All Sciences Proceedings
http://as-proceeding.com/

1st International Conference on Pioneer and Innovative Studies

June 5-7, 2023 : Konya, Turkey

© 2023 Published by All Sciences Proceedings

ICPIS

https://as-proceeding.com/index.php/icpis

Sentiment Analysis and Emojification of Tweets

Enes Cerrahoğlu[1], Pınar Cihan [1*]

[1]*Computer Engineering Department, Tekirdağ Namık Kemal University, Turkey*

[*]*pkaya@nku.edu.tr*

Abstract **–** Social media platforms have become a prevalent means for individuals to share their emotions and thoughts. With millions of tweets being posted on Twitter every day, these tweets provide us with a vast dataset. Conducting sentiment analysis on this dataset can be a valuable method to obtain meaningful insights about societal trends. For this purpose, a sentiment analysis model and a web interface that emojifies emotions were developed using the Python programming language. This model works on tweets shared on Twitter and utilizes natural language processing techniques to determine the sentiment of the tweets. In this study, 168.274 English tweets were collected using the Twitter API. The collected tweets underwent a cleaning process where URLs, hashtags, mentions, and emojis were removed. Then, the TextBlob Python library was employed to label the tweets as positive, negative, or neutral. The labeled tweets were subjected to classification accuracy testing using Gradient Boosting, Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machines machine learning models. The findings revealed that logistic regression achieved the highest classification accuracy with 94%. Lastly, a web interface was developed, which retrieves the last 50 tweets of a queried user's profile and appends a relevant emoji based on the sentiment of each tweet.

Keywords – Sentiment Analysis, Natural Language Processing, Twitter, Emojification, Machine Learning

I. Introduction

Sentiment analysis is a process used to determine the emotional tone and sentiment of texts by employing natural language processing and machine learning techniques [1]. This analysis is utilized to identify positive, negative, or neutral emotional expressions.

Sentiment analysis utilizes various techniques to determine the emotional tone and sentiment of texts. These techniques allow emotions to be classified as positive, negative, or neutral. These methods generally involve natural language processing and machine learning algorithms. Among the methods used for sentiment analysis are techniques such as word distribution, word association, word sensitivity, and topic modeling [2]. These techniques analyze the frequency of words in the text, their co-occurrence patterns, sensitive words, or different topics to determine the emotional tone of the text.

Sentiment analysis is an important tool used in various industries. For example, it can be used to measure the impact of a brand's social media campaigns. By analyzing the tweets shared by customers, a brand can assess how effective their campaigns are on customers. This analysis can help the brand gauge customer satisfaction and interest in their products, aiding them in determining future marketing strategies.

Political parties or candidates can analyze the opinions of voters on social media to measure the popularity of their policies and the motivation of their supporters. This analysis can help them identify the most discussed topics and issues among voters, enabling them to shape their policies and campaigns accordingly.

The healthcare sector can utilize information about health-related topics shared by patients on social media to improve healthcare policies or disease management. For instance, tweets discussing the stress or anxiety caused by symptoms of a particular disease can help in providing psychological support or counseling services for disease management.

When examining the studies conducted in this regard, Putri et al [3]. used the Support Vector Machine (SVM) method for sentiment analysis of tweets in Jakarta, Bandung, and Medan, which are three major cities in Indonesia. The tweets were classified as positive, negative, or neutral. The research was conducted in four stages: preprocessing, target-dependent classification, SVM method, and opinion classification. As a result, the target-dependent approach was successfully applied for sentiment analysis in Bandung, Jakarta, and Medan cities. Subsequently, sentiment classification was performed using the SVM method. Furthermore, it was observed that the target-dependent approach influenced the number of positive, negative, and neutral opinions regarding Bandung, Jakarta, and Medan cities.

Xia et al [4]. conducted a study to investigate the effectiveness of ensemble learning method in sentiment analysis. The study concluded that traditional text classification approaches, such as Bag-of-Words (BOW), were not suitable due to their tendency to overlook certain words. Therefore, the study employed two types of features, namely POS and Word relationships, and three classifiers including NB, MaxEnt, and SVM. Three types of ensemble classifiers were proposed and evaluated: weighted clustering, fixed clustering, and meta-classifier clustering. The results obtained from the study demonstrated that ensemble learning methods led to significant improvements compared to individual classifiers, indicating their effectiveness in sentiment analysis.

Akgül et al [5]. conducted a study titled "Emotional Twitter" in which tweets collected with a specific keyword are automatically labeled as positive, negative, or neutral. Both a dictionary and an n-gram model were used for this purpose. The developed system is reported to allow individuals and organizations to create custom dictionaries for themselves. The study resulted in an approximate success rate of 70% for the dictionary-based method and 69% for the character-based n-gram method.

Osmanoğlu et al [6]. scaled 6059 feedbacks using the triadic Likert method for online materials such as books, audio books (mp3), videos, and interactive tests in their study. Subsequently, sentiment analysis was performed using machine learning techniques. The study resulted in an accuracy rate of 0.775 using the logistic regression algorithm.

Ilhan and Sağaltıcı [7], aimed to perform sentiment analysis on 1,578,627 classified tweets obtained from Twitter. Various machine learning techniques were used for this purpose. In the study, the analysis of positive and negative sentiments was carried out using the N-gram method, and performance comparisons of relevant classifiers were made using commonly used machine learning methods such as Naïve Bayes and Support Vector Machines. According to the results obtained, the successful method was found to be the Support Vector Machines classifier.

In this study, Gradient Boosting, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machines machine learning methods were used to perform sentiment analysis on 168,274 classified tweets obtained using the Twitter API. Subsequently, using the developed interface, the tweets of the desired profile were analyzed based on the most successful method, and the results were returned as emojis.

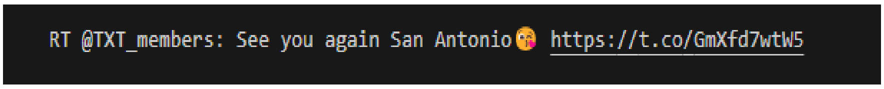
II. Materials and Method

A. *Data Acquisition*

In this study, up-to-date data was obtained to perform sentiment analysis on tweets. For this purpose, 168,274 English tweets were collected using the Twitter API. The Twitter API is a highly useful tool that allows direct access to Twitter data, providing a large data source for various analyses. The use of the Twitter API enables researchers to retrieve tweets related to specific keywords and track the current sentiment of the community. This is an important resource for both academic and commercial purposes. However, in order to create these data sets accurately, specific data analysis techniques need to be applied. These techniques should be carefully applied, especially in areas such as accurate classification, feature selection, and data cleansing. Nevertheless, properly prepared data sets can be used in various fields, such as sentiment analysis, and serve as a foundation for future research.

B. *Data Pre-processing and Classification*

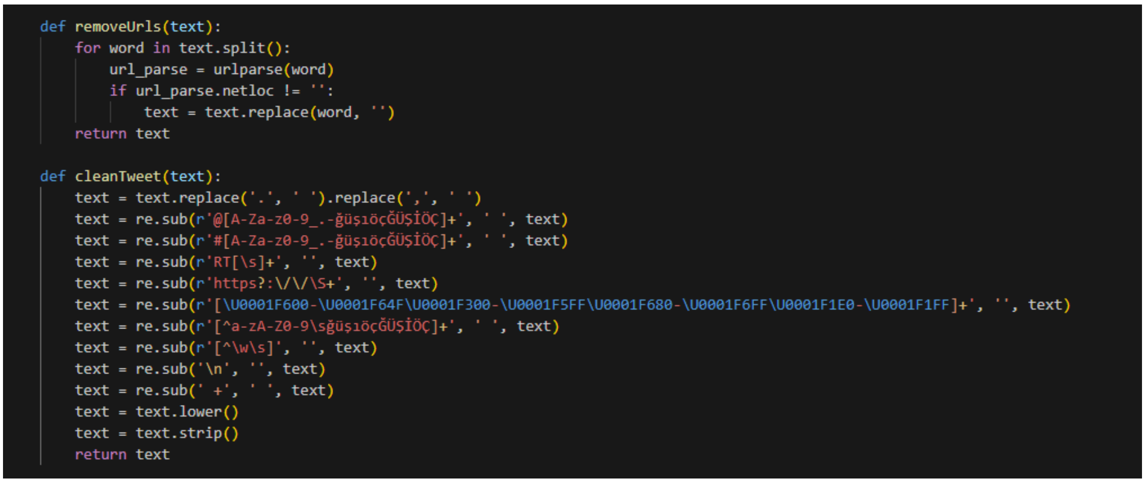
The dataset used in the study contains tweet data, which includes information such as URLs, emojis, hashtags, and mentions that do not affect sentiment analysis. Fig. 1 presents an example tweet content.



```
RT @TXT_members: See you again San Antonio😘 https://t.co/GmXfd7wtW5
```

Fig. 1 Example of a tweet content

Various functions were written to clean these information from the tweets. The code snippets for the functions responsible for data cleaning are provided in Fig. 2.

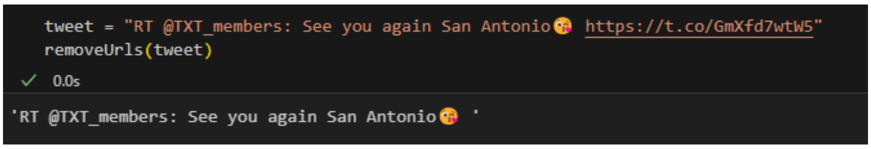


```python
def removeUrls(text):
    for word in text.split():
        url_parse = urlparse(word)
        if url_parse.netloc != '':
            text = text.replace(word, '')
    return text

def cleanTweet(text):
    text = text.replace('.', ' ').replace(',', ' ')
    text = re.sub(r'@[A-Za-z0-9_.-ğüşıöçĞÜŞİÖÇ]+', ' ', text)
    text = re.sub(r'#[A-Za-z0-9_.-ğüşıöçĞÜŞİÖÇ]+', ' ', text)
    text = re.sub(r'RT[\s]+', '', text)
    text = re.sub(r'https?:\/\/\S+', '', text)
    text = re.sub(r'[\U0001F600-\U0001F64F\U0001F300-\U0001F5FF\U0001F680-\U0001F6FF\U0001F1E0-\U0001F1FF]+', '', text)
    text = re.sub(r'[^a-zA-Z0-9\sğüşıöçĞÜŞİÖÇ]+', ' ', text)
    text = re.sub(r'[^\w\s]', '', text)
    text = re.sub('\n', '', text)
    text = re.sub(' +', ' ', text)
    text = text.lower()
    text = text.strip()
    return text
```

Fig. 2 Functions that ensure data cleaning

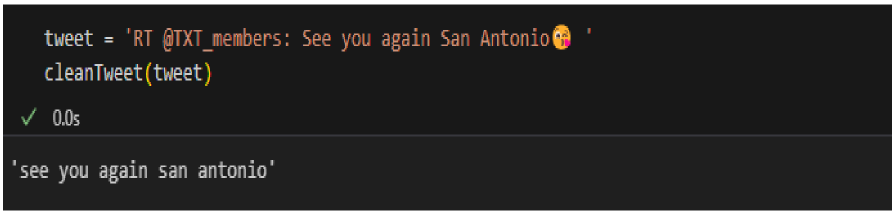
Thanks to the implemented functions, URL information in the tweets has been cleaned (Fig. 3).



```
tweet = "RT @TXT_members: See you again San Antonio😘 https://t.co/GmXfd7wtW5"
removeUrls(tweet)
✓ 0.0s
'RT @TXT_members: See you again San Antonio😘 '
```

Fig. 3. Tweet content with URLs cleaned

After removing URLs from the tweet data, a cleaning process was applied to eliminate other factors that needed to be cleaned, such as emojis, hashtags, mentions, punctuation marks, and numerical values. Fig. 4 presents an example of the cleaned tweet content.



```
tweet = 'RT @TXT_members: See you again San Antonio😘 '
cleanTweet(tweet)
✓ 0.0s
'see you again san antonio'
```

Fig. 4. Cleaned tweet content without other factors

After the data cleaning process, the dataset has been prepared for further processing and classification. The training dataset requires labeled tweets for classification. There are two options to accomplish this. The first option is to manually classify each tweet as positive, negative, or neutral. However, this process can be laborious and time-consuming. Additionally, it can be challenging to classify tweets objectively. The second option is to use pre-trained models such as TextBlob and Bert, which generate sentiment classification outputs, to classify the dataset. In this study, TextBlob was used to classify each tweet in the dataset as positive, negative, or neutral. TextBlob determines the negativity or positivity of a text based on its polarity value, ranging from -1 to +1. Each tweet in the dataset was classified as negative if the polarity value was less than zero, neutral if it was equal to zero, and positive if it was greater than zero.

The example tweets from the classified dataset are provided in Table 1.

Table 1. Example Tweets from the Classified Dataset

Tweet	Sentiment
see you again san Antonio	Neutral
today is going so bad	Negative
i am very happy to be promoted in my company	Positive

III. Results

After retrieving 168,274 tweets using the Twitter API, data cleaning processes were applied. The cleaned tweets were then classified into positive, negative, or neutral sentiments using the TextBlob model. The classified data was used to test the classification performance of five different machine learning models: Gradient Boosting, Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machines. During the training phase, 80% of the dataset was used, while the remaining 20% was used to evaluate the models' performance. The accuracy scores and training times of the models are presented in Table 2.

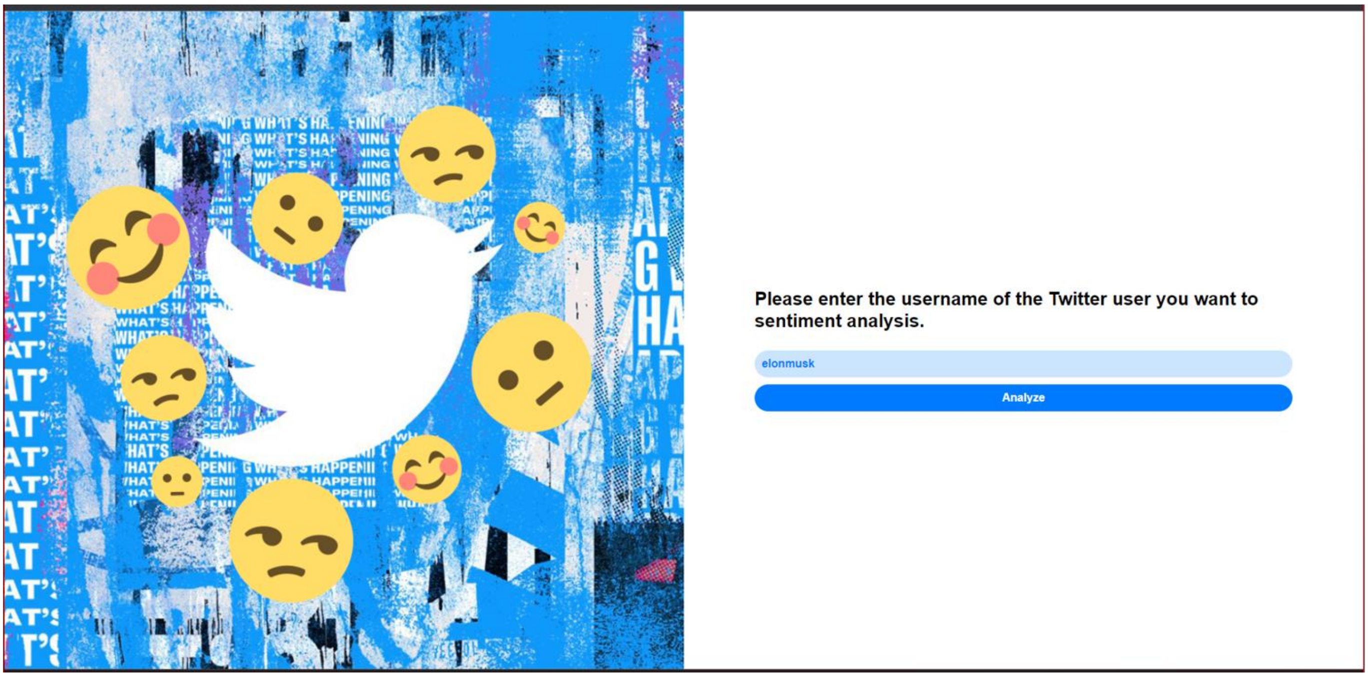
Table 2. Model test performance and training times

Model	Accuracy (%)	Training Time (s)
Gradient Boosting	75.52	267.7
Logistic Regression	94.05	18.3
Naive Bayes	73.62	5.4
Random Forest	82.80	12.9
Support Vector Machines	88.70	>6600

When examining the results, it can be observed that the Logistic Regression method outperforms the other methods and has a relatively faster training process.

The model trained using the Logistic Regression method is saved as a 'pkl' file using the pickle library. The saved model is then utilized in a function developed with Google Cloud Functions to create an API. When a POST request is sent to the URL provided by Google Cloud Functions, with the JSON data in the body section in the format {"username": "username"}, it returns a JSON result containing the profile information of the entered Twitter username, the latest 50 tweets and retweet contents shared by that account, and the prediction results for these tweet and retweet contents.

Using this cloud function, a website is developed using the Python Flask framework. The website consists of two pages. The first page is where the user enters their Twitter username (Fig. 5).

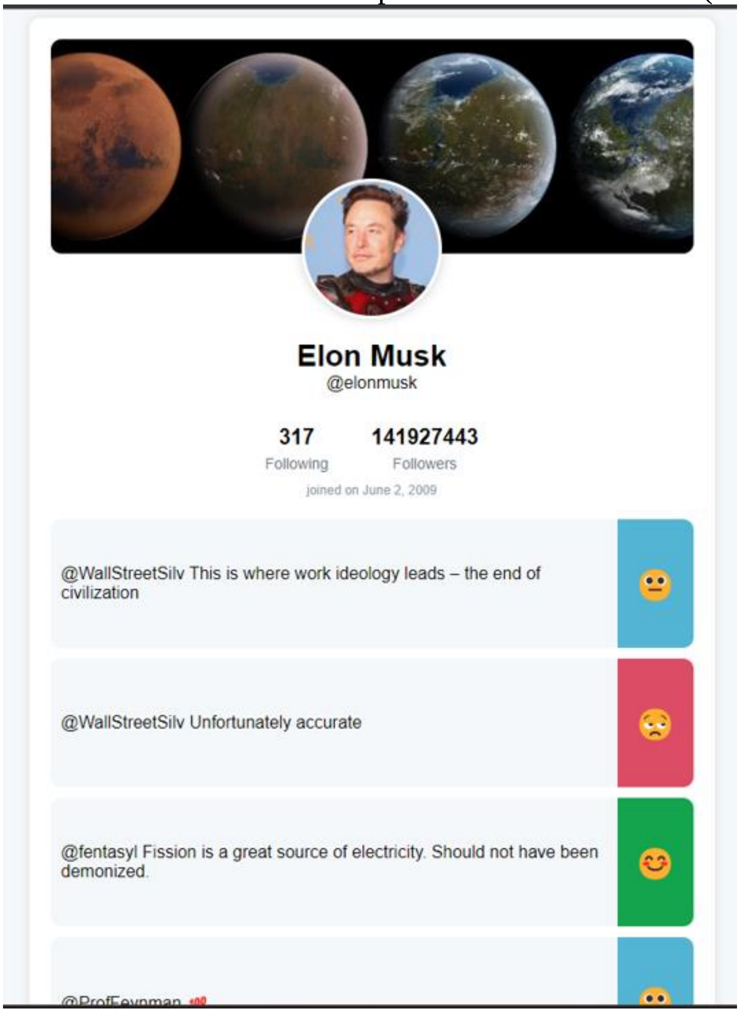


Please enter the username of the Twitter user you want to sentiment analysis.

elonmusk

Fig. 5. Homepage for entering the Twitter username

After entering the Twitter username on the page shown in Fig. 5, 50 tweets belonging to the user are retrieved. These retrieved tweets are then classified using our successful method, Logistic Regression. Based on the resulting classification, emojis are placed next to the tweets (Fig. 6).



Elon Musk
@elonmusk

317 Following
141927443 Followers
joined on June 2, 2009

@WallStreetSilv This is where work ideology leads – the end of civilization 😐

@WallStreetSilv Unfortunately accurate 😟

@fentasyl Fission is a great source of electricity. Should not have been demonized. 😊

@ProfFeynman 💯

Fig. 6. Page where tweets are emoji-fied based on tweet content

IV. DISCUSSION

In this study, a website has been developed where tweets are classified using logistic regression for sentiment analysis, and these sentiments are represented with emojis. There are several potential benefits of a web page where tweets are emoji-fied based on their content, including:

Visual Expression and Emotional Conveyance: The use of emojis allows for better conveyance of emotional expressions and tone. Adding emojis to tweets based on their content enriches the emotional context and makes communication more effective.

Quick Understanding and Perception: People can derive meaning faster with visual symbols. Supporting tweets with emojis can facilitate easier understanding of the content and quick perception.

Attention Grabbing and Increased Engagement: The use of emojis can help tweets grab attention and generate interest. Visual symbols can encourage users to show more interest in and share the content.

Personalization and Brand Image: The use of emojis can contribute to personalizing a web page or brand. Unique emoji sets or styles can be created for a specific brand or web page, enhancing brand image and recognition.

Social Media Interaction: The use of emojis can encourage interaction among social media users. Reactions, comments, and shares can be generated with emojis that are appropriate to the content of the tweets, thereby increasing social media engagement.

However, there are still some challenges that need to be addressed to further improve the developed model. For example, factors such as the complexity of emotional expressions and the multi-layered nature of texts can hinder the accuracy of sentiment analysis results. Therefore, future studies should focus on using more advanced natural language processing and machine learning techniques and expanding the dataset to achieve more accurate results.

V. CONCLUSION

In this study, the Twitter API was used to fetch up-to-date tweets. The obtained tweet data was cleaned by removing URL, hashtag, mention, and emoji information using the implemented functions. Then, the tweets were labeled as positive, negative, or neutral using the TextBlob Python library. The prediction performance of Gradient Boosting, Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machines methods in sentiment analysis was tested.

The results showed that the most successful method was logistic regression with an accuracy of 94%. Finally, a web page was developed using the Python Flask framework. Through this site, the latest 50 tweets of a Twitter user were fetched, and the tweets were emoji-fied based on the sentiment analysis result using the logistic regression method.

This study presents a detailed design, implementation, and results of a sentiment analysis model developed using the Python programming language. This model works on Twitter data and uses natural language processing methods to detect the sentiment of tweets. The results obtained in the study showed that the Logistic Regression method achieved the highest accuracy rate of 0.9405.

REFERENCES

[1] Mullen, T., and Collier, N. (2004, July). Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 412-418).

[2] Kwon, H. J., Ban, H. J., Jun, J. K., and Kim, H. S. (2021). Topic modeling and sentiment analysis of online review for airlines. Information, 12(2), 78.

[3] Putri, T. T. A., Mendoza, M. D., and Alie, M. F. (2020). Sentiment Analysis On Twitter Using The Target-Dependent Approach And The Support Vector Machine (SVM) Method: Sentiment Analysis On Twitter Using The Target-Dependent Approach And The Support Vector Machine (SVM) Method. Jurnal Mantik, 4(1), 20-26.

[4] Xia, R., Zong, C., and Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information sciences, 181(6), 1138-1152.

[5] Akgül, E. S., Ertano, C., and Diri, B. (2016). Twitter verileri ile duygu analizi. Pamukkale University Journal of Engineering Sciences, 22(2).

[6] Osmanoğlu, U. Ö., Atak, O. N., Çağlar, K., Kayhan, H., and Talat, C. A. N. (2020). Sentiment analysis for distance education course materials: A machine learning approach. Journal of Educational Technology and Online Learning, 3(1), 31-48.

[7] İlhan, N., and Sağaltici, D. (2020). Twitter'da duygu analizi. Harran Üniversitesi Mühendislik Dergisi, 5(2), 146-156.