JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Sentiment analysis and clustering of trending tweets from Twitter including emojis

[1]Vaishnavi S. Manthalkar, [2]Sudhakar R. Barbade

[1]PG Student, [2]Assitant Professor
[1]Department of Electronics Engineering,
[1]Walchand Institute of Technology, P.A.H. Solapur University, Solapur, India

Abstract : This Twitter (Now X) is a major source of modern-day connectedness to social, political and entertainment waves of events and reactions from the global community. It also is a unique source of personal opinions and interests which in the form of digital information acts as an opportunity to extend the understanding of impact of freedom of digital speech. The aim of this work is to develop capacity to work with live data from this source to analyze the sentiment of people's reactions and incorporating the subtle and at times strong or complex change in sentiment by inclusion of emojis. This work was conducted in India and in recent times Twitter has gained a wider userbase in the country which is a host to people who react in multilingual phrases and tweets. This work acknowledges the popular languages in use and segregate content based on their content similarity using K-means clustering. The use of unsupervised learning led to an average silhouette score of 0.78 on a sample live data sourced from Twitter which indicated that the quality of the clusters is good and clusters are clearly distinguished. The paper also talks about challenges and opportunities to take this work forward.

***IndexTerms* - Sentiment analysis, clustering, multi-lingual data analysis, silhouette analysis, emoji interpretation, elbow curve.**

I. INTRODUCTION

With the rise of internet technologies and social media sites like Twitter (now X), people are able to easily and quickly share their thoughts, feedback, and opinions with others. Twitter is a widely used social networking platform where millions of users share their thoughts and opinions on a variety of topics through tweets. These tweets often include hashtags, retweets, and user mentions to categorize and share content with others. With a restriction of 280 characters per tweet, users can quickly share their thoughts on current events, government policies, sports tournaments, and more.

On Twitter people express in volumes as much as half a billion tweets daily as of last published volume in 2014 [1] which is equivalent of around 8 TB new data added every day, the userbase has since increased multifold. If this data is analyzed, information can be extracted by mining specific aspects of concurrent events, trends and opinions. Twitter has become a popular platform for evaluating consumer opinions on products or services through sentiment analysis. This involves using machine learning algorithms to classify tweets as positive, negative, or neutral based on the viewpoint shared. However, creating a dataset for sentiment analysis can be complex and time-consuming, with issues such as overfitting and bias. Despite these challenges, sentiment analysis remains an important tool for companies and influential figures to understand public opinion. Python now has multiple libraries of Natural Language Processing, NLP, which has unleashed the potential to provide functionalities of parsing, tagging and semantic classification, it can help analyze linguistic structures and thereby derive sentiment towards the queried topics. Analyzing Twitter trends and public sentiment through tweets can provide valuable insights for organizations and companies. By understanding how the public feels about certain topics, products, or events, organizations can make informed decisions based on this data.

In the scope of this article, tweets are collected using the Twitter API and analyzed using counting methods and various machine learning algorithms. This analysis can provide valuable insights into public sentiment and help organizations make informed decisions based on current trends. There has been a significant amount of research done on Twitter sentiment analysis. Researchers have used various algorithms and approaches to analyze tweets and understand public sentiment on a wide range of topics. For example, one study introduced an efficient system for Twitter sentiment analysis that used machine learning to detect positive and negative tweets [2].

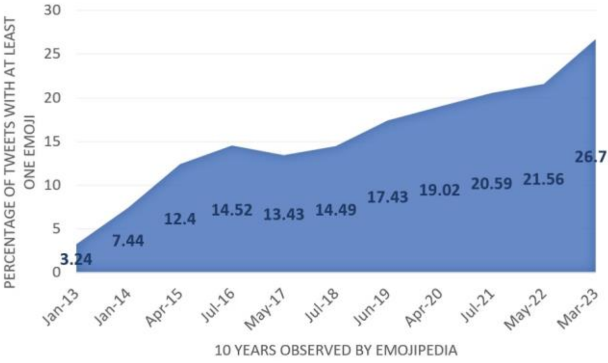

Figure 1 Rise of share of posts on Twitter containing emojis from 2013 to 2023

Another study provided an overview of the algorithms and approaches used for sentiment analysis on Twitter, categorizing them into four categories based on their approach [3]. In the educative article on using machine learning to analyze Twitter sentiment, classifiers such as Logistic Regression, Bernoulli Naive Bayes, and SVM, along with techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) were advised [4].

Ahuja and Dubey [5] also did sentiment analysis on twitter data with dictionaries named AFINN and TextBlob for polarity and subjectivity for individual tweets. They included emoticons in their study with unsupervised techniques. Their main focus was on clustering based on the polarity and subjectivity and deployed K-means clustering for this purpose. Subjectivity was based on capacity to identify a specific tweet as a personal opinion based on relating the tweet with personal nouns and pronouns. They limited their analysis to graphical method of optimization.

Till the time the current work, under the scope of this article, was conducted, the content on Twitter has evolved and the need for missing insights based on emojis has risen due to its extensive usage amounting to 26.7% of the Tweets containing emojis in 2023 as against less than 4.25% back in 2013 [6]. The current work focuses on using live data inclusive of emojis from Twitter and analyzing it to get insights, on sentiment that the trending tweets carry, employing clustering based on unsupervised learning by K-means clustering and the outcome is evaluated for its efficacy using silhouette analysis.

II. Materials and Methods

Twitter API is used as the data source to collect the information in the form of tweets and this data was then processed as described in the flow chart in **Fig. 2**. Detail on each individual steps follows hereafter. Trends are geographically differently focused. In this work, the focus will be tweets popular and trending in India. For India, the WOEID (Where-On-Earth-ID) is 23424848, which is considered when sourcing data from Twitter.

2.1 Primary Trend Analysis

In this paper, Tweepy [7], a python library, is used to analyze the data from Twitter. The trending tweets were sourced from Twitter API for the country, India, and tweet volume was calculated. The trending hashtags were also sourced from the same API to create a criterion for top trends. An attempt was made to understand the bias of source of creation the tweets in terms of device used viz. Android, Web app, iPhone; political state of users' location.

2.2 Fetch tweets for Top Trends

The trending hashtags were used to query a thousand tweets from the collection of tweets that were sourced. A word cloud was generated to visualize the gamut of content available for analysis.

2.3 Language detection

From the top trending tweets, it was necessary to understand the language in which the tweets were written. Langdetect module [8] was used to differentiate between languages/scripts of the tweets. The most frequent language was then to be considered for further analysis.

2.4 Data cleaning and pre-processing

Sourcing custom written text doesn't often lead to clean, well-formatted data and more so while developing a machine learning project. In data pre-processing, the intent was to convert raw data into something that can be conveniently used by the machine learning model. We will look into text processing in further detail.

2.4.1 Text Processing

2.4.1.1 Tokenization

Within this step, text as sourced from Twitter was broken down into smaller units, viz. Tokens. Tokenization helps create raw text data for further processing.

2.4.1.2 Lower casing

All tokenized text data is converted into single casing text structure, in this case, lower case. Implementing a common case reduces the complexity of processing multiple cases.

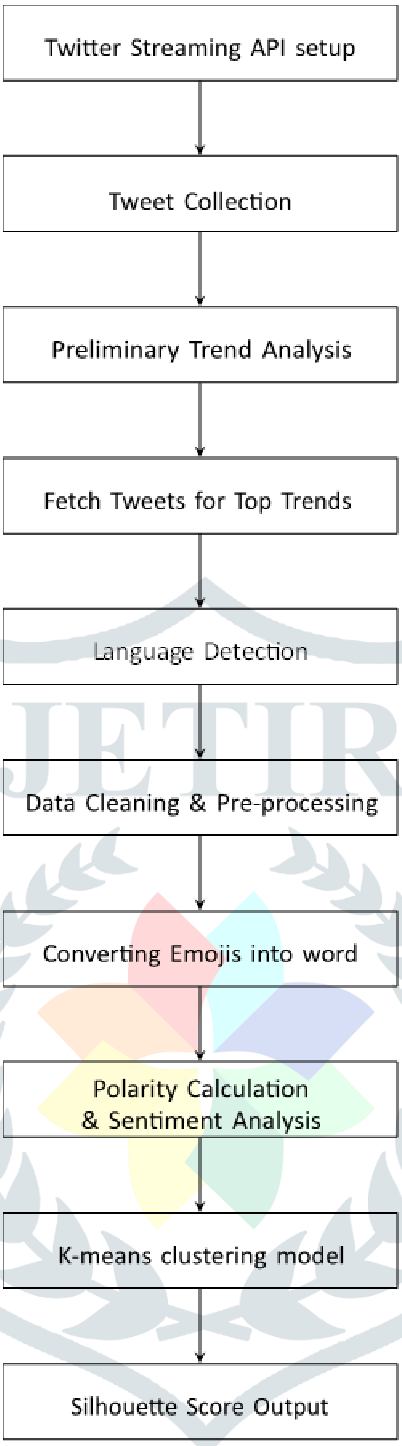
Twitter Streaming API setup
Tweet Collection
Preliminary Trend Analysis
Fetch Tweets for Top Trends
Language Detection
Data Cleaning & Pre-processing
Converting Emojis into word
Polarity Calculation & Sentiment Analysis
K-means clustering model
Silhouette Score Output

Figure 2 Process flow chart in use in this work for Twitter Sentiment Analysis

2.4.1.3 Punctuation removal

The lower-case tokens are then processed to eliminate the punctuation marks used in the text because since the text is broken into tokens, punctuations will not carry useful information there onwards.

2.4.1.4 Stop word removal

Stop words are nothing but most common use words which add very little value in the overall meaning of the text. Removing stop words ease the process of training the model on text data.

2.4.1.5 Removal of special repeated words/characters

On Twitter, ‘Rt’ and ‘@’ are unique repeated words and characters which are removed under this category. Apart from this, web links are identified and removed, since the intent is to redirect the users to different webpage without the text containing exact or complete information.

Table 1 Emojis and their meaning in text

Emoji symbol/image	Closest meaning
🍎	Apple
😋	Delicious
👧	Girl face
🙉	Hear no evil monkey
🎭	Performing Art

2.4.2 Using TD-IDF vectorizer

TDF-IDF vectorizer is a text vectorization algorithm which transforms text into vectors. Python's Pandas library is used for managing the data in this work, which stores the data in the form of arrays but the natural language processing would only understand vectors and hence the text from the arrays is hereby converted into vectors using TF-IDF vectorizer. Here TF means Term Frequency and IDF means Inverse Document Frequency. Term frequency corresponds to a calculated ratio of total count of instances for which a particular term is repeated in a document to the total number of words in the document. Document Frequency would mean the total count, a particular term would be present, in a set or number of documents in which and thus Inverse Document Frequency would mean the ratio of number of documents containing a particular term to the Document Frequency of that term. A log of this ratio is referred for calculation purposes and presented as Inverse Document Frequency. In Python, the authors have used sklearn module [9] containing the method, TfidVectorizer(), to transform the text into vectors.

2.5 Converting emojis into text

People often use emoji as a part of their expression in digital writing and to complement their text message. Emoji is a small text size image popularly in use to express explicit or implicit emotions when used along the text messages. A sample of a few emojis is presented in Table 1 with their inferred closest meaning.

Emoji can help interpret the context and sentimental implication of the messages which becomes more sensitive when it may be regarding a product review. Without having the capacity to interpret the emoji, valuable part of the information will be lost. In this work, it was chosen to convert these emojis to text using emoji python package [10] so that the information which was contained in the emoji could be preserved and converted into a format which the natural language learning models could easily interpret.

2.6 Polarity calculation and Sentiment analysis

In natural language processing, sentiment is analyzed using NLTK VADER sentiment analyzer [11], which is driven by the usage of positive and negative words and phrases and thus deriving the sentiment carried in the statements. It is based on syntactic and lexical features of the phrases. The term 'VADER' itself stands for 'Valence Aware Dictionary for sEntiment Reasoning', meaning that it derives its sensitivity to positive and negative polarity based on dictionary of words which together with the intensity of emotions provides a valence score into consideration. Valence score has a scale from -4 as minimum for the most negative sentiment to +4 as the maximum, carrying the most positive sentiment. The median of this scale is the neutral sentiment. VADER is chosen because it is an effective and resource light model, it doesn't require huge data to train the model and thus opted as best alternative for streaming tweets.

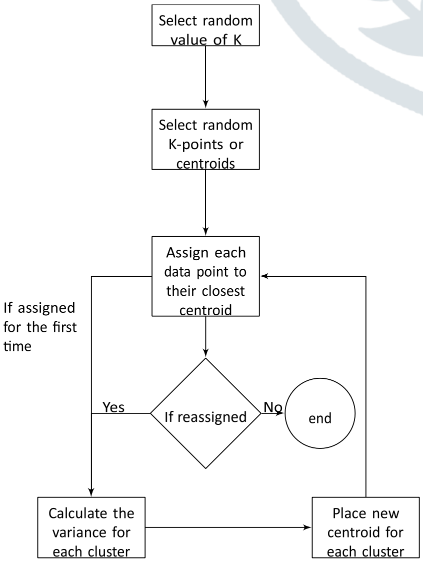
Select random value of K
Select random K-points or centroids
Assign each data point to their closest centroid
If assigned for the first time
Yes
If reassigned
No
end
Calculate the variance for each cluster
Place new centroid for each cluster

Figure 3 Method for K-means clustering

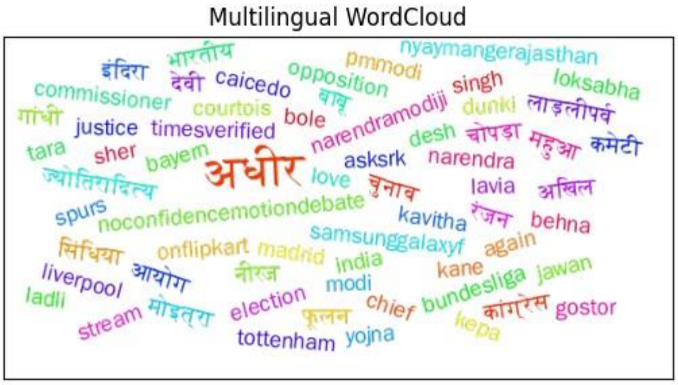

Figure 4 A word cloud of trending tweets collected

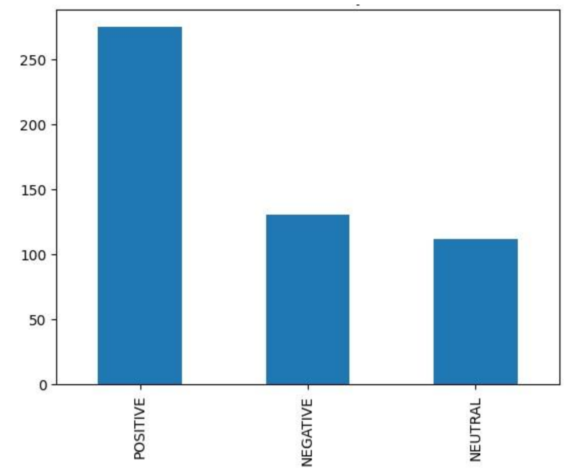

Figure 5 Histogram of polarity assigned into positive, negative and neutral sentiments after sentiment analysis

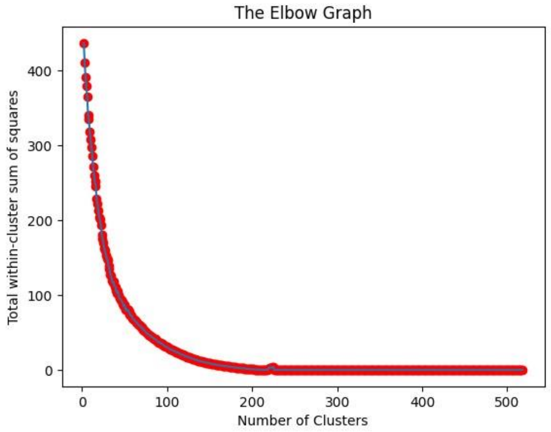

Figure 6 Elbow curve

2.7 K-means Clustering

Among the unsupervised learning methodologies, K-means clustering [9] was chosen as it iteratively divides data into K clusters by minimizing the variance in each cluster. This is done to segregate the data into groups which are comparably similar to each other but different from the other groups. The character K is also used as a means to understand the number of such groups to be formed. If K=4, then four clusters must exist. But the value of K is not guessed but it is deduced to be optimum as per given data. The method is described in the form of a flow chart in **Fig. 3**.

K-means clustering is relatively simple and convenient to implement, easier to understand, hence chosen to use for the required clustering task. It also is computationally efficient and considering the limited resources used in this work, it becomes an optimum choice. K-means clustering was also capable of handling large dataset; hence it is scalable to future usage for the designed task, considering evermore increasing data from Twitter. Apart from Twitter, it can be used for studying search engines, sensor network, diagnostic systems and academic performances as well. So, a similar approach of this work can be utilized in other applications. An optimal method used to determine the value of K, elbow method or elbow technique is used. This method requires iteration from an initialized value of K=1 till K=n, where n is a hyperparameter. In current work, n is chosen to be 500. For every value of K, the algorithm calculates Within-Cluster-Sum-of-Squares (WCSS). WCSS is the sum of square of distance between the centroids (K) and each point, which helps refine the reassignment of the data point to a new centroid. WCSS is highest for the initialized value of K, and it decreases thereafter. When visualized, it forms an elbow shape on a 2-D graph between K and number of iterations (or value of n) since the rate of decrease of WCSS keeps decreasing. The value of K, for which the rate of decrease is negligible, is chosen to be the optimal value of K

2.8 Silhouette analysis

It is a graphical technique used to evaluate the quality of the clusters evolved by a clustering algorithm, which in this case is the K-means clustering outcome. It requires the calculation of the silhouette coefficient [9] for each data point and using histogram for the visualization. On such histograms, histograms get more weighted or wide if the size of the cluster is bigger owing to smaller clusters being banded together into a bigger cluster and on the other side, the narrow histograms are considered an indication of separated clusters.

Silhouette coefficient is also referred to as silhouette score, which is considered as an indicator of efficacy of clustering algorithm’s implementation. The silhouette score can have value ranging from -1 to 1, where 1 corresponds to clusters which are distinct, well apart and clearly distinguished; 0 corresponds to clusters which are close to each other and their relative bias is indifferent; -1 corresponds to clusters which are misassigned and do not have coherence.

III. RESULTS

As tweets were fetched in India, the tweets fetched included a representation of Indian scripts together with English tweets composing the trending tweets. Trending hashtag tweet words are converted into a word cloud (**Fig. 4**) for a relative understanding of the readers.

The most frequent language was English, so all further analysis was conducted on tweets in the English language. Using NLTK-VADER sentiment analyzer, the tweets were then classified into three sentiment polarities, i.e. Positive, Negative and Neutral. Based on the instant when the data was sourced, majority of tweets were of positive sentiment, followed by negative and neutral in the same order. The sum of count of these sentiment polarity tweets is reflected in **Fig. 5**.

The trending tweets were then processed to cluster the tweets based on the topic, meaning and sentiment using Kmeans clustering algorithm. The optimum number of clusters was identified using the elbow-method. For the instance of data sourced from twitter, the optimum number of clusters came out to be 196, which means that K is 196 for this instance. The elbow curve is depicted in **Fig. 6**, which shows the red filled circles as the values of WCSS for each value of K starting from 1 to 500. A blue trendline is drawn based

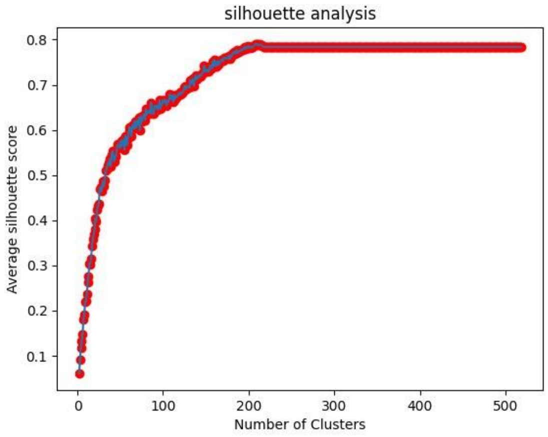

Figure 7 Silhouette score analysis

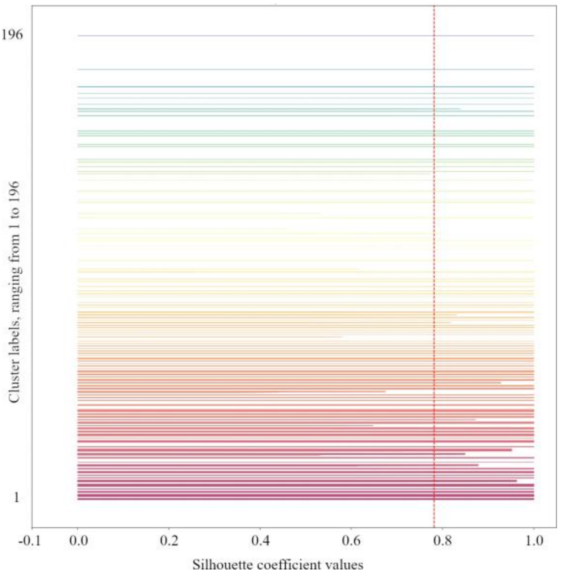

Figure 8 Silhouette plot for the clusters

on the trend followed by the data plot. The trendline follows an optimal elbow shape and the knee point is at K=196 from where onwards the WCSS does not change much thereafter.

The outcome of the elbow curve was then validated by conducting a silhouette analysis. Silhouette score was calculated for each combination of number of clusters formed and the silhouette score kept climbing as the number of clusters increased and again in accordance with the elbow curve, the silhouette score also followed the similar pattern (**Fig. 7**). The rate at which the silhouette score increased dropped significantly after K=196, hence 196 was confirmed to be the optimum number of meaningfully segregated clusters.

There is another graphical method to assess the outcome of the silhouette analysis, which is by plotting the silhouette score on a scale of -1 to +1, and an optimal number of clusters will have been achieved if the silhouette score of all the clusters individually would fall on the positive end of the scale, i.e. being more than 0. The silhouette score is thereby plotted for K=196 in **Fig. 8**, which clearly shows that all clusters have positive silhouette scores, making 196 an optimum number of clusters.

At K=196, the average silhouette score was 0.7817 (rounded up to the fourth decimal place), which is close to 1, meaning that the clusters had distinct meanings and can be considered for further analysis as per interest of the users.

IV. Discussion

This work was conducted with limited academic computational resources and hence the machine learning algorithms are chosen which are not computationally intensive. The presented work was accomplished using a computer with Intel Core i7-4710HQ CPU and 16GB RAM, and some of the time-intensive computational capacity was sourced from Google Colaboratory platform (free limited access) [12]. Since the algorithms had been chosen to be computationally light, the use of higher computational capacities will lead to faster and responsive sentiment classification and clustering for live trending tweets from the platform at smaller intervals of time. Twitter has been changing with time, but very rapidly, and access to tweets from API is also changing. Recently, free access is limited to posting tweets and developer access is now a paid access. For extending the current work, readers will have to follow the most recent terms of access provided at Twitter Development web resources page [13]. Since the count of Indian script in use in Twitter (X) is on the increasing trend, the authors encourage the further development of this methodology to assess the content in Indian scripts as well.

V. Conclusions

In this work, the authors intended to devise a collective methodology to source tweets from the Twitter (X) developer platform and extract information in the form of sentiment and classify them into clusters with distinct or similar areas of interest and meanings. The implicit meaning carried in the modern communications along with emoticons were assessed using NLTK-VADER itself, but the meaning carried by the emojis were also included by converting the emojis into their equivalent text forms. The methodology hereby created was successfully implemented on more than 1.8 lac live sourced tweets from Twitter (X) and an optimal number of 196 clusters were formed for the top tweets, with a silhouette score of 0.78.

REFERENCES

[1] Tweets per day. Available online: https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html (accessed on Nov 13, 2023)

[2] Fouad, M.M., Gharib, T.F., Mashat, A.S. Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble. Advances in Intelligent Systems and Computing, AMLTA 2018, vol 723; pp. 516-527.

[3] Adwan, O. Y., Al-Tawil, M., Huneiti, A., Shahin, R., Abu Zayed, A., & Al-Dibsi, R. Twitter Sentiment Analysis Approaches: A Survey. International Journal of Emerging Technologies in Learning (iJET) 2020. 15(15); pp. 79–93.

[4] Twitter Sentiment Analysis With Python | Introduction & Techniques. Available online: https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/ (accessed on Nov 13, 2023)

[5] Ahuja, S., Dubey G. Clustering and sentiment analysis on Twitter data. 2nd International Conference on Telecommunication and Networks (TEL-NET) (2017); pp. 1-5.

[6] K. Broni, Global emoji usage on twitter. URL https://blog.emojipedia.org/10-years-ofemojipedia-10-years-of-record-breaking-emojipopularity/ (accessed on Nov. 13, 2023)

[7] Tweepy Documentation. Available online: https://docs.tweepy.org/en/stable/ (accessed on Nov 13, 2023)

[8] Langdetect Python Package. 2021. Available online: https://pypi.org/project/langdetect/ (accessed on Nov 13, 2023)

[9] 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., Duchesnay, E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011, vol 12; pp. 2825-2830.

[10] Emoji Python Package. Available online: https://pypi.org/project/emoji/ (accessed on Nov 13, 2023)

[11] Hutto, C., & Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media 2014, 8(1), pp. 216-225.

[12] Google Colaboratory, Release 2023-11-08. Available online: https://colab.research.google.com (accessed on Nov 13, 2023)

[13] Twitter Developer Portal. Available online: https://developer.twitter.com (accessed on Nov 13, 2023