71.92 |

As in Table 3. for the Cluster labels, the accuracy achieved by various models such as K-Nearest Neighbor, Logistic Regression, Support Vector Machine, and Decision Tree, ranges from around 96.93% to 98.97%. For manual labels, the accuracy was achieved from around 58.51% to 67.93%. VADER Sentiment labels, the accuracy achieved ranges from around 58.87% to 71.92%.

The models trained on Cluster labels generally exhibit higher accuracies compared to those trained on Manual labels and VADER Sentiment labels. This suggests that the clustering algorithms might have captured underlying patterns in the data more effectively.

Manual labels are typically assigned by human annotators based on domain knowledge or specific guidelines, making them more interpretable and possibly more reliable in certain contexts. VADER Sentiment labels are derived from sentiment analysis techniques and may capture sentiment-related information in the text but might not necessarily align with manual annotations.

## 4. Results

These measures quantify the similarity between two sets of labels based on their intersection and union [20]. A higher similarity score indicates a closer resemblance between the labels. According to the (Jaccard and Cosine) similarity, and Accuracy measures, the results obtained are as follows:

### 4.1. Jaccard Similarity

values range from 0 to 1, with 1 indicating complete similarity and 0 indicating no similarity. In this case, all Jaccard Similarity scores are 1.0, This suggests that the category names in every three labels are the same.

### 4.2. Cosine Similarity

values range from -1 to 1, with 1 indicating perfect similarity, 0 indicating no similarity, and -1 indicating complete dissimilarity (orthogonal). In this case, the cosine similarity between (VADER labeling and Manual labeling) was (0.979), and between (Cluster labeling and Manual labeling) was (0.868), suggesting that the first pair distributions have a stronger similarity than the second pair distributions.

### 4.3. Accuracy measures

the difference in accuracy between each labeling method and manual labeling was computed across all models. A smaller difference in accuracy indicates that the labeling approach is closer to the manual label [21].

In the cluster labeling approach, the average difference in accuracy compared to manual labeling is approximately 33.92%, while for the VADER labeling approach, the average difference in accuracy compared to manual labeling is approximately 3.26%. Since the average difference in accuracy for the VADER labeling approach is smaller compared to cluster labeling, it indicates that the VADER labeling approach is much closer to the manual label than the cluster labeling approach, those results are shown in Figure 4.

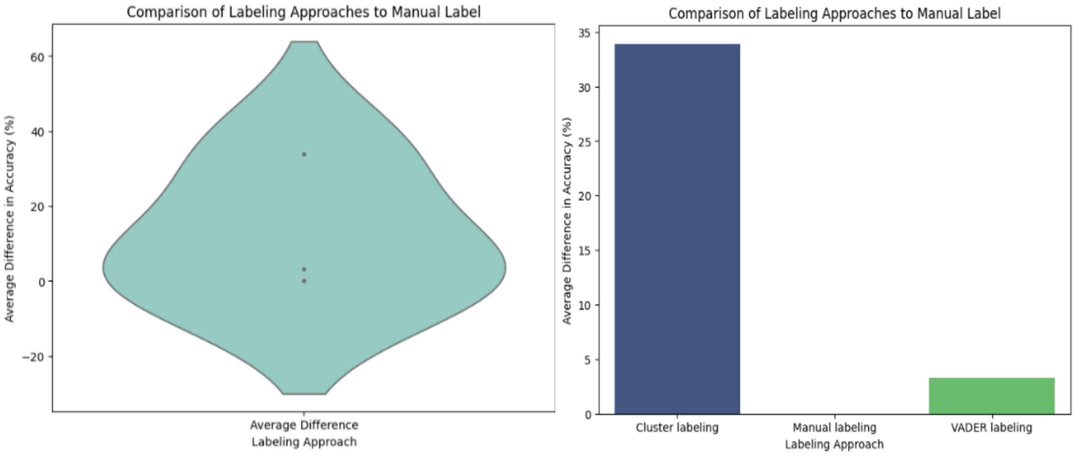

**Figure 4:** The average difference in accuracy compared to the manual labeling.

According to the results obtained in all three criteria, the automatic label NLTK_VADER has a closer analysis rate to the manual analysis than the cluster analysis rate. Therefore, in case you need to make a quick assessment and analyze several topics, you can use NLTK_VADER to label texts like manual labels without using complex algorithms.

## 5. Conclusion

Many techniques and studies have been tried and tested in text classification, but what makes our paper different from past works, this paper presents a sentiment analysis conducted on Twitter data related to the Turkey Earthquake 2023. Twitter is a popular platform, where users can also express their opinions on a variety of themes related to their everyday lives by writing tweets.

This study analyzed 2,800 tweets posted during and after the earthquake indicating which had positive and negative sentiments among citizens, then used a machine learning classification approach to determine which labeling types of accuracy were much closer to the manual labeling.

Finally, VADER labeling was found to be more effective and suitable for determining the emotional tone or sentiment expressed in social media texts, especially tweets. Because manual labeling requires a lot of time and human expertise, and the accuracy of VADER labeling is much closer to manual labeling accuracy.

## 6. Future work

Although the results obtained in this study were not highly accurate, they could be useful and improved by using more appropriate criteria and methods in the future.

## Acknowledgments

Thanks to those who helped us in this study, especially Professor Dr. Oguz Findk in Turkey who supervised this paper, (Dr. Aras Asad in the UK and Mr. Mario Caesar) in Indonesia for their comments, and My brother Engineer Amanj Kamal in Iraq for the linguistic review.

## References

[1] Shehu, H. A., Tokat, S., Sharif, M. H. & Uyaver, S. *Sentiment Analysis of Turkish Twitter Data*.

[2] Agrali, Ö., Sökün, H. & Karaarslan, E. Twitter Data Analysis: Izmir Earthquake Case. (2022).

[3] Varol Malkoçoğlu, A., Orman, Z. & Şamlı, R. Comparison of Different Heuristics Integrated with Neural Networks: A Case Study for Earthquake Damage Estimation. *Acta Infologica* **6**, (2022).

[4] Contreras, D., Wilkinson, S., Alterman, E. & Hervás, J. Accuracy of a pre-trained sentiment analysis (SA) classification model on tweets related to emergency response

and early recovery assessment: the case of 2019 Albanian earthquake. *Natural Hazards* **113**, 403–421 (2022).

[5] Juhász, P. L., Stéger, J., Kondor, D. & Vattay, G. A Bayesian approach to identify Bitcoin users. *PLoS One* **13**, (2018).

[6] Toraman, C., Kucukkaya, I. E., Ozcelik, O. & Sahin, U. Tweets Under the Rubble: Detection of Messages Calling for Help in Earthquake Disaster. (2023).

[7] Yufei Xie, st & Rodolfo Raga Jr, nd C. *Convolutional Neural Networks for Sentiment Analysis on Weibo Data: A Natural Language Processing Approach*.

[8] Qu, Z., Wang, F., Chen, X., Wang, X. & Zhou, Z. Rapid report of seismic damage to hospitals in the 2023 Turkey earthquake sequences. *Earthquake Research Advances* **3**, 100234 (2023).

[9] Harald Hornmoen & Klas Backholm. *Social Media Use in Crises and Risks: An Introduction to the Collection*. (2018).

[10] Despoina Antonakaki, Paraskevi Fragopoulou & Sotiris Ioannidis. A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *ScienceDirect* (2020).

[11] Yue, L., Chen, W., Li, X., Zuo, W. & Yin, M. A survey of sentiment analysis in social media. *Knowl Inf Syst* **60**, 617–663 (2019).

[12] GABRIEL PREDA. Turkey Earthquake Tweets. (2023).

[13] Wankhade, M., Rao, A. C. S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* **55**, 5731–5780 (2022).

[14] Meng, Y. *et al.* Text Classification Using Label Names Only: A Language Model Self-Training Approach. (2020).

[15] Friedman, L., Prokopenko, V., Djanian, S., Katrychuk, D. & Komogortsev, O. V. Factors affecting inter-rater agreement in human classification of eye movements: a comparison of three datasets. *Behav Res Methods* **55**, 417–427 (2023).

[16] Abou-Kassem, T., Alazeezi, F. H. O. & Ertek, G. A Data Analytics Methodology for Benchmarking of Sentiment Scoring Algorithms in the Analysis of Customer Reviews. in *Lecture Notes in Networks and Systems* vol. 693 LNNS 569–581 (Springer Science and Business Media Deutschland GmbH, 2023).

[17] Al Mahmoud, R. H., Hammo, B. H. & Faris, H. Cluster-based ensemble learning model for improving sentiment classification of Arabic documents. *Nat Lang Eng* (2023) doi:10.1017/S135132492300027X.

[18] Rahman, H. et al. Multi-Tier Sentiment Analysis of Social Media Text Using Supervised Machine Learning. *Comput. Mater. Contin* **74**, 5527–5543 (2023).

[19] Hui Lin & Ming Li. *Practitioner's Guide to Data Science*. (2023).

[20] Joyinee Dasgupta, Priyanka Kumari Mishra, Selvakuberan Karuppasamy & Arpana Dipak Mahajan. A Survey of Numerous Text Similarity Approach. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 184–194 (2023) doi:10.32628/cseit2390133.

[21] Cascante-Bonilla, P., Tan, F., Qi, Y. & Ordonez, V. *Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning*. www.aaai.org (2021).