Original Research

COVID-19 Vaccine –Public Sentiment Analysis Using Python's Textblob Approach

P. Sivalakshmi[1], P. Udhaya Kumar[2], M. Vasanth[2], R. Srinath[2], M. Yokesh[1]

[1]Assistant Professor, Department of ECE, RMK College of Engineering and Technology, RSM Nagar, Gummidipoondi Taluk, Puduvoyal, Thiruvallur, Tamil Nadu 601206 India; [2]RSM Nagar, Gummidipoondi Taluk, Puduvoyal, Thiruvallur, Tamil Nadu 601206 India; [2]B.E Student, Department of ECE, RMK College of Engineering and Technology, RSM Nagar, Gummidipoondi Taluk, Puduvoyal, Thiruvallur, Tamil Nadu 601206 India.

ABSTRACT

Background: In the present scenario, weeks over social media creates a high impact on the verdict of individuals& organizations. Opinions in the form of tweets reflect one's attitude and emotions towards a specific person or an event. Also, Companies can benefit from this massive platform by collecting data related to opinions on them.

Objective: To infer the public opinion towards the tag Covid-19 Vaccine. Which is one of the natural language processing.

Methods: The TextBlob approach is used to extract emotions and visualize them from the raw data collected from Twitter. Initially, tweets were collected on Covid -19 Vaccine after preprocessing the collected data set, the TextBlob approach classifies polarity of textual data in positive, strongly positive, weakly positive, neutral, negative, strongly negative & weakly negative categories.

Results: The sentiment scores for collected tweets is calculated and shown under the results section. Which projects the emotions of all the people using social media towards covid-19 vaccination.

Conclusion: There are surplus opportunities in future for exploring trend sentiments over some time. Also analysis over the different location of the world. Based on which necessary measures could be taken by the government or any organizations to create positivity among the public.

Key Words: Emotion analysis, Natural Language Processing, Social media, TexBlob

INTRODUCTION

People started sharing different varieties of emotions very first in the year 1970.[1,2] Thereafter, people started analyzing that information in bits and pieces, has various application in turns like predicting the election results, realizing the society's attitudes towards an event or a specific person. It has been shown that how social media expresses collective wisdom which, when properly used, can yield and accurate predictions over any issues.[3] Meanwhile, the advent of technologies these obtained more popularity due to their high accessibility.[1] Now a days, people use social media sites to share opinion. Twitter is one such platform in which users send, read post known as tweets and interact with different community people. Users exchange their opinion on daily lives, brands and places. Sentiment analysis, which is called opinion mining, is a field of study that analyses the opinions, evaluations, attitudes, and feelings of individuals according to entities such as products, services, organizations, individ-

uals, events, issues, and their characteristics.[4] Beliefs are important almost in every human activity. At present, any small scale or large scale businesses and organizations want to find out consumer's viewpoints of their products and services. On the ballot, voters want to know other's opinions about the candidates before they vote. In the past, individuals used to ask their family and friends about their viewpoints. Small & large scale businesses and organizations used to post a questionnaire when they want to know about the individual's opinion.[4]Business marketing, public relations, and political advertising companies for a long time with an increase in usage of the social networks, individuals and organizations increasingly use the content on these networks and make decisions.[4] The EAS system had been designed to extract emotions in social network data. To check the performance of the designed system, the data of Twitter gathered in the range of the Iran presidential election in 2017.[5,6] Amongst the earlier social networks studies, we chose Twitter because of its importance.[7] For fetching tweets data from Twitter. Initially, an

API request was made to Twitter which was later approved. Afterwards, as explained by Shah et al. python's tweepy library which is specifically developed for retrieving tweets data from Twitter is used along with Twitter streaming API and authentication keys (Consumer key, consumer token key, access token and access token secret.) provided by twitter.[8,9] For authentication purpose, the Tweepy library used the AuthHandler function for verification of authentication keys. Once, the authentication request is approved it starts fetching the tweets. Then panda's library was used to fetch all the collected tweets in a data frame. Next, pre-processing of a tweet using a regular expression library performed through which unnecessary data removed from tweets. In the last stage of pre-processing the textual data changed into numerical form named count vectorization. It is a type of encoding that helps in performing tokenization. Which involves crucial steps performed as a part of natural language processing (NLP). Through which sentences were tokenized into tokens of each word to form a feature set. Once the dataset pre-processed, in the next stage sentiment scores were calculated using Python's TextBlob Library, the same discussed in detail in further sections.

RELATED WORK

A system designed for real-time visualization of Twitter microblogs and their analysis. For offering enhanced semantic insights, a weighted tag network has been designed.[8] There is a system to find semantic patterns through heterogeneous data and without any social network structure like Instagram to detect events.[9] Not only Instagram users' posts, but also the combination of Instagram and Twitter users' posts employed to improve event detection quality.[10] The numbers of four comment categories of Trump and Clinton, supporters and opponents have been studied.[11] Usually classification accomplished by hashtags. The study made one day earlier U.S Presidential ballot demonstrated that Clinton's supporters were more than trump. Since 60% of Instagram users were in the age between 18 and 35, their prediction was different from the real results. This shows clearly the inability to predict the result of Instagram to the whole society. Mohamed studied the Malaysian politician's storytelling.[11,12] In this, all the post of politicians, including video, picture, and text content, have been analyzed. The post is divided into six categories, and the results are compared to each other. Recently, it has been shown the high correlation between different Indonesian party influencers and their presidential candidates on Twitter.[13] Emotion analysis using text processing techniques also applied on the social network in, city event detection using expandable in initial event keywords has been studied.[14-16] In this study, it found that Twitter is much better than Instagram in analyzing & predicting city events.[17]

Sentiment Classification Using Natural Language Processing (NLP)

As articulated by Lobur, natural language processing (NLP) is the domain in machine learning which is used in analytics.[18] NLTK which is called the Natural language toolkit is a part of python's library belonging to natural language processing. Natural language processing not only deals with text analytics but also plays an important part with research based on analysis of human languages. Preparing models for research based on human languages comes in computational linguistics. The major advantage of using NLTK is that it allows even a beginner programmer to understand concepts of natural language processing saving a lot of time from gathering information about it. Numerous advantage associated with using NLTK is it contains 60 corpora belonging to real-world data, collections of grammar, models which have been trained, functions which provide a path for performing general natural language processing tasks. The corpora used in the Natural language toolkit (NLTK) are generally divided into different categories for assisting its users. Though in other programming languages, natural language processing task can be accomplished. The major points which take python apart from other languages is a follows:

- Better reading ability.
- User-friendly object-oriented technique.
- Ease of extensibility.
- Better Unicode assistance.
- A functionality-rich library.

NLTK has a vast source of libraries that are being updated with new functionalities over the period. This paper has provided a deep understanding of the functioning of the NLTK library. Tasks such as summarization of text, extraction of information, machine translation are performed by Natural Language Process as depicted in work from Zitnik.[20] here, the author has carried out sentiment analysis using natural language processing toolkit to detect the language of the text and extract meaning out of it. For it, first, the language dataset has been cleaned in the pre-processing stage which was then followed by language detection and evaluation of the result. Though, this natural language toolkit library. The major limitation of this work is that it does not compares the performance of this library with other natural language toolkit libraries. Though, as an advantage, this library can be used for natural language processing courses for educational purposes. Moreover, this work has provided an understanding of the use of natural language processing in language detection which is used in this research article.

Feature Extraction for Sentiment Classification

As stated by Zhang, Jin and Zhon in their work that one of the most important models utilized for the categorization of the object is Bag of Words (BoW).[21] The concept behind the Bow model is forming visual words by quantizing every extracted

key point. After this, each picture is shown using visual words histogram. Joachims also worked upon the BoW model.[23] He showed that the BoW model depicts the count of every word present in textual data. Ma et al.[24] showed that a matrix depicting the count of words in textual data is created in the BoW model. Afterwards, the frequency of occurrence of these words is used as features to train the classifier. Thang Luong conducted a research where it is observed that the BoW model performed considerably well in compassion with other models on Chinese English language translation data.[22] All these works have helped in understanding the concept behind BoW Model for feature extraction. Janani, emphasized various steps being taken while preprocessing the dataset.[23] Various steps which were taken for the pre-processing dataset are stopped words removal, determination of sentence boundary, tokenization and stemming. Tokenization is one of the most important steps while pre-processing a dataset. It works in a manner that textual data is divided into small tokens. Each token represents a word from the textual document or language.

There are numerous libraries available in python such as NLTK word tokenize, Mila tokenizer, TextBlob tokenizer etc. which are used for tokenization. This work has helped in understanding the in-depth functioning of the Text Blob library for pre-processing phase.

TextBlob Approach Algorithm for Sentiment Classification

One of python's libraries that use API for accessing methods to perform Natural language processing is called as Text-Blob. A common challenge for work based on sentiment analysis is missing spelt words. This problem is addressed by Manushree, Adarsh and Kumar here, authors have compared TextBlob and SentiWordNet approach. Firstly, the dataset was pre-processed by removing stop words and unrequired data which could result in added computational cost in the performance of models.[24,25] It was followed by aspect selection and based on it sentence extraction was done. Both the models were then used to calculate sentiment polarity and categorize the reviews into positive, negative and neutral categories. This work just focused on sentiment analysis of miss spelt words in the English language. The advantage of this work was that it performed sentiment analysis on miss spelt words in the English language. However, the limitation of this work is that it was unable to perform sentiment analysis on miss spelt words in other language using TextBlob.

Moreover, this work has helped in-depth understanding regarding the implementation of the TextBlob approach for the research project. The pre-processing of the textual data is of very importance in sentiment analysis as it reduces the size of textual data which is given as input to the model. Various steps are followed while pre-processing the textual data. The various pre-processing tasks performed for cleaning textual data are the determination of boundary of sentences, removal

of stop words from natural language, stemming and tokenization. Tokenization involves splitting a sentence into tokens of each word belonging to the respective sentences. Janani et al.[25] carried out work on certain tokenize and read the tokenized words. The advantage of this work is that it compared various good tokenization tools and it distinguished TextBlob from them. However, it was unable to read tokenized special characters which turned out as its limitation. This work has helped in understanding the major limitation of TextBlob.[26]

Python's Regular Expression (REGEX/RE) Library

Stolee, concentrated mainly on regex, which is also called a regular expression, is a reflection of specific words search which helps in the identification of text through recognition of patterns in place of exact strings. REGEX library is commonly utilized for parsing textual data belong to general language. Regex is also called Python's module. Even though regex is considered a versatile and powerful library it could be difficult to understand, this is one of its limitations. According to Spishak, Dietl and Ernst, The major advantage of python's regular expression library is that it has a variety of applications as it has a powerful ability to fetch meaningful information from the given sentence.[27] Regular expression is applicable in per-processing the data, MY QL injection, generation of test cases and intrusion detection in networks etc. According to Ganesh and Yeole, the major advantage of a regular expression library is that it has fast processing speed in terms of code execution, and it has very compressed code which reduces efforts of writing long codes for pre-processing of the dataset.[28,29] Advantage fast processing and compresses coding.

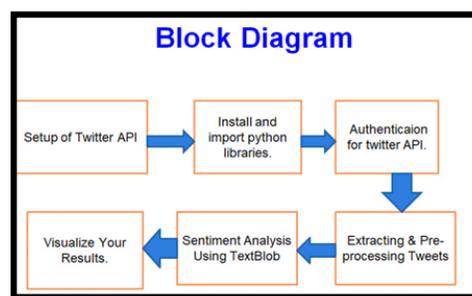## ANALYSIS BASED ON THE LITERATURE REVIEW:

Figure 1: Block Diagram of the analysis to be performed over tweets.

Step: 1 Setup of Twitter Application Programming Interface

To utilize Twitter API, we first need a Twitter Developer Application Page, to create a Developer account. After the application

is approved, head to Developer Dashboard, Click on 'Projects & Apps' -> 'Overview'(in that section) click the 'New Project' button to create a new project.[30] Then again return to overview click on the Key icon to access the Keys and Tokens of the new project, upon clicking generate /regenerate we will have API Key, API Secret, Access Token & Access Token Secret to be popup, saved them for later use (**Figure 1-4**)
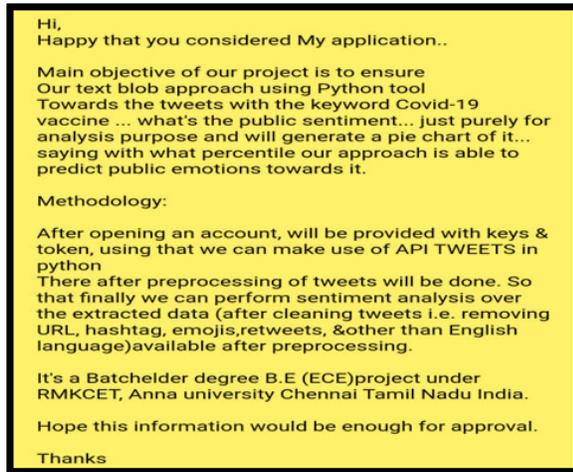

Figure 2: Requisition Letter to Twitter, to create a developer portal.

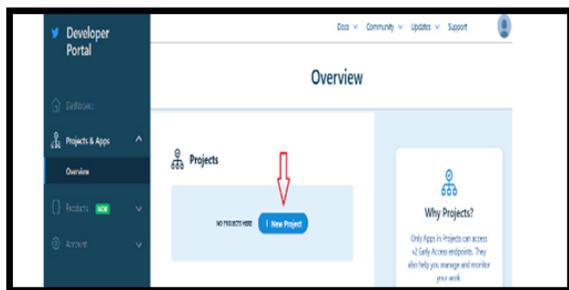**Figure 3:** Approval Letter from Twitter, to create a developer portal.

Figure 4: Developer Portal to fetch tweets.

Step:2 Extracting & Pre-processing Tweets

Install and Import Libraries Before analysis, you need to install TextBlob and tweepy libraries using! Pip install command. Using the accessing credentials-4 codes generated earlier, we can set up the Twitter API authentication.

api_key = 'your_api_key_here'

api_key_secret = 'your_api_key_secret_here'

access_token = 'your_access_token_here'

access_token_secret = 'your_access_token_secret_here'

auth = tweepy.OAuthHandler(api_key, api_key_secret)

auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

Step 3: Tweets Extraction

Key Considerations:

1. Rate limit of 900 API calls every 15 minutes
2. Twitter allows tweet extraction of the past 7 days alone. So need to extract each day for the record of the last 7 days.
3. The search term should be COVID -19 Covid vaccine.
4. Excluded retweets &tweets other than English.

from tweepy import

import pandas as pd

import csv

import re

import string

import preprocessor as p

consumer_key = <enter your consumer key>

consumer_secret = <enter key>

access_key= <enter key>

access_secret = <enter key>

auth=tweepy.OAuthHandler(consumer_key, consumer_secret)

auth.set_access_token(access_key, access_secret)

api = tweepy.API(auth,wait_on_rate_limit=True)

csvFile = open('file-name', 'a')

csvWriter = csv.writer(csvFile)

search_words = "#"   # enter your words

new_search = search_words + " -filter:retweets"

*for tweet in tweepy.Cursor(api.search,q=new_*

```
search,count=100,
        lang=”en”,
    since_id=0).items():
```

csvWriter.writerow([tweet.created_at, tweet.text. encode('utf-8'),tweet.user.screen_name.encode('utf-8'), tweet.user.location.encode('utf-8')])

The output of the above code is a csv file.

Step 4: Tweet Pre-Processing

import preprocessor as p

 # Clean tweet text with tweet-preprocessor

tweets_df['text_cleaned'] = tweets_df['text'].apply(lambda x: p.clean(x))

By using this tweet preprocessor, we can remove URL's, Hashtag, Emoji's, reserved words and mentions in the tweets.

Step 5: Sentiment Analysis Using TextBlob

from textblob import TextBlob

 # Obtain polarity scores generated by TextBlob

tweets_df['textblob_score'] = tweets_df['text_cleaned]. apply(lambda x: TextBlob(x).sentiment.polarity)

 # Set threshold to define neutral sentiment

neutral_thresh = 0.05

 # Convert polarity score into sentiment categories

tweets_df['textblob_sentiment'] = tweets_df['textblob_ score'].apply(lambda c: 'Positive' if c >= neutral_thresh else ('Negative' if c <= -(neutral_thresh) else 'Neutral'))

Step 6:Visualize Your Results.

Using Pie-Chart will have the output of the analyzed tweets.

RESULTS AND DISCUSSION

Results on Covid-19 Vaccine (20 tweets)

The below mentioned figure represents the reaction of people towards the vaccination campaign for Covid 19, the initial analysis takes into account the analysis results from 20 tweet reaction, as observed it's clear that around 13 tweets i.e. around 65% of the people reacted neutral indicated by the yellow portion of the pie chart, followed by 4 people accounting to 20% of the total had weakly positive mindset indicating that they were in an confusion to either to be on the positive side or to have neutral opinion indicated by the light green portion of the pie chart and also it can be observed that around 10% of the people reacted positively and another feeble percentage of people accounting to 5% of the total tweet had a weakly positive opinion. The overall results

are deliberately indicating that majority of the opinions were on the neutral side. These results were obtained with total runtime of around 12sec (**Figure 5**).
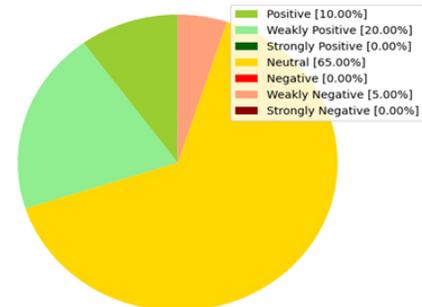

Figure 5: Results on Covid-19 Vaccine (20 tweets).

Covid-19 Vaccine Results for 500 Tweets

Widening the survey range to 500, the results followed up have been indicated in the below pie chart, still results indicated a neutral opinion from most of the people accounting to 56.80% of the survey count, followed up by around 22.60% having a weakly positive opinion, these people accounted the majority of total survey count. The positive tweets arose from around 9.40% of the people. The rest of the opinions accounted for only a minimal lest off a percentage of the total survey count. These results were obtained with total runtime of around 28 sec (**Figure 6**).
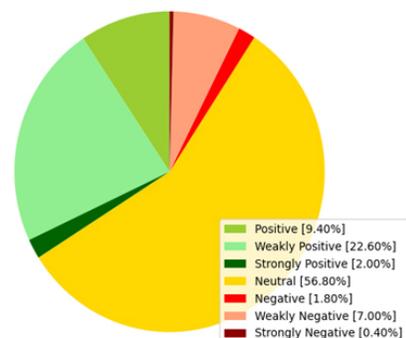

Figure 6: Results on Covid-19 Vaccine (500 tweets).

Covid-19 Vaccine Results for 1000 Tweets

The final survey results were taken accounting for the opinion that arose from a total of 1000 tweets, which were no different from the former results, and had the neutral reaction from most the people accounting for 54.20% of the total survey count, followed by weakly positive opinion from 23.40% of the people, followed by positive, weakly negative, strongly positive and negative accounting to the most

feeble amount of the total. These results were obtained with total runtime of around 52 sec (**Figure 7**).
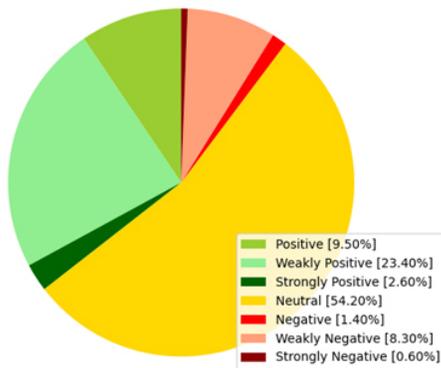

Figure 7: Results on Covid-19 Vaccine (1000 tweets).

Investigated Result of 1000 Tweets

Since all the above analysis shows not much of change in neutral polarity. We had an idea of analyzing with another tag of words to ensure the working of the Programme in all polarities. Another most commented sector is food, from reviews to opinions, the following pie chart represents the results we got from surveying around 1000 tweets related to the opinion of the people reacting to the McDonalds chicken, the results conveyed the follows, mostly there was an equal distribution of people liking and disliking to the tasty of the particular recipe, the pie chart indicated a majority of the distribution of people reacted with the weakly negative and strongly negative opinion that accounted to around 62% of the total survey count, the negative reactions included up to 18.40%, followed up by weakly positive people accounting to 13.60% and rest of the feeble count being distributed among positive and strongly positive. The pie chart strongly indicates that there was a negative opinion by most of the people showing an unwillingness to the recipe. These results were obtained with a total runtime of around 50 sec (**Figure 8**).
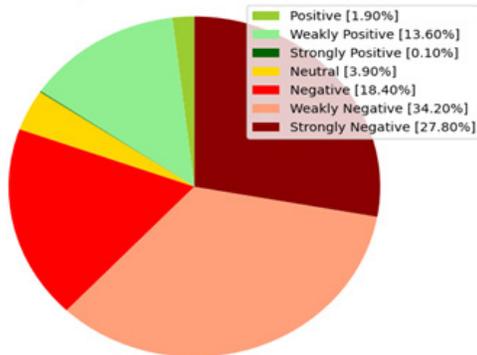

Figure 8: Results on McDonald Chicken (1000 tweets).

Similar to the previous one, KFC being a rival to McDonald's for several decades, we tested for KFC to compare both the results for 1000 tweets. These results though showed a surge of opinion towards positive, thereby account for around 30.80% of the whole survey count, all though there were equally negative opinions. These results were obtained with total runtime of around 55 sec (**Figure 9**).
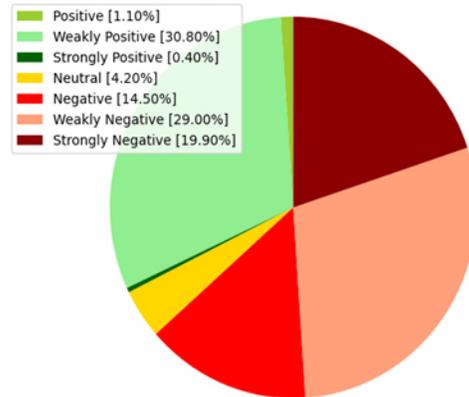

Figure 9: Results on KFC Chicken (1000 tweets).

Comparison of Two previous Outputs

By comparing the above previous results we can observe that one is better than the other based on people's opinion. Similarly, we can compare various aspects using this algorithm. Thereby ensuring the change in neutral polarity. So basically we can conclude that around 35% of people are feeling positive, around 10% of people feeling negative & around 54% of people having neutral emotion on Covid-19 Vaccine. Through the investigated result, we can conclude that more than 60% of people have negative opinion or sentiment on KFC & around 80% of people using this social platform has negative emotion/opinion or sentiment on McDonald's. Strong positive emotions were found less than 1% reflecting the health consciousness of the public.

CONCLUSION

In this article, we have analyzed Covid-19 Covid vaccine tweets via Twitter API and implemented them using TextBlob. Likewise, there are surplus opportunities in future for exploring trend sentiments over some time. Also analysis over the different location of the world. Based on which necessary measures could be taken by the government or any organizations to create positivity among the public.

DECLARATION OF INTEREST: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Kiaei SF, Dehghan RM, Farzi Sa. Designing and Implementing an Emotion Analytic System (EAS) on Instagram Social Network Data. Int J Web Res. 2019;2(2):9-14.
2. Boyd DM, Ellison N. Social Network Sites: Definition, History, and Scholarship. J Comp-Mediated Comm. 2007;13(1):210-230.
3. Asur S. Huberman, BA. Predicting the Future with Social Media. Proceedings of 2010 IEEE/WIC/ACM Int Conf on Web Intelligence and Intelligent Agent Tech. 2010;31:492-499.
4. Martinez-Pecino R, Garcia-Gavilán M. Likes and Problematic Instagram Use: The Moderating Role of Self-Esteem. Cyberps Behav Soc Netw. 2019;22(6):412-416.
5. Liu B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. 2012;5: 362-368.
6. Li X, Peng Q, Sun Z, Chai L, Wang Y. Predicting Social Emotions from Readers' Perspective. IEEE Trans on Affective Comp. 2019;10 (2):255-264.
7. Rao Y, Li Q, Wenyin L, Wu Q, Quan X. Affective topic model for social emotion detection. Neural Netw. 2014;58: 29-37.
8. Dhar V. Big Data and Predictive Analytics in Health Care. Big Data. 2014;2(3):113-116.
9. Tsou M. Research challenges and opportunities in mapping social media and Big Data. Cartograp Geograp Inf Sci. 2015; 42(sup1):70-74.
10. Al-Bahrani, Abdullah A, Patel D. Incorporating Twitter, Instagram, and Facebook in Economics Classrooms. J Econ Educ. 2014;46(1):56-67,
11. Schmidbauer H, Rösch A, Stieler F. The 2016 US presidential election and media on Instagram: Who was in the lead? Compt Human Behav. 2018;81:148-160.
12. Mohamed S. Instagram and Political Storytelling among Malaysian Politicians during the 14th General Election. Malaysian J Comm. 2019; 35(3):353-371.
13. Dhanesh G, Duthler G. Relationship management through social media influencers: Effects of followers' awareness of paid endorsement. Pub Rel Rev. 2019; 45(3):101765.
14. Mozafari F, Tahayori H. Emotion Detection by Using Similarity Techniques. 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems. 2019;12:1-5.
15. Ghanbari-Adivi, F& Mosleh, M. Text emotion detection in social networks using a novel ensemble classifier based on Parzen Tree Estimator (TPE). Neural Comput Appl. 2019;7:1-13.
16. Liu S, Jansson P. City event detection from social media with neural embeddings and topic model visualization. Int Conf on Big Data. 2017;4:4111-4116,
17. Chaplin Xia, Raz Schwartz, Ke Xie, Adam Krebs, Andrew Langdon, Jeremy Ting, and Mor Naaman. CityBeat: real-time social media visualization of hyper-local city-data. In Proceedings of the 23rd Int Conf on World Wide Web (WWW '14 Companion).2014 (167–170).
18. Lobur, MA. Romanyuk F, Romanyshyn M.Using NLTK for educational and scientific purposes. 11ᵗʰ Int Conf The Experience of Designing and Application of CAD Systems in Microelectronics. 2011: 426-428.
19. Shah B, Agarwal V, Dubey D, Correia S, Twitter Analysis for Disaster Management. Fourth Int Conf on Compt Comm Cont Autom.2018:1-4.
20. S. Žitnik, D. Draskovic, B. Nikolić and M. Bajec, nutIE- A modern open-source natural language processing toolkit. 25th Telecommunication Forum (TELFOR).2017:1-4.
21. Zhang, YJ, Rong AS, Zhou, ZH. Understanding bag-of-words model: A statistical framework. Int J Mach Learn Cybern. 2010;1:43-52.
22. Christopher MT, Manning D. Effective Approaches to Attention-based Neural Machine Translation. arXiv: 1508.04025 v5 [cs. CL]. 2015 Sep.
23. Joachims T. Text categorization with support vector machines: Learning with many relevant features. In European Conf Mach Learning. 1998;21:137-142.
24. Ma, S., Sun, X., Wang, Y. and Lin, J. Bag-of-Words as Target for Neural Machine Translation. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.2018 (2);332–338.
25. S.Vijayarani & R.Janani.Text Mining: open Source Tokenization Tools. An Analysis. Advanced Computational Intelligence: An Int J.2016 3(1):.37-47
26. Tanushree AM, Adarsh MJ, Kumar PR. A comparative method for different aspect based products features in online reviews of different languages. 2nd IEEE Int Conf on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017;12:1836-1840.
27. Eric Spishak, Werner Dietl, and Michael D. Ernst. A type system for regular expressions. In Proceedings of the 14th Workshop on Formal Techniques for Java-like Programs 2012:20–26.
28. Adam Kiezun, Vijay Ganesh, Shay Artzi, Philip J. Guo, Pieter Hooimeijer, and Michael D. Ernst. HAMPI: A solver for word equations over strings, regular expressions, and context-free grammars. ACM Trans. Softw. Eng. Methodol. 2013:21(4):1103-1128.
29. Yeole AS, Meshram BB. Analysis of different technique for detection of SQL injection. In Proceedings of the Int Conf & Workshop on Emerging Trends. 2011:963–966.
30. Twitter API Documentation [Internet]. Developer.twitter.com. 2021 [cited 5 May 2021]. https://developer.twitter.com/en/docs/twitter-api