Advancement of artificial intelligence techniques based lexicon emotion analysis for vaccine of COVID-19

Mohammed Rasool jawad[1], Maysaloon Abed qasim[2], Gerardo Cazzato[3], Musadaq Mahir Abdel Zahra[4a,b]
Prabhakara Rao Kapula[5], Noreddine Gherabi[6], Refed Adnan Jaleel[7]

[1] dept. of Network technology and information systems, University of Babylon, Babylon, Iraq
[2] Technical Engineering Collage, Northern Technical University (NTU), Mosul, Iraq
[3]dept. of Emergency and Organ Transplantation (DETO), University of Bari "Aldo Moro", Bari, 70124, Italy
[4a]Computer Techniques Engineering Department, Al-Mustaqbal University College, Hillah 51001, Iraq
[4b]Electrical Engineering Department, College of Engineering, University of Babylon, Hilla, Babil, Iraq
[5]Professor, dept. of ECE, B V Raju Institute of Technology, Narsapur. Telangana State
[6]Sultan Moulay Slimane University, ENSA Khouribga
[7]dept. of Information and Communication Engineering, Al-Nahrain University, Baghdad, Iraq

ABSTRACT

Emotions are a vital and fundamental part of life. Everything we do, say, or do not say, somehow reflects some of our feelings, perhaps not immediately. To analyze a human's most fundamental behavior, we must examine these feelings using emotional data, also known as affect data. Text, voice, and other types of data can be used. Affective Computing, which uses this emotional data to analyze emotions, is a scientific fields. Emotion computation is a difficult task; significant progress has been made, but there is still scope for improvement. With the introduction of social networking sites, it is now possible to connect with people from all over the world. Many people are attracted to examining the text available on these various social websites. Analyzing this data through the Internet means we're exploring the entire continent, taking in all of the communities and cultures along the way. This paper analyze text emotion of Iraqi people about COVID-19 using data collected from twitter, People's opinions can be classified based on lexicon into different separate classifications of feelings (anticipation, anger, trust, fear, sadness, surprise, disgust, and joy) as well as two distinct emotions (positive and negative), which can then be visualized using charts to find the most prevalent emotion using lexicon-based analysis.

Keywords:	Machine learning, Sentiment analysis, Emotion analysis, Natural language processing NLP, API

Corresponding Author:

Refed Adnan Jaleel
Information and Communication Engineering
Al-Nahrain University
Baghdad, Iraq
E-mail: Iraq_it_2010@yahoo.com

1. Introduction

A series of security incidents have recently happened around the world, demonstrating the wide range of crises in which today's ordinary people are effectively using their mobile communication devices [1]. The majority of significant news events now include real-time social media comments [2]. Social networks have emerged as a research topic in which experts from all backgrounds seek inspiration. As a matter of fact, social networks, particularly social network analysis (SNA), which is backed by computer science, offer the opportunity to widen other fields of knowledge. Many fields have established the notion of social network and social network analysis, including scientist or other professions cooperation networks, family networks, student friendship networks, company director networks, consumer networks, labor market, public health, psychology, and so on. It has recently become a part of a new branch of science known as computational

social science [3]. Social networks are characterized by social scientists as a group of people who share a common interest, such as familial relationships, political action, information, views, or geographic location [4]. Microblogging is defined as the activity of posting small amounts of digital content, such as text, images, links, small videos, or any other type of media, to the internet. They, like other social networking sites, attempt to build a sense of online community. These platforms let users to exchange information about their lives, activities, opinions, and status in a light-weight, simple manner. Twitter is one of the most widely used microblogging sites [5]. Recent events have brought Twitter to the fore as a new media to study. Users can follow or be followed on Twitter. The relationship of following and being followed on does not need reciprocation like most online social networking sites, such as Facebook or MySpace,. A user can follow any other user without having to follow them back [6].

2. Artificial intelligence (AI)

Several computer systems have been constructed over the last few decades to do numerous human mental functions such as arithmetic, designing computer programs, and interpreting languages, all of which are thought to require "intelligence." Some computer systems are able to analyze electronic circuits, solving formulas, diagnosing diseases, and comprehending a limited amount of human speech and natural language texts. The majority of the work in developing these systems has been done in the field of "AI" [7]. AI is the branch of engineering and science related to the theory and practice of creating systems that possess the characteristics we associate with intelligence in human behavior, such as NLP, perception, issue planning and solving, adaptation & learning, and environmental action [8]. The AI field of NLP is a branch of computer science. It is related with how computers interact with human natural languages, and in particular with programming computers to analyze huge amounts of natural language data [9].

3. Social media

Any medium of communication that allows for two-way engagement is referred to as social media (SM). It is a allows users to share and consume content in a variety of formats, including text, image, and video. People utilize social media in their daily lives in a variety of ways, from text messaging to online dating. Users of social media can contact with friends, family, and organizations all around the world using interactive services [10]. Users of these services have decided to form type of virtual society known as online social networks (OSN). They're also known as virtual communities [3]. These social networks have grown in popularity in recent years, offering a more efficient and user-friendly method to maintain social connections and communicate information in a variety of formats and mediums, including microblogging, status updates, mobile text alerts, blogs, instant messaging, and forums. Microblogging is defined as the activity of posting little amount of online content to the internet, which can take the shape of text, images, links, brief videos, or any other form of media. Microblogging has become extremely popular among groups of friends and professional colleagues who often update their material and follow one other's updates . Because they are brief and easy to analyze, this style of blogging is seen to be more informative and accurate for marketers. Twitter, Jaiku, and Pownce are just a few of the services that provide microblogging. These platforms let users to exchange information about their lives, activities, opinions, and status in a light-weight, simple manner. Twitter is one of the most widely used microblogging sites [5]. Twitter keeps track of the most popular phrases, words, and hashtags and posts them under the heading "trending topics" on a regular basis. A hashtag is a Twitter protocol for starting and following a discussion thread by prefixing a word with the '#' character. Twitter displays a list of the top ten trending topics on every user's homepage in a right sidebar [6]. This data is open to the public. As a result, it can be used as raw data primarily for opinion extraction, customer satisfaction analysis, and grading alternative government schemes, as well as sentiment analysis [11].

3.1. Twitter application programming interfaces (APIs)

APIs are the modern software systems' electrical sockets. It defines how software components should communicate with one another. This interface, at a high level, contains a list of commands that a first component can use to access functionality in a second component, as well as the particular format in which the first component should provide those commands to the second component. The user can see some program components, such as the user interface of a web browser. There are many more components that are hidden but serve important duties. Various software components, for example, are in charge of delivering and receiving web page data over the Internet, interpreting that data and rendering it in a graphical style, and

handling persistent data (e.g., browser cookies) kept by websites. The relationship between these components is defined via APIs [12]. There are two forms of API sorting available on Twitter. Developers can read and write Twitter data using the REST-APIs. These APIs are useful for researchers because they allow them to search for messages that have been posted recently and meet certain criteria, such as the inclusion of specific keywords, hashtags, or user names. Developers can use the streaming APIs to access Twitter's global data stream. These APIs are useful for academics since they enable for the real-time capture of data matching certain criteria. Researchers must first construct a Twitter application that handles requests to Twitter's database in order to gain access to these APIs. To obtain the login details required to access data using Twitter's API, you must first create a Twitter application. The steps for establishing a Twitter application are outlined below. It must first have a Twitter account that is active. Then go to Twitter's app registration page and follow the instructions there. Following the creation of the application, it must generate four tokens that will allow the scripts to collect data on Twitter. Access token, API-secret, API-key, and Access token secret are the four. These keys are extremely important, and the user should treat them as if they were email passwords. Any use of these credentials to access Twitter's databases can be traced [13].

3.2. Gathering data from twitter

To retrieve data from Twitter, a user must first have access to the Twitter API. Twitter may alter the procedure for granting API access to users [14]. The OAuth package in R is used to perform Twitter API authentication. The processes for using OAuth to access the Twitter API are shown in Figure 1. To use Twitter API, you'll need to establish a Twitter application.
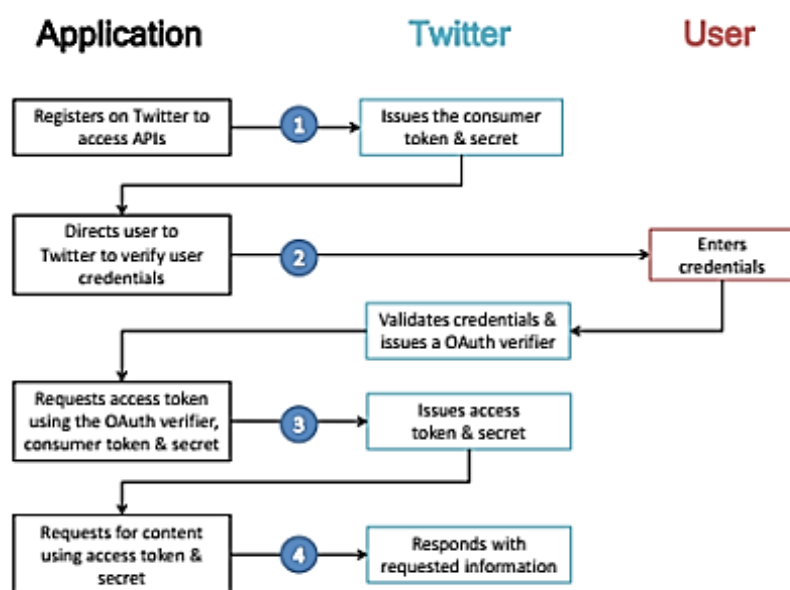
Figure 1. The architecture of gathering tweets from twitter

- Consumers ought to sign up with Twitter. It continues to provide the consumer with a key and secret key that can be used to authenticate the application.

- These keys will be used to construct a Twitter link that will start the authentication procedure. Twitter verifies the user's identification and issues PIN 3 (also known as a verifier). This pin must be provided by the user to the application.

- The next application step uses this PIN to obtain from the Twitter API an Access Secret and Token that are unique to the user.

- The information about the token and secret key is cached for future use. GetUserAccessKeySecret can be used to accomplish this.

3.3. Gathered data preprocessing

By reducing data errors, preprocessing the data can improve sentiment analysis. It's a method of removing undesirable parts from data. Sentiment analysis methods that do not use data preprocessing may miss crucial terms, reducing the accuracy of the results. Preprocessing, on the other hand, can result in the loss of critical information. The elimination of punctuation, which could be valuable to the analysis, is one example of erroneous over-preprocessing. There are many different types of general preprocessing procedures which are:

1) Filtering: Non Arabic tweets are removed using the filtering method. The filter () function in the dplyr package in the Rstudio environment is used to subset a data frame, keeping the rows that satisfy the constraints. It works with both grouped and ungrouped data; most data operations are performed on variables-defined groups. group by() turns an existing table into a grouped table in which actions are carried out "by group." The function ungroup () eliminates grouping. On ungrouped data, filtering is frequently more faster. Simple typing errors, such as repeated letters and misspellings, are corrected by the filtering mechanism. It used dictionaries in this operation. Predefined dictionary words replace acronyms and abbreviations. The following are some examples of simple forms of errors:

a) Errors created by the sound of the language, such as "ظلام" can be transcribed as "ضلام" and errors generated by switching letters, such as "بيت" can be written as "بيت"

b) Eliminate the repetition of characters, such as " كثييييييير " being replaced by " كثير " by deleting the vowels repetition.

c) Correct spelling errors such as:

- Issues with the space bar; either no space (as in " كيفحالك "), or incorrect space (as in "مر حبا").
- Letter closeness; for example, the word "كنت" might be written as "منت" because the letters and are close to each other on the keyboard.
- The confusion produced by similar characters, such as the words "كتير" and "كثير" [15].

2) Removing URL's: Links to other webpages and websites can be found in URLs. During the analysis, they supply no information. These were of no use, thus they were eliminated from all tweets using R's tm_map() function. Blank spaces were used to replace all sentences and subparts of sentences that began with http [16].

3) Removing emoticons, numbers, and punctuations: The different emoticons punctuations in the tweets are removed in this phase [17]. One might wonder why emoticons were removed. It was removed because when the emoticons were retrieved, they appeared as square boxes rather than genuine emoticons. The "gsub" function in R was used to delete the unwanted emoticon values [16].

4) Removing stopwords: It's a method for removing commonly used terms that are nonsensical and unhelpful for text classification. This decreases the size of the corpus without sacrificing crucial information [18].

5) Stemming: Stemming is a necessity for many Natural Language Processing operations. In most Information Retrieval systems, it is critical. The basic goal of stemming is to reduce a word's various grammatical forms, such as its noun, adjective, verb, adverb, and so on, to its root form. The purpose of stemming is to reduce a word's inflectional forms and, in some cases, derivationally related forms to a common base form. R's "stem" function is used to complete this task [19].

3.4. Sentiment and emotion analysis

Sentiment analysis is a procedure that applies NLP to automate the extraction of opinions, attitudes and emotions from text, audio, perspectives, tweets, and sources of DB. Subjectivity analysis, opinion mining, and assessment extraction are other terms for It [20]. A sentiment classifier can detect whether a sentence has a positive or negative connotation by determining its polarity. A general opinion on a topic can be derived by averaging the polarity of individual texts given a sample of texts that discuss the same issue. For example, a consensus view on a product can be determined by gathering a collection of reviews, i.e., whether the product is popular with consumers or not. The state-of-the-art in sentiment categorization is generally split into two approaches, one lexicon-based and the other learning-based classifier [21]. To discover the emotions contained in a text, emotion analysis employs NLP, analysis of text, and a variety of techniques for computational. This analysis can be done on a number of levels, including document, sentence, word, and

aspect levels. The basic two important approaches in sentiment analysis for classifying emotions are: emotional dimensions and emotional categories [22].As shown in Figure 2, the analysis of emotion of certain input data consists of the following steps:
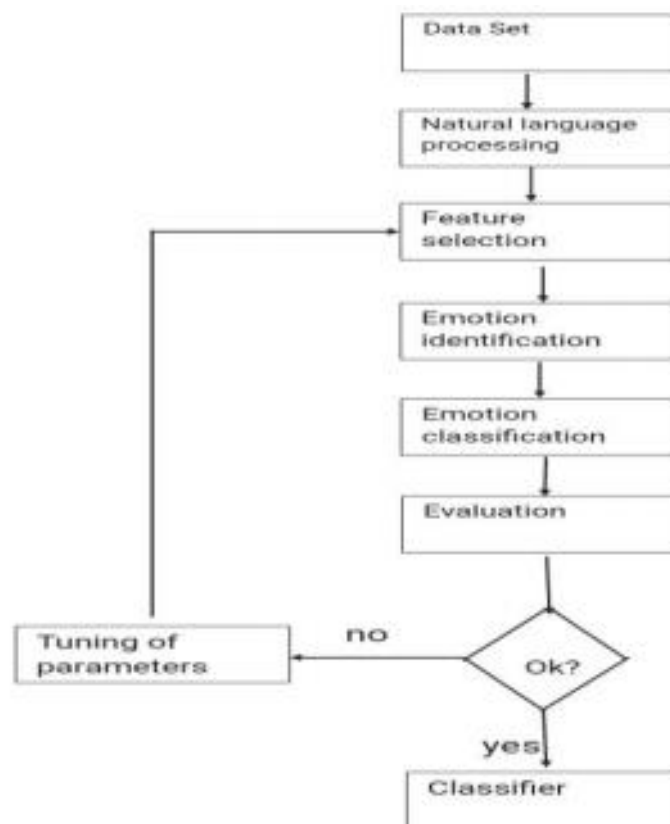

Figure *2*. Procedure for analyzing emotions

3.5. Experiment

This research is based on the use of lexicon-based analysis as an analytical tool for natural language processing. The text was pre-processed and filtered after collecting data from Twitter. Then take a series of procedures to demonstrate individual sentiments as well as the most prevalent emotion among Iraqis regarding Corona virus vaccine. Figure 3 shows the algorithm that was used to analyze the sentiment of the Iraqi populace.

The analysis procedure was as follows:

1- The keys are used to generate tokens that are used to authenticate the browser after receiving the Twitter API and Google Map API. The Iraqi trend "كورونا" on November 2020 is used in this research. This hashtag is used to perform a Twitter search and collect data.

2- Non Arabic tweets are removed from the fetched tweets.

3- Remove URLs, numbers, punctuation, emoticons, and stopwords from the filtered data.

4- To get the words back to their roots, a stemming process is used.

5- Make a list of the most frequently used words.

6- Create the TDM (term-document matrix), which describes the frequency of words found in the cleaned tweets.
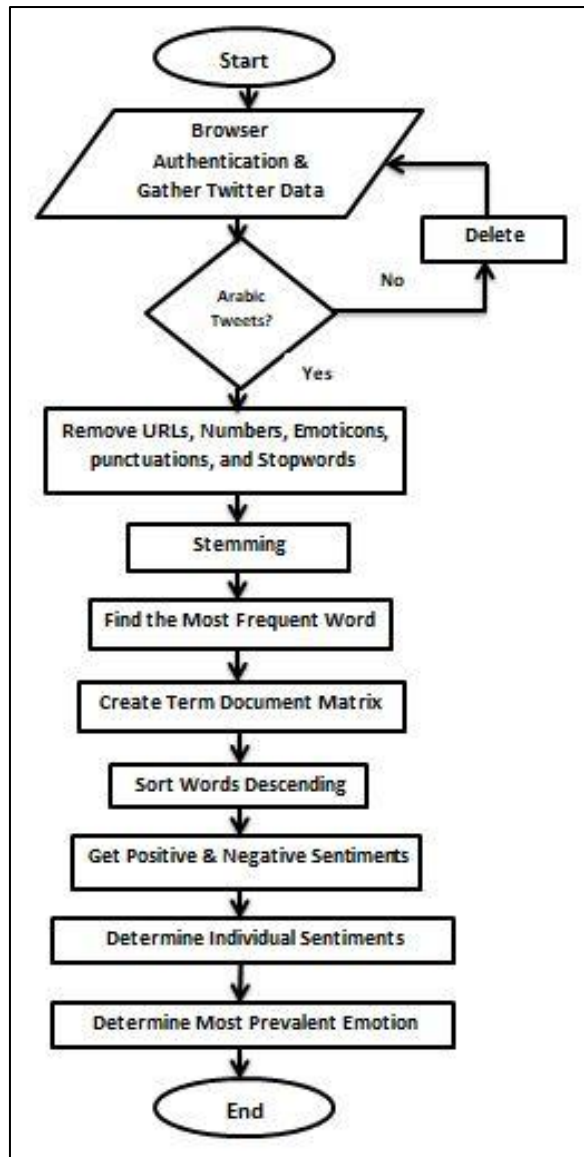
7- Arrange the words in decreasing order.

Figure *3*. Lexicon-based sentiment analysis

8- Use the NRC lexicon to compare the most frequently used terms to automatically determine positive and negative sentiments and display them in a chart.

Symbol of NRC lexicon can be shown in figure 4 which shows each\ word with its sentiment:

Figure *4*. Symbol of NRC lexicon

The comparison process between the most frequent collected words with the used lexicon to determine the positive and negative sentiments is shown in figure 5:

```
> v_text_sentiment1
        word index v_original_n v_positive_negative v_original_sentiment anger anticipation disgust fear
1    account     0            1            positive                trust     0            0       0    0
2      ahead     2            1            positive             positive     0            0       0    0
3    analyst     0            1            positive         anticipation     0            1       0    0
4    analyst     0            1            positive             positive     0            0       0    0
5    analyst     0            1            positive                trust     0            0       0    0
6      angel     0            1            positive         anticipation     0            1       0    0
7      angel     0            1            positive                  joy     0            0       0    0
8      angel     0            1            positive             positive     0            0       0    0
9      angel     0            1            positive             surprise     0            0       0    0
10     angel     0            1            positive                trust     0            0       0    0
11       art     0            1            negative              sadness     0            0       0    0
12       art     0            1            positive         anticipation     0            1       0    0
13       art     0            1            positive                  joy     0            0       0    0
14       art     0            1            positive             positive     0            0       0    0
15       art     0            1            positive             surprise     0            0       0    0
16    assist     2            1            positive             positive     0            0       0    0
17    assist     2            1            positive                trust     0            0       0    0
18    author     0            1            positive             positive     0            0       0    0
19    author     0            1            positive                trust     0            0       0    0
20     award     1            1            positive         anticipation     0            1       0    0
21     award     1            1            positive                  joy     0            0       0    0
22     award     1            1            positive             positive     0            0       0    0
23     award     1            1            positive             surprise     0            0       0    0
24     award     1            1            positive                trust     0            0       0    0
25    banker     2            1            positive                trust     0            0       0    0
26      base     0            1            positive                trust     0            0       0    0
27     beach     2            1            positive                  joy     0            0       0    0
28     beast     2            1            negative                anger     1            0       0    0
29     beast     2            1            negative                 fear     0            0       0    1
30     beast     2            1            negative             negative     0            0       0    0
31   benefit     2            1            positive             positive     0            0       0    0
32     birth     2            1            negative                 fear     0            0       0    1
33     birth     2            1            positive         anticipation     0            1       0    0
34     birth     2            1            positive                  joy     0            0       0    0
35     birth     2            1            positive             positive     0            0       0    0
36     birth     2            1            positive                trust     0            0       0    0
37  birthday     1            1            positive         anticipation     0            1       0    0
38  birthday     1            1            positive                  joy     0            0       0    0
39  birthday     1            1            positive             positive     0            0       0    0
40  birthday     1            1            positive             surprise     0            0       0    0
```

Figure 5. Symbol of compared words with NRC lexicon

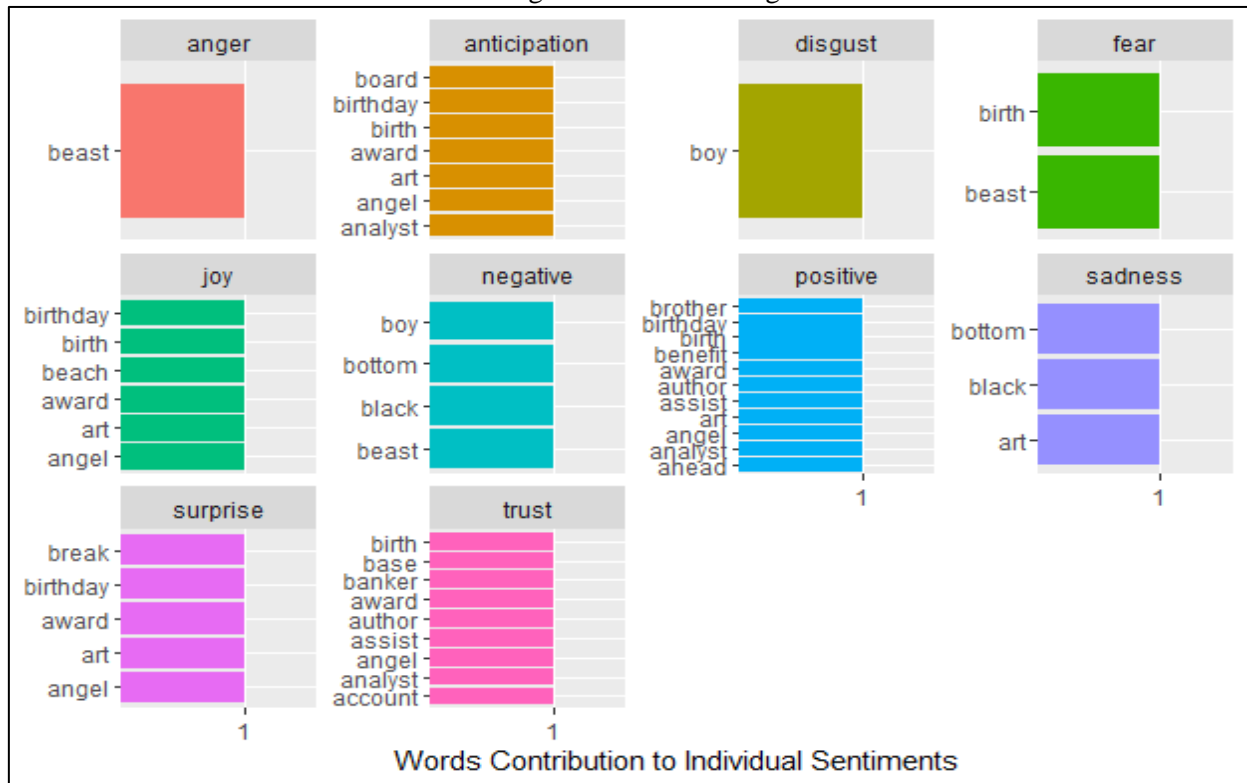The individual sentiments of the chosen hashtag can be shown in figure 6:

Figure 6. The most contributed individual sentiments

9- The most frequent emotion for the chosen hashtag is calculated and displayed in a plot by summarizing the sentiments.

The most positive and negative feelings can be shown in figure 7. Positive sentiments can be found in the right-hand corner. The negative sentiments are displayed in the bottom left corner.
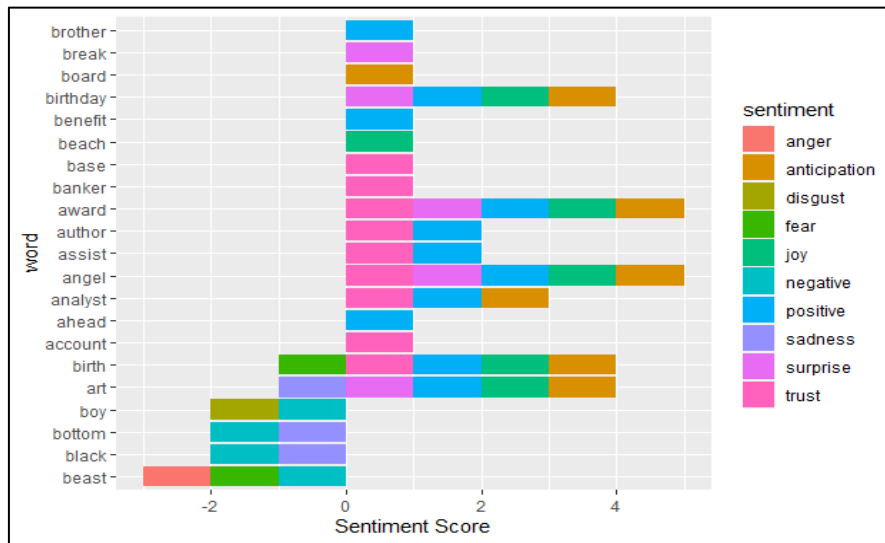

Figure 7. Most populated positive and negative emotions

Finally, using the hashtag "كورونا" we were able to determine the most prevalent emotion for Iraqi population opinion about corona vaccine. They were feeling trust about it. Figure 8 depicts the situation.
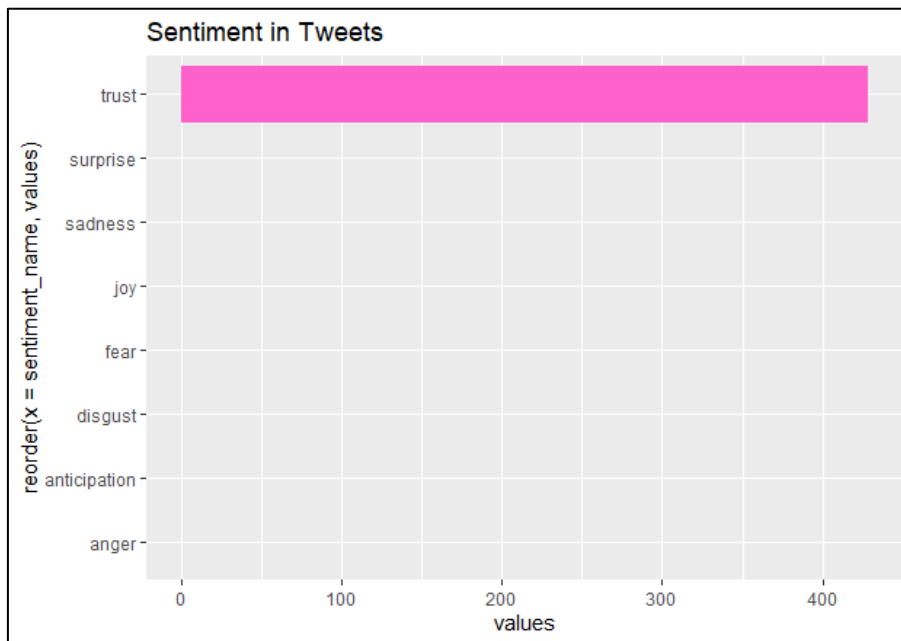
Figure 8. Most prevalent emotion

4.   Conclusions

In this research, a sentiment analysis method was developed to examine Iraqis' perceptions of the corona virus vaccine. In our work, we could: Create a Twitter developer account and a Twitter application that gives us access to the Twitter API, allowing us to get Twitter data by selected hashtag. Examine the public's positive and negative sentiments, also their most prevalent emotions on the chosen issue.

References

[1]   S. Andrews, S. Yates, B. Akhgar, and D. Fortune, "The ATHENA project: Using formal concept analysis to facilitate the actions of responders in a crisis situation," Strateg. Intell. Manag. Natl. Secur. Imp. Inf. Commun. Technol., pp. 167–180, 2013.

[2]   H. Gibson, S. Andrews, K. Domdouzis, L. Hirsch, and B. Akhgar, "Combining big social media data and FCA for crisis response," Proc. - 2014 IEEE/ACM 7th Int. Conf. Util. Cloud Comput. UCC 2014, pp. 690–695, 2014, doi: 10.1109/UCC.2014.112.

[3]   K. Musiał and P. Kazienko, Social networks on the Internet, vol. 16, no. 1. 2013.

[4]   S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson, and T. Seymour, "The History of Social Media and its Impact on Business," Management, vol. 16, no. 3, pp. 79–91, 2011, [Online]. Available: http://search.proquest.com.eproxy.ucd.ie/docview/889143980.

[5]   R. Yazdanifard, W. Khalid Obeidy, W. Fadzilah, W. Yusoff, and H. R. Babaei, "Social Networks and Microblogging;The Emerging Marketing Trends & Tools of the Twenty-first Century," Int. Conf. Comput. Commun. Manag., vol. 5, no. May 2014, pp. 577–581, 2011.

[6]   H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," Proc. 19th Int. Conf. World Wide Web, WWW '10, pp. 591–600, 2010, doi: 10.1145/1772690.1772751.

[7]   N. J. Nilsson, "Principles of artificial intelligence," Morgan Kaufmann, 2014.

[8]   G. Tecuci, "Artificial intelligence," Wiley Interdiscip. Rev. Comput. Stat., vol. 4, no. 2, pp. 168–180, 2012, doi: 10.1002/wics.200.

[9]   and A. S. Shmelova, T., Y. Sikirda, "Artificial Intelligence Applications in the Aviation and Aerospace Industries," Eng. Sci. Ref., 2019.

[10]   M. N. O. Sadiku, M. Tembely, and S. M. Musa, "Social Media for Beginners," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 8, no. 3, p. 24, 2018, doi: 10.23956/ijarcsse.v8i3.564.

[11]   V. Vasudha Rani and K. Sandhya Rani, "Twitter Streaming and Analysis through R," Indian J. Sci. Technol., vol. 9, no. 45, 2016, doi: 10.17485/ijst/2016/v9i45/97914.

[12]   P. Sagdeo, "APPLICATION PROGRAMMING INTERFACES AND THE STANDARDIZATION-VALUE APPROPRIATION PROBLEM," Harv. J. Law Technol., vol. 33, no. 1, 2019.

[13]   "A Tutorial for Using Twitter Data in the Social Sciences : Data Collection , Prepartion , and Analysis."

[14]   S. Hill, R. Scott, and S. Chain, "Developing an Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R," Inf. Syst. Educ. J., vol. 15, no. 3, p. 15, 2017.

[15]   Ayyoub, Essa, and Alsmadi, "Lexicon-based sentiment analysis of Arabic tweets Mahmoud Al-Ayyoub * and Safa Bani Essa Izzat Alsmadi," Int. J. Soc. Netw. Min. 2(2) 101, vol. X, no. June 2014, 2015, doi: 10.1504/IJSNM.2015.072280.

[16]   R. Gull, U. Shoaib, S. Rasheed, W. Abid, and B. Zahoor, "Pre Processing of Twitter's Data for Opinion Mining in Political Context," Procedia Comput. Sci., vol. 96, no. September, pp. 1560–1570, 2016, doi: 10.1016/j.procs.2016.08.203.

[17]   M. Tare, I. Gohokar, J. Sable, D. Paratwar, and R. Wajgi, "Multi-Class Tweet Categorization Using Map Reduce Paradigm," Int. J. Comput. Trends Technol., vol. 9, no. 2, pp. 78–81, 2014, doi: 10.14445/22312803/ijctt-v9p117.

[18]   A. Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," IISA 2016 - 7th Int. Conf. Information, Intell. Syst. Appl., no. July, 2016, doi: 10.1109/IISA.2016.7785373.

[19]   Anjali, G. Jivani, and M. Anjali, "A Comparative Study of Stemming Algorithms," October, vol. 2, no. 2004, pp. 1930–1938, 2007.

[20]   V. A. and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," Int. J. Comput. Appl., vol. 139, no. 11, pp. 5–15, 2016, doi: 10.5120/ijca2016908625.

[21]   V. D. Nguyen, B. Varghese, and A. Barker, "The royal birth of 2013: Analysing and visualising public sentiment in the UK using Twitter," Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 2013, no. i, pp. 46–54, 2013, doi: 10.1109/BigData.2013.6691669.

[22]   N. M. Hakak, M. Mohd, M. Kirmani, and M. Mohd, "Emotion analysis: A survey," 2017 Int. Conf. Comput. Commun. Electron. COMPTELIX 2017, no. July, pp. 397–402, 2017, doi: 10.1109/COMPTELIX.2017.8004002.