Multimedia Tools and Applications (2023) 82:27819–27838
https://doi.org/10.1007/s11042-023-14514-x

A new sentiment analysis method to detect and Analyse sentiments of Covid-19 moroccan tweets using a recommender approach

Youness Madani[1] · **Mohammed Erritali**[1] · **Belaid Bouikhalene**[1]

Received: 24 February 2021 / Revised: 19 September 2022 / Accepted: 31 January 2023 /
Published online: 22 February 2023

Abstract
Since the beginning of the covid-19 crisis, people from all over the world have used social media platforms to publish their opinions, sentiments, and ideas about the coronavirus epidemic and their news. Due to the nature of social networks, users share an immense amount of data every day in a freeway, which gives them the possibility to express opinions and sentiments about the coronavirus pandemic regardless of the time and the place. Moreover, The rapid number of exponential cases globally has become the apprehension of panic, fear, and anxiety among people. In this paper, we propose a new sentiment analysis approach to detect sentiments in Moroccan tweets related to covid-19 from March to October 2020. The proposed model is a recommender approach using the advantages of recommendation systems for classifying each tweet into three classes: positive, negative, or neutral. Experimental results show that our method gives good accuracy(86%) and outperforms the well-known machine learning algorithms. We find also that the sentiments of users changed from period to period, and that the evolution of the epidemiological situation in morocco affects the sentiments of users.

Keywords Sentiment analysis · Covid-19 · Recommendation system · Collaborative filtering · Classification

1 Introduction

COVID-19, originally known as CoronaVirus disease of 2019, was declared a pandemic by the World Health Organization (WHO) on March 11, 2020. Unprecedented pressures

✉ Youness Madani
younesmadani9@gmail.com

Mohammed Erritali
m.erritali@usms.ma

Belaid Bouikhalene
B.BOUIKHALENE@usms.ma

[1] Sultan Moulay Slimane University, Beni Mellal, Morocco

have mounted on every country to make conditions imperative to control the population by assessing cases and properly using available resources. The rapid number of exponential cases around the world has become the apprehension of panic, fear, and anxiety among people. We see that the mental and physical health of the world's population is directly proportional to this pandemic. The current situation has reported 225070 people tested positive in Morocco as of November 2, 2020.

The COVID-19 epidemic has had a huge impact on the general lifestyles of people around the world. People express their views on COVID-19 more frequently on social media when cities are on lockdown. In Twitter, for example, millions of users share their opinions and sentiments about coronavirus every day and in all languages, which produces a large amount of data.

Since the beginning of this crisis, many studies have been done by many researchers on different domains(from medicine to social science). However, the exploration of Covid-19 and social media is limited, which motivates us to analyze the sentiments of people and their evolution in the face of this public health crisis based on Twitter tweets.

Social network analysis SNA (sentiment analysis on social media) is a domain that consists of analyzing social media content and extracting opinions, sentiments, and attitudes. SNA can extract sentiments in two manners: either by classifying a social content into classes such as positive, negative, and neutral; or by extracting the degree of polarity. If we take a deep look into the literature of sentiment analysis methods, we find that the majority of the existing approaches could be classified into three classes: Dictionary-based approaches (SentiWordNet, SenticNet...etc) [20], Machine learning approaches(Naive Bayes, SVM, Random forest...etc) [8], and hybrid approaches that combine between the two last approaches [10].

Given the epidemiological situation in Morocco linked to covid-19, which begins on March 2, 2020, and is still experiencing a remarkable evolution until today, the feelings of Moroccans on this epidemic are changing with the cases recorded (they reached 225070 cases on November 2, 2020).

From all that, this work aims to analyze and extract sentiments from the tweets published in morocco, to get a general idea about the feelings of Moroccan users during the crisis of covid-19. For extracting the tweets to analyze, we based on the keywords **"covid-19"**, **"coronavirus"**, and **"moroccoCovid-19"** for the tweets published between march and October 2020.

The proposed method is a new sentiment analysis approach that takes advantage of a dictionary approach and a proposed recommender approach to improve the accuracy of the classification. Our method is a collaborative filtering approach that uses four tweet features for finding the k-nearest neighbors. For finding these four tweet features, we based on a dictionary approach using the SenticNet dictionary, TextBlob library, and the Natural Language processing methods. Our model is multilingual(supports four languages: Arabic, English, Spanish, and French), and it takes into account the spelling check(corrects spelling errors).

In its recent sense, collaborative filtering is underlying recommendation systems. It brings together techniques that aim to make a selection on (filtering) the elements to present to users(target users) based on the behavior and expressed tastes of a large number of other users (collaboration). In our work, the tweets to classify play the role of target users, and the other users are represented by a labeled dataset of tweets. The idea is to find the k-nearest neighbors labeled tweets of the target tweet, which will help us to find the sentiment(class) or in other words recommend the relevant class for the target tweet.

Collecting information for collaborative filtering plays a crucial role in the process, it can be:

- Explicit: The user assigns ratings to the products or indicates their appreciation (like).
- Implicit: Collection based on behavior (purchases, clicks, duration on a page).

In our work, the collection of information for our collaborative filtering CF model is done using a dictionary-based approach based on the Senticnet dictionary, and the Textblob python library to find four sentiments(four features of the tweet), that will help our CF model to assign the relevant class(recommend a class) for the target tweet.

The majority of methods existing in the literature are based on machine learning algorithms which necessitate annotated datasets to implement the models, empirically we demonstrate that the method based on recommendation approach and a dictionary-based approach outperforms methods based on machine learning. Additionally, our method takes into account four languages(Arabic, French, English, and Spanish), Also it supports spelling check and correction. All that improves the classification rate. The choice of using our method is based on the result that the recommendation methods give in a lot of domains (e-learning, data science, industry...etc), and also based on some existing works that demonstrate the strength of using the dictionary-based approach in sentiment analysis domain. All that motivates us to propose a new approach based on a recommendation approach and a dictionary-based approach.

The rest of this paper is organized as follows: in Section 2, we describe the literature review related to our domain. Section 3 presents our approach by describing all the steps needed to classify tweets. In Section 4, we implement the experimental results to show the performance of our model. And finally, it's the conclusion of our article.

2 Literature review

During this year and due to the rapid propagation of the coronavirus in the world, a lot of scientific articles have been published in all domains From medicine to literature. In the domain of sentiment analysis, many researchers proposed new approaches to analyze and extract the sentiment of users from social media platforms related to the covid-19 epidemic.

Authors in [7] used automated extraction of COVID-19–related discussions from social media and a natural language process (NLP) method based on topic modeling to uncover various issues related to COVID19 from public opinions. They investigate how to use LSTM recurrent neural network for sentiment classification of COVID-19 comments. For that, researchers present a systematic framework based on NLP that is capable of extracting meaningful topics from COVID-19– related comments on Reddit, and for the classification of the comments, they propose a deep learning model based on Long ShortTerm Memory (LSTM) for sentiment classification of COVID-19–related comments. Experiments demonstrated that the research model achieved an accuracy of 81.15% – a higher accuracy than that of several other well-known machine-learning algorithms for COVID-19–Sentiment Classification.

Aslam et al. [1] published a new article extract and classify sentiments and emotions from 141,208 headlines of global English news sources regarding the coronavirus disease (COVID-19). Authors take into account news with keyword coronavirus between the time frame 15 January 2020 to 3 June 2020 from top rated 25 English news sources. Each headline is classified into three classes: positive, negative, or neutral. For the classification of

each headline, authors used the R package "sentiment" by relying on lists of words and phrases with positive and negative connotations.

In the article of [19], researchers identify public sentiment associated with the pandemic using Coronavirus specific Tweets and R statistical software, along with its sentiment analysis packages. The authors used a number of machine learning algorithms to extract sentiments from tweets(Linear Regression Model, Naïve Bayes Classifier, Logistic Regression) and also some textual methods. The aim of this article is to demonstrate insights into the progress of fear-sentiment over time as COVID-19 approached peak levels in the United States. Experimental results show a strong classification accuracy of 91% for short Tweets, with the Naïve Bayes method. We also observe that the logistic regression classification method provides a reasonable accuracy of 74% with shorter Tweets, and both methods showed relatively weaker performance for longer Tweets.

The study of [18] focuses on the sentiment analysis of tweets of the Twitter social media using Python programming language with Tweepy and TextBlob library. Authors collect tweets based on two specified hashtags keywords: $\#COVID-19$ and $\#coronavirus$ from the users who shared their location as 'Nepal' between 21st May 2020 and 31st May 2020. The result of the study concluded that while the majority of the people of Nepal are taking a positive and hopeful approach, there are instances of fear, sadness and disgust exhibited too.

In the work [5]; the authors analyze discussions on Twitter related to COVID-19. researchers used tweets originating exclusively in the United States and written in English during the 1-month period from March 20 to April 19, 2020. For the classification, They applied machine learning methods to classify tweets into three classes(positive, negative, and neutral). For a dataset of 902,138 tweets, the proposed model classified 434,254 (48.2%) tweets as having a positive sentiment, 187,042 (20.7%) as neutral, and 280,842 (31.1%) as negative.

In [16], Muthusami et al. analyze and visualize the influence of coronavirus (COVID-19) in the world by executing such algorithms and methods of machine learning in sentiment analysis on the tweet dataset to understand very positive and very negative opinions of the public around the world. To analyse tweets, They used machine learning approaches, and results show that the LogitBoost algorithm performed better with accuracy of 74

Authors of [14] use Twitter to analyse sentiments related to covid-19 epidemic. For the collection of the tweets to classify, researchers are based on two specified hashtag keywords, which are ("COVID-19, coronavirus") using the tweepy library. The date of searching data is seven days from 09-04-2020 to 15-04-2020. And by usingTextBlob library in python, the sentiment analysis operation has been done.

The work of Chakraborty et al. [2] presents a new research on analysing sentiments in tweets during the period of covid-19 epidemic. The datasets used are obtained by searching using the keywords: $\#corona$, $\#covid19$, $\#coronavirus$, coronavirus and $\#covid-19$. For the classification, authors propose a model using deep learning classifiers with admissible accuracy up to 81%, and an implementation of a Gaussian membership function based fuzzy rule base to correctly identify sentiments from tweets.The accuracy for the said model yields up to a permissible rate of 79%.

In [11] authors analyse the sentiments and their evolution of people in the face of this public health crisis based on Chinese Weibo. For constructing the dataset of work, authors obtained the top 50 hot searched hashtags from January 10, 2020 to May 31, 2020, and collected 1,681,265 Weibo posts associated to the hashtags regarding COVID-19. For the classification, They use 7 classes(fear, anger, disgust, sadness, gratitude, surprise, and optimism) to annotate each Weibo post. To detect sentiments of users in Weibo, researchers use

three methods, i.e., LSTM, BERT, and ERNIE, and experimental results show that ERNIE classifier has the highest accuracy and reaches 0.8837.

Lamiaa Mostafa in the paper [15], propose a Sentiment Analysis Model that will analyze the sentiments of students in the learning process within their pandemic using Word2vec technique and Machine Learning techniques. The proposed model use a method that starts with the processing process on the student's sentiment and selects the features through word embedding then uses three Machine Learning classifiers which are Naïve Bayes, SVM and Decision Tree. Results show that the Naïve Bayes classifier gives best results with an accuracy equal to 87% by using the DF word embedding method, and 91% using skip-gram method.

In [17], authors analyse sentiments of users in the Twitter platform using the keywords 'covid' and 'coronavirus'. The proposed model use advantages of Natural Language Processing and the Recurrent Neural Network algorithm. The model gives good results at the level of accuracy in comparison with the TextBlob library.

In [3], researchers collected a total of 410,643 tweets in English related to covid-19 in india from March 22 to April 21, 2020. This article aims to detect and analyse sentiments during the lockdown.

The work of Heras-Pedrosa et al. [4] examines how social media has affected risk communication in uncertain contexts and its impact on the emotions and sentiments derived from the semantic analysis in Spanish society during the COVID-19 pandemic. This research collected data directly in real time from the main media and digital ecosystems: Twitter, YouTube, Instagram, official press websites, and internet forums, during March and April 2020. Research results also demonstrate a lot of mixed feelings. It is observed that the same news, information or media communication generated peaks in different emotions, indicating that they are very mixed between sadness, disgust, anger, and fear.

IMRAN et al. in their article "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets" [6] analyze reaction of citizens from different cultures to the novel Coronavirus and people's sentiment about subsequent actions taken by different countries. Deep long short-term memory (LSTM) models used for estimating the sentiment polarity and emotions from extracted tweets have been trained to achieve stateof-the-art accuracy on the sentiment140 dataset.

In [9], authors analyze Twitter messages (tweets) collected during the first months of the COVID-19 pandemic in Europe with regard to their sentiment. In their approach, researchers use a method that is implemented with a neural network for sentiment analysis using multilingual sentence embeddings. Authors analyzed around 4.6 million tweets, of which around 79,000 contained at least one COVID-19 keyword. The analysis is done by contry. As a result, authors find for example, that lockdown announcements correlate with a deterioration of mood in almost all surveyed countries, which recovers within a short time span.

In the paper of [21], authors use 999,978 randomly selected COVID-19 related Weibo posts from 1 January 2020 to 18 February 2020. For the classification, the unsupervised BERT (Bidirectional Encoder Representations from Transformers) model is adopted to classify sentiment categories (positive, neutral, and negative) and TF-IDF (term frequency-inverse document frequency) model is used to summarize the topics of posts. This study demonstrates how public sentiment on social media evolves as COVID-19 spreads.

All the methods cited above are based on machine learning and deep learning approaches to detect sentiments related to covid-19 which requires massive data sets(labeled datasets) to train the models for making the classification. Also, massive data needs enough time

to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy, without forgetting that with machine learning algorithms sometimes the used datasets for training the algorithms affect negatively the classification. Our proposed method takes advantage of recommendation systems and also of dictionary-based approaches to propose a new approach for analyzing tweets. It does not necessitate any dataset for training, which optimizes the classification time and also it demonstrates a high classification rate in comparison with approaches that use machine learning and deep learning.

3 Research methodology

As presented earlier the main objective of this work is to analyze the Moroccan tweets during the period of the Covid-19 epidemic. The idea is to propose a new approach for improving the results of a sentiment analysis approach, that will classify each tweet into three classes(positive, negative, and neutral). Our proposed method is based on the notions of Natural language processing methods (text preprocessing, spell check, multilingual approach...etc), and a new collaborative filtering approach that uses a dictionary-based approach and the Textblob python library.

We want to analyze the sentiments of Morrocan users from the beginning of the covid-19 crisis in morocco until the end of August 2020. For that, we collect tweets from March 2020 until October 2020 in the form of periods. For example, we collect the tweet published from 1st March to 15 March 2020 as the first period, and from 16 March to 30 March 2020 as the second period, until the period from 16 October to 30 October as the last period. The idea is to analyze the sentiments of users in each period to find the relation between the sentiments of users and the Epidemiological situation in morocco.

The proposed approach is a multilingual method that takes into account four languages (Arabic, French, English, and Spanish). This choice is validated by the fact that most Moroccan tweets are written in one of these 4 languages. Also, our approach uses a spelling check method to clean the tweets from wrong words, and some Natural language preprocessing methods.

In the next subsections, we will explain in detail the necessary steps of our method.

3.1 The datasets of work

As mentioned earlier, our work needs old tweets from march to October 2020. For that, we developed a new python program that gives us the possibility to retrieve old tweets(because some libraries like Tweepy allow users to collect tweets for only one week later). Our program is based on GetOldTweets3(A Python 3 library and a corresponding command-line utility for accessing old tweets). The collected tweets are divided into periods, for example, the first period is from March,1st to march,15 2020, and the last period is from October,1st to October 15, 2020. In each period we based on three keywords to collect tweets which are: **“Covid-19”, “coronavirus”, and “moroccoCovid-19”**. Also for working only with the Moroccan tweets, we used the “near” option of the GetOldTweets3 library that consists in retrieving tweets in an exact location using latitude and longitude.

Figure 1 shows the different steps for collecting the tweets to analyze.

As shown in Fig. 1, our model can collect old tweets(from March 2020) and store them in files(each period contains all the published tweets related to the coronavirus epidemic and located in Morocco).

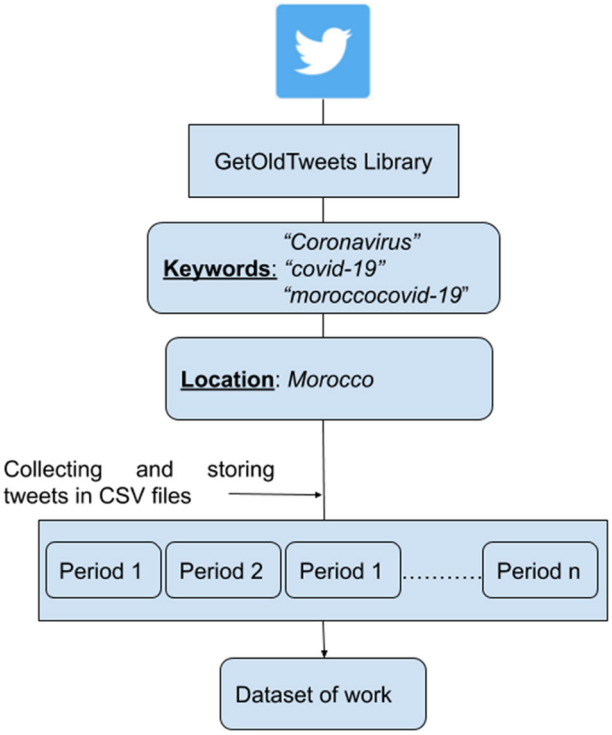

Fig. 1 Steps for collecting the tweets

After the construction of our dataset of work. Another important step in this article consists in choosing a labeled dataset that will help us to analyze the performance of our proposed model and also in the experimental results step. From all that in this article, we worked with the Twitter US Airline Sentiment TUAS dataset which consists in Analyzing how travelers in February 2015 expressed their feelings on Twitter. This dataset contains tweets with several features and also with a target column that accepts three values: 1 for positive, -1 for negative, or 0 for neutral. TUAS dataset is used in the problem related to sentiment analysis, it contains 14640 entries written in English.

3.2 Tweets preprocessing

The proposed approach is based on the tweet text, an important step consists in preparing the tweet for the analysis. The published tweets can contain unimportant information or wrong words. The text preprocessing methods clean the tweet and prepare it for the classification, which will improve the results of the analysis [12].

Our text preprocessing methods begin with the detection of hashtags and then removing the "#" character from them. The next step consists in removing the words with the "@" character, which are not necessary for the classification, and finally, we remove URLs and numbers.

After these three steps, we apply several text preprocessing methods such as :

- Tokenization: This consists in splitting the tweet to classify into individual words.

- Removing stopwords: Stopwords are words that are unnecessary for detecting sentiments. Examples of stopwords include articles and pronouns. In our works, we used the NLTK Python library which provides a complete list of stopwords (in 4 languages: French, Arabic, English, and Spanish) to remove from each tweet to classify.
- Stemming: is the process of reducing inflected (or sometimes derived) words to their root, base, or root form. Our approach uses the NLTK python library for stemmers(English stemmer, French stemmer, Arabic stemmer, and Spanish stemmer).

After applying all these methods, we find clean tweets with only important words for the analysis. But before the application of our approach, there are two important methods to apply to each tweet.

3.2.1 Spell check

Sometimes users on Twitter publish some tweets with spell errors. These errors can negatively affect the analysis and the classification of tweets. To avoid these problems our proposed approach supports spell check methods that detect and correct spell errors in each tweet to classify. In this work, our spell check approach uses the pyspellchecker python library that gives us the possibility to detect spell errors and correct them. This approach goes through the tweet word by word and if it finds a spelling error, it corrects it and replaces the wrong word with the correct one.

Algorithm 1 shows the different steps of our spell check model.

```
lang ← Langdetect(tweet_text)
if lang =’en’ then
    spell ← SpellChecker()
else if lang =’ar’ then
    spell ← SpellChecker(language=’ar’)
else if lang =’fr’ then
    spell ← SpellChecker(language=’fr’)
else if lang =’es’ then
    spell ← SpellChecker(language=’es’)
end if
clean_tweet=[]
for all w ∈ tweet_text do
    clean_tweet.append(spell.correction(w))
end forclean_tweet return
```

Algorithm 1 Spell_check function.

Where:

- Langdetect(): is a function that returns the language of the tweet. It accepts as a parameter a tweet's text.
- SpellChecker(): is a function that initializes our spell check model automatically for English.
- SpellChecker(language=‘ar’): a function that initializes our spell check model for the Arabic language.
- clean_tweet.append(spell.correction(w)): is an instruction that adds to the clean_tweet table each word of the tweet after the application of the spell check and correction.

3.2.2 Language detection and translation

As presented earlier, our approach is multilingual, and it supports four languages: French, Spanish, Arabic, and English. Our idea is that in every step of our work, either in the collection of tweets, in the application of the natural language processing methods, or even in the step of spell check, we work with the four languages. For example, in the collection step, our query of search takes into account tweets written in the four languages by adding a new property with four possibilities: Ar for Arabic, Fr for french, Es for Spanish, and En for English. In the step of text preprocessing methods, we based on English and if we find a tweet in another language, we translate it to English to apply all our preprocessing methods. Before we apply the text preprocessing methods and the translation, we apply our method of spell check to improve the quality of tweets.

Our method of language detection and translation is based on langdetect model for the detection of the tweets' language, and the googletrans python library to translate tweets. To improve the quality of translation, our method is based on individual words using the results of the application of the text preprocessing methods.

Figure 2 shows the tweet preprocessing methods to prepare each tweet of our dataset for classification.

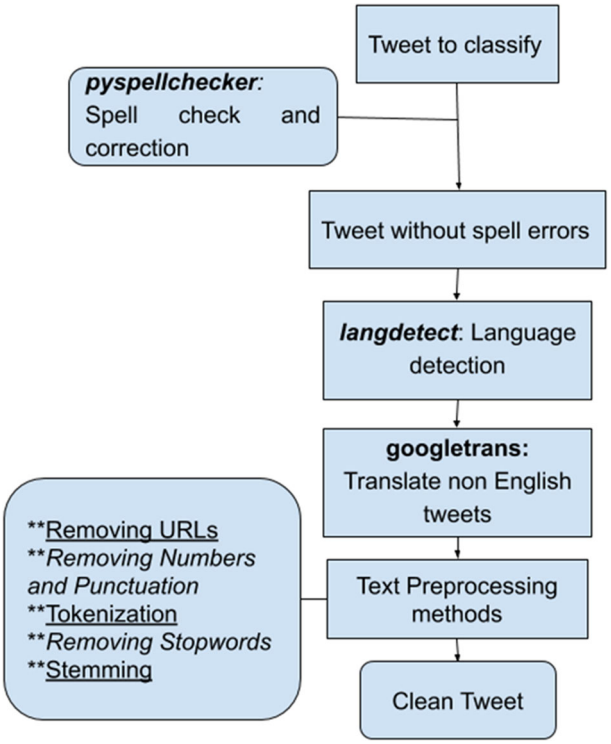

Fig. 2 Tweet preprocessing methods

3.3 Description of our method

In this work, The proposed approach is a new collaborative filtering recommendation approach that uses a dictionary based-approach using the SenticNet dictionary[1] and the TextBlob python library. The proposed method uses the advantages of recommendation systems, dictionary-based approaches to extract sentiment polarity from tweets, and natural language processing methods. Our approach is multilingual, it supports four languages. Moreover, it uses a spell check method on the tweets' text before the classification.

Collaborative filtering (CF) recommends ratings or products for a user based on the rating preferences of similar users. CF is based on the assumption that users with similar tastes have similar preferences to products or items, it is divided into a memory-based approach and a model-based approach. In the model-based CF, models are developed using different algorithms (neural network, machine learning algorithms...) to predict and recommend relevant products or services. On the other hand, Memory-based CF approaches are based on the calculation of the similarity between users or items. they are divided into user-based CF and item-based CF [13].

Based on the definition of CF, and as a comparison with our proposed method. Our approach consists in recommending the relevant class(positive, negative, or neutral) for the tweet to classify(target tweet) based on the k-nearest neighbor's tweets in the labeled dataset. CF approaches recommend relevant content to the target user based on his ratings of some products and also the ratings of other users. The idea is that the system tries to find similar users of the target user and based on their preferences, the system predicts and recommends the relevant content to the target user.

From the definitions above-mentioned, the ratings of the target tweet and the labeled tweets(from the labeled dataset) are obtained by extracting sentiment polarities using a combination between tweets content, natural language processing methods, Textblob library, and a dictionary-based approach.

Our proposed collaborative filtering approach begins with the calculation of ratings of the target tweet and the tweets of the labeled dataset(in which we will look for the k-nearest neighbors tweet of the target tweet). Each tweet in this work will be presented in the form of a vector of four elements as presented below:

$$Tweet\_vector = [sentiment, sentiment\_hashtag, sentiment\_sentic\_text, sentiment\_sentic\_hashtag]$$

where:

- **Sentiment**: a new feature added to the tweet. The calculation of this element of the tweet's vector is based on individual words of the cleaned tweet and the TextBlob python library to extract the sentiment of the tweet's text. This element accepts three values: 1 for positive, -1 for negative, and 0 for neutral.

$$Polarity = \frac{\sum_{i=1}^{N} P_i}{n} \tag{1}$$

$$sentiment = \begin{cases} 1 & \text{if } Polarity \geqslant 0.1 \\ -1 & \text{if } Polarity \leqslant -0.1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

[1] https://sentic.net/

With:

- n: is the number of words in the tweet
- Pi: polarity of the word i
- **Polarity**: is the total polarity of the tweet.

- **sentiment_hashtag**: The calculation of this element of the tweet's vector is based on individual words extracted from the tweet's hashtags and the TextBlob library. It accepts three values: 1 for positive, -1 for negative, and 0 for neutral.
- **sentiment_sentic_text**: This element consists in extracting sentiment from the tweet's text using the SenticNet dictionary. This approach goes through the tweet's words by calculating the polarity intensity of each word which has a value between -1 and 1. And to find the final sentiment of the tweet we calculate the average of the tweet's words polarity as presented in the following formulas:

$$Polarity\_Intense = \frac{\sum_{i=1}^{N} SenticNet(i)}{n} \tag{3}$$

$$sentiment\_senticnet = \begin{cases} 1 & \text{if } Polarity\_Intense \geqslant 0.1 \\ -1 & \text{if } Polarity\_Intense \leqslant -0.1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

With:

- n: is the number of words in the tweet
- SenticNet(i): represents the sentimental degree of the tweet's word i using the SenticNet dictionary.
- *Polarity_Intense*: is the sentimental degree of the tweet.
- *sentiment_senticnet*: is the sentiment of the tweet using the SenticNet dictionary.

- **sentiment_sentic_hashtag**: To calculate this element, we based on hashtags' words and the SenticNet dictionary to extract sentiments in the form of three values 0, 1, or -1.

The threshold used in the last formulas(0.1 and -0.1) means that if the final sentimental degree is greater or equal to 0.1 the tweet is positive, if its value is less or equal to -0.1 the tweet is considered negative, and otherwise the tweet is neutral. Empirically this threshold gives the best results in the step of classification, it's for that reason that we have chosen it.

The Definition of tweets' vectors(by calculating these 4 new features for the tweet to classify and the tweets of the labeled dataset) is the first step of our approach. The vectors' elements are like the product ratings in a normal CF system. Figure 3 shows the necessary steps to construct tweets' vectors based on the four proposed elements.

Calculating these four new features is equivalent to giving a rating to each one. In the dataset of training and test(Twitter US Airline Sentiment), for each tweet, we construct a vector with the new features and also with the label feature that have three values: 0 for neutral, -1 for negative, and 1 for positive, which means that every tweet's vector of the dataset of training and test will have the following format:

$$Tweet\_vector = [sentiment, sentiment\_hashtag, sentiment\_sentic\_text, sentiment\_sentic\_hashtag, label]$$

After we have the necessary ratings(vectors) to start our recommendation approach, the next step in our new sentiment analysis collaborative filtering approach is the search of the k-nearest neighbor's tweets(from the dataset of training and test) of the target tweet(tweet

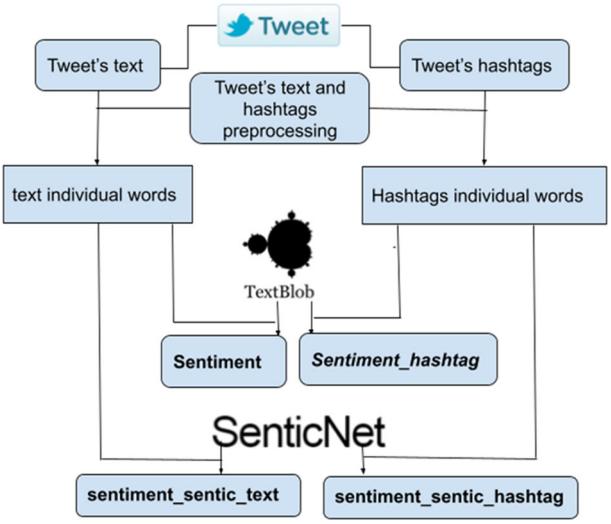

Fig. 3 Steps to construct tweets' vectors

to classify), that is to say, finding the similar vectors (from the vectors of the tweets of the Twitter US Airline Sentiment dataset) of the target tweet's vector. For calculating the similarity between vectors we based on the cosine similarity [22].

$$Similarity(A, B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5}$$

By using a threshold, we keep as similar vectors to the target tweet's vector the ones that have a similarity value greater than or equal to 0.5. These retrieved similar vectors are the k-nearest tweets' vectors to the target tweet vector (Fig. 4).

The final step of our CF approach is to recommend the relevant class(sentiment) to the target tweet. For that, we based on the labels of similar vectors by searching the majority class. For example, if we find that the majority of vectors(from the list of the similar vectors) have a label with 1 value, we recommend to the target tweet the positive sentiment, if the

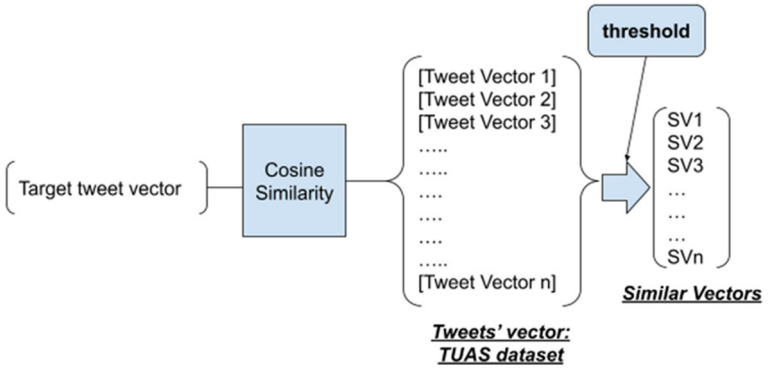

Fig. 4 Similar vectors

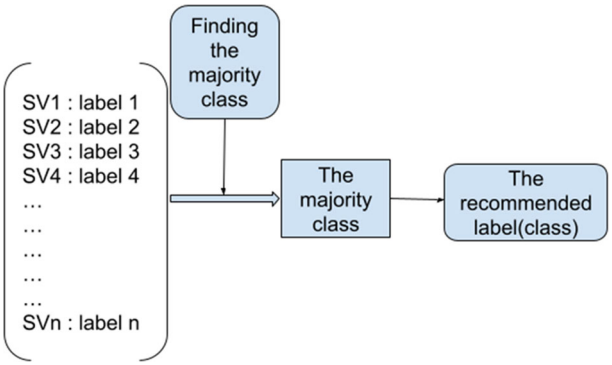

Fig. 5 Recommended label

majority of vectors have the value -1 as a label, the target tweet is classified as negative, and the target tweet is neutral if the majority of vectors have the value 0 (Fig. 5).

As presented in Fig. 5, after we find the k-nearest tweets' vectors we look for the majority class that will give us the relevant class of the target tweet.

Figure 6 presents a general description of our model.

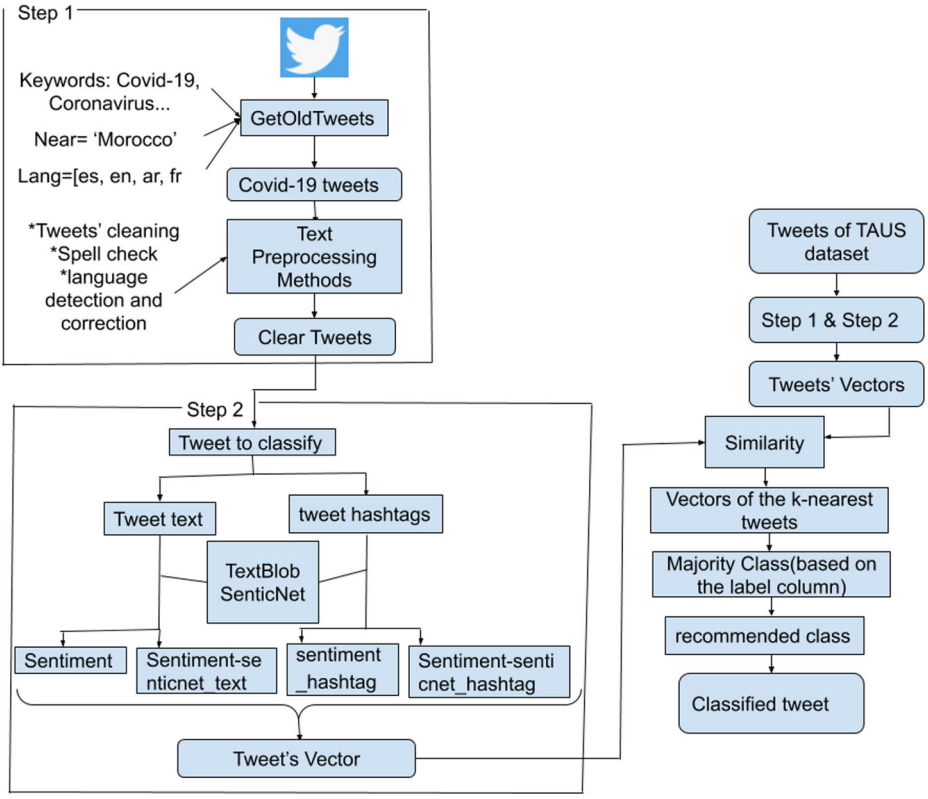

Fig. 6 General description of our model

Algorithm 2 presents the different steps and functions of our approach:

Ensure: a tweet
Require: tweet's sentiment
tweet_text ← GEtOldTweets(tweet)
clean_tweet ← Spell_check(tweet_text)
tweet_lang ← LangDetect(clean_tweet)
tweet ← Text_preprocessing(clean_tweet)
tweet_hashtag ← Retrieve_hashtags(clean_tweet)
tweet_text[] ← Split(tweet)
tweet_hashtag[] ← Split(tweet_hashtag)
if tweet_lang='en' **then**
Sentiment ← ***TextBlob_Sent***(tweet_text)
Sentiment_hashtag ← ***TextBlob_Sent***(tweet_hashtag)
Sentiment_senticnet_text ← ***SenticNet_sent***(tweet_text)
Sentiment_senticnet_hashtag ← ***SenticNet_sent***(tweet_hashtag)
else
translated_tweet ← Translate(tweet_text, dest='en')
translated_hashtag ← Translate(tweet_hashtag, dest='en')
Sentiment ← ***TextBlob_Sent***(translated_tweet)
Sentiment_hashtag ← ***TextBlob_Sent***(translated_hashtag)
Sentiment_senticnet_text ← ***SenticNet_sent***(translated_tweet)
Sentiment_senticnet_hashtag ← ***SenticNet_sent***(translated_hashtag)
end if
tweet_vector=[**Sentiment**, **Sentiment_hashtag**, **Sentiment_senticnet_text**, **Sentiment_senticnet_hashtag**]
K-nearest_neighbors ← Cosine_similarity(tweet_vector, training_vectors)
Sentiment ← Majority_class(K-nearest_neighbors)

Algorithm 2 Sentiment analysis using a recommendation approach.

Where:

- GetOldTweets(): this proposed GetOldTweets model helps to retrieve old tweets(because some libraries like Tweepy allow only to retrieve tweets of the last week). GetOldTweets retrieves tweets with several features, we used the GetOldTweets() function to work only on the tweets' text.
- Spell_check: is our proposed function described in the Spell check subsection.
- Text_preprocessing(): applies the different text preprocessing methods on the tweet's text to prepare it for the classification.
- Retrieve_hashtags(): Retrieves the hashtags from the tweet's text without the character '#' and save them as a sentence.
- Translate(): a function that translates a tweet to a specific language.

The following two algorithms(3 and 4) show the TextBlob_Sent function(which is based on the TextBlob python library) and the SenticNet_sent function(that uses a dictionary-based approach to retrieve sentiments).

```
lang ← Langdetect(tweet_text)
if lang = 'en' then
    Sent ← TextBlob(tweet_text)
else if lang! ='en' then
    translation ← Translate(tweet_text, dest='en')
    Sent ← TextBlob(translation)
end if
if Sent.polarity ≥ 0.1 then
    return 1
else if Sent.polarity ≤ -0.1 then
    return -1
else
    return 0
end if
```

Algorithm 3 TextBlob_Sent function.

Where:

- SenticNet(): is a function that initializes the Senticnet dictionary for the English language.
- BabelSenticNet(): is a function that initializes the Senticnet dictionary for any language other than English.
- sn.polarity_value(w): returns the degree of polarity(sentimental degree) of the word w using the Senticnet dictionary.
- TextBlob(): initializes the Textblob library using the tweet's text. sent.polarity: returns the sentimental degree of the tweet's text.

4 Experimental results

To verify the advantages of our approach, this section will present experimental results. The majority of papers published in the literature use machine learning algorithms and classification methods to find sentiments of sentences or social networks content (tweets, Facebook comments, and publications, etc), as presented in the last section our method uses a recommendation approach based on Textblob and a dictionary-based method by proposing four new tweets' features to retrieve sentiments from covid-19 tweets in morocco.

The proposed approach is based on the notions of the collaborative filtering method which is based on the preferences of neighbors of the target tweet to find the relevant class. Each tweet to classify will be presented with a vector with four elements, and each element accepts 3 values (1 for positive, -1 for negative, and 0 for neutral). Based on the tweets vector we look for its neighbors in a labeled dataset and based on the labels of the k-nearest neighbors obtained, we recommend for our target tweet the relevant class (relevant sentiment).

```
lang ← Langdetect(tweet_text)
if lang =’en’ then
    sn ← SenticNet()
else if lang =’ar’ then
    sn ← BabelSenticNet(’ar’)
else if lang =’fr’ then
    sn ← BabelSenticNet(’fr’)
else if lang =’es’ then
    sn ← BabelSenticNet(’es’)
end if
S ← 0
C ← 0
for all w ∈ tweet_text do
    Sent ← sn.polarity_value(w)
    S ← S+Sent
    C ← C+1
end for
Moy=S/C
if Moy ≥ 0.1 then
    return 1
else if Moy ≤ -0.1 then
    return -1
else
    return 0
end if
```

Algorithm 4 SenticNet_sent function.

To demonstrate the strengths of our proposed approach, we compare it with four machine learning algorithms (SVM support vector machine, NB naive Bayes, RF random forest, DT decision tree) by calculating the accuracy. Figure 7 shows the results obtained.

According to this figure, the proposed approach using a recommendation method with the four proposed new measures(new tweets' features) gives good results with an accuracy that reaches 86%. Our model outperforms the well-known machine learning algorithms(the best of them gives only 65% accuracy).

Figure 8 shows how using four new features improves our approach (see Fig. 6 for more information about the four proposed features). We conclude that by increasing the number of features the error rate of our model decreases. By using 2 features the error rate is 42%, Its value is reduced to 36% using 3 features, and by using the four features, we find the best result with an error rate equal to 14%. All that demonstrates why we choose to work with the four features.

As mentioned at the beginning of this article, our work aims to extract and analyze the sentiments of Moroccan people during the covid-19 pandemic. For doing that, we collect tweets based on periods. For example, the first period to analyze is from 20 mars 2020 until 30 mars 2020.

In Fig. 9, we present an overview of the different analyzed periods.

According to Fig. 9, we use 12 periods from March to the end of August. In each period we extract all the tweets related to the covid-19 epidemic and published in morocco(using

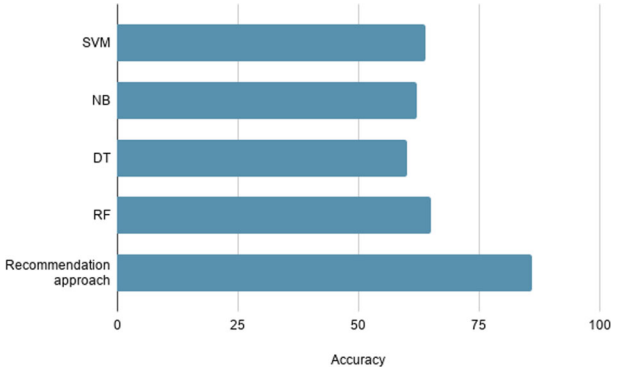

Fig. 7 The comparison result

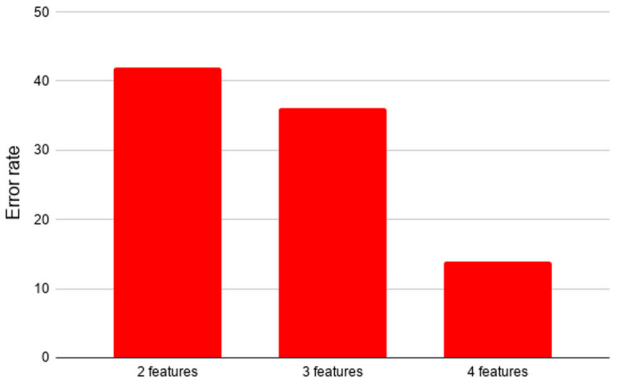

Fig. 8 How using 4 new features improve our approach

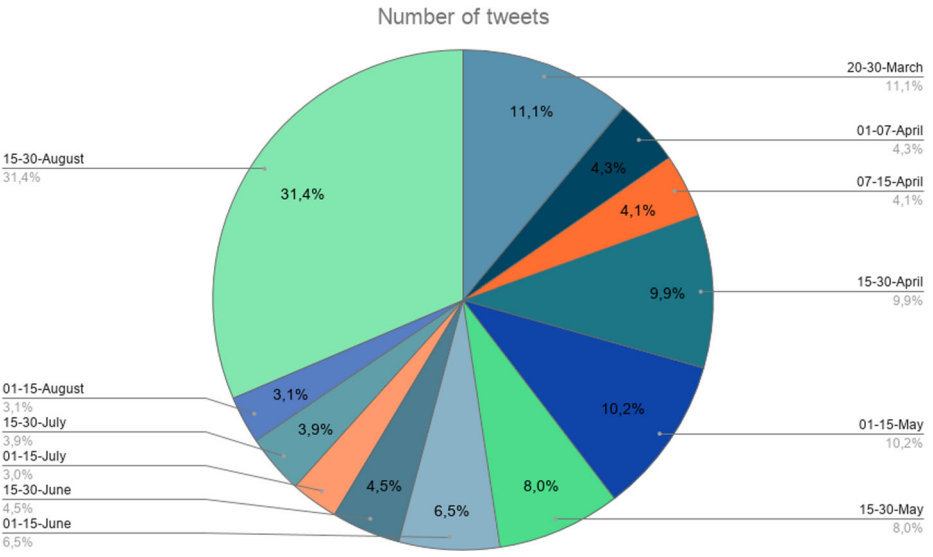

Fig. 9 Different analyzed periods

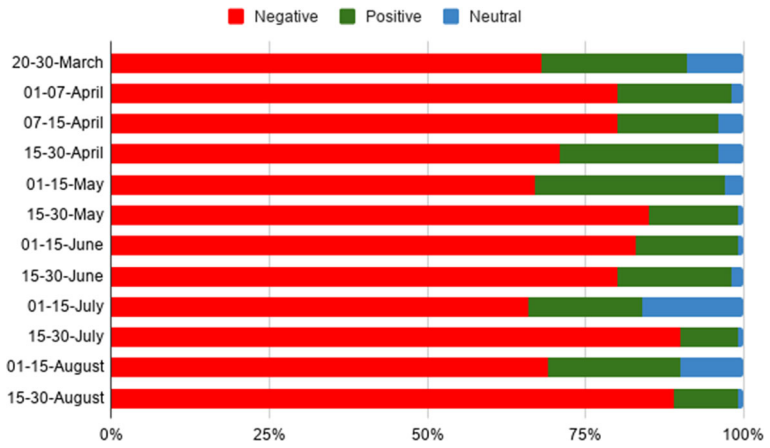

Fig. 10 Classification results

geo-coordinates). This figure shows the number of covid-19 tweets published in each period. For example, in the first period(from 20 to 30 March 2020) we have 11.1% of the total of tweets published in all periods, and in the last period(from 15 to 30 August) we find a percentage of 31.4%.

To extract the sentiments of people on the covid-19 epidemic, we apply our proposed approach to the tweets of each period, that is applying all the steps described earlier from the spell check and the text preprocessing methods to the step of classification based on the proposed four features(sentiment, sentiment_hashtag, sentiment_sentic_text, sentiment_sentic_hashtag) and our recommendation approach. We classify each tweet into three classes: positive, negative, or neutral.

Figure 10 shows the results obtained after we classify the tweets of all the periods using our proposed approach.

The remarks that we can extract from Fig. 10 are as follows:

- In all periods, the degree of negativity is greater than the value of positivity, which means that people have a negative feeling with a lot of fear of this new epidemic.
- The percentage of negativity is changed from a period to another, and that is due to the change of the figures (number of cases contaminated by the virus in Morocco in a given period ... etc).
- In the last periods, The percentage of negativity reached 90% and that is due to the second wave of the virus in Morocco.

5 Conclusion

In addition to the medical and health crisis due to the spread of the coronavirus, there is also a negative development of the feelings of people all over the world. Analyzing how the covid-19 crisis affects the sentiments of users on Twitter is becoming an important research axis from the beginning of this crisis until today, especially at the start of the second wave. In this article, we have proposed a new approach for analyzing the sentiments of Moroccan users on Twitter from the beginning of the covid-19 crisis in Morocco in March 2020 until the end of August 2020.

The proposed method is based on a new collaborative filtering approach that uses four new tweets' features using the TextBlob python library and a dictionary-based approach with the SenticNet dictionary. Our approach outperforms the well-known classification methods with an accuracy equal to 86%. By applying our method on Moroccan covid-19 tweets, we find that the majority of published content on Twitter related to covid-19 is negative.

The future work consists in proposing a new approach that deals with the elongation of words in tweets, and using it as a new parameter to improve the quality of classification of tweets.

Data Availability The datasets generated during and/or analysed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

Research involving human participants and/or animals This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Aslam F, Awan TM, Syed JH, et al (2020) Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. Humanit Soc Sci Commun 7(23). https://doi.org/10.1057/s41599-020-0523-3
2. Chakraborty K, Bhatia S, Bhattacharyya S, Platos J, Bag R, Hassanien AE (2020) Sentiment analysis of COVID-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media, Applied Soft Computing 97 Part A
3. Das S, Dutta A (2020) Characterizing public emotions and sentiments in COVID-19 environment: a case study of India. J Hum Behav Soc Environ. https://doi.org/10.1080/10911359.2020.1781015
4. de las Heras-Pedrosa C, Sánchez-Núñez P, Peláez JI (2020) Sentiment analysis and emotion understanding during the COVID-19 pandemic in Spain and its impact on digital ecosystems. Int J Environ Res Public Health 17(15):5542. https://doi.org/10.3390/ijerph17155542
5. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, Hon SD, Park J, Dang P (2020) Lipsky MS social network analysis of COVID-19 aentiments: application of artificial intelligence. J Med Internet Res 22(8):e22590
6. Imran AS, Daudpota SM, Kastrati Z, Batra R (2020) Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. In: IEEE Access, vol 8, pp 181074–181090. https://doi.org/10.1109/ACCESS.2020.3027350
7. Jelodar H, Wang Y, Orji R, Huang S (2020) Deep sentiment classification and topic discovery on novel Coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. IEEE J Biomed Health Inform 24(10):2733–2742. https://doi.org/10.1109/JBHI.2020.3001216
8. Kaur S, Sikka G, Awasthi LK (2018) Sentiment analysis approach based on N-gram and KNN classifier, 2018 first international conference on secure cyber computing and communication (ICSCCC), pp–4. https://doi.org/10.1109/ICSCCC.2018.8703350
9. Kruspe A, Häberle M, Kuhn I, Zhu XX (2020) Cross-language sentiment analysis of European Twitter messages duringthe COVID-19 pandemic, arXiv:2008.12172
10. Kumar S, Kumar K (2018) IRSC: integrated automated review mining system using virtual machines in cloud environment, 2018 conference on information and communication technology (CICT), pp 1–6, https://doi.org/10.1109/INFOCOMTECH.2018.8722387
11. Lyu X, Chen Z, Wu D, Wang W (2020) Sentiment analysis on chinese Weibo regarding COVID-19. In: Zhu X, Zhang M, Hong Y, He R (eds) Natural language processing and chinese computing. NLPCC 2020. Lecture Notes in Computer Science, vol 12430. Springer, Cham. https://doi.org/10.1007/978-3-030-60450-9_56

12. Madani Y, Erritali M, Bengourram J et al (2020) A multilingual fuzzy approach for classifying Twitter data using fuzzy logic and semantic similarity. Neural Comput & Applic 32:8655–8673. https://doi.org/10.1007/s00521-019-04357-9
13. Madani Y, Ezzikouri H, Erritali M et al (2020) Finding optimal pedagogical content in an adaptive e-learning platform using a new recommendation approach and reinforcement learning. J Ambient Intell Human Comput 11:3921–3936. https://doi.org/10.1007/s12652-019-01627-1
14. Manguri KH, Ramadhan RN, Mohammed Amin PR (2020) Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan J Appl Res 5(3):54–65
15. Mostafa L (2021) Egyptian student sentiment analysis using Word2vec during the Coronavirus (Covid-19) Pandemic. In: Hassanien AE, Slowik A, Snášel V, El-Deeb H, Tolba FM (eds) Proceedings of the international conference on advanced intelligent systems and informatics 2020. AISI 2020. Advances in Intelligent Systems and Computing, vol 1261. Springer, Cham. https://doi.org/10.1007/978-3-030-58669-0_18
16. Muthusami R, Bharathi A, Saritha K (2020) Covid-19 outbreak: tweet based analysis and visualization towards the influence of coronavirus in the world. Gedrag en Organisatie 33(2). https://doi.org/10.37896/GOR33.02/062
17. Nemes L, Kiss A (2020) Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication. https://doi.org/10.1080/24751839.2020.1790793
18. Pokharel BP (2020) Twitter sentiment analysis during Covid-19 outbreak in Nepal. Available at SSRN: https://ssrn.com/abstract=3624719 or https://doi.org/10.2139/ssrn.3624719
19. Samuel J, Ali GG, Rahman M, Esawi E, Samuel Y (2020) Covid-19 public sentiment insights and machine learning for tweets classification. Inf 11(6):314
20. Sharma S, Kumar P, Kumar K (2017) LEXER: LEXicon based emotion analyzeR. In: Shankar B, Ghosh K, Mandal D, Ray S, Zhang D, Pal S (eds) Pattern recognition and machine intelligence. PReMI 2017. Lecture Notes in Computer Science, vol 10597. Springer, Cham
21. Wang T, Lu K, Chow KP, Zhu Q (2020) COVID-19 Sensing: negative sentiment analysis on social media in china via BERT model. In: IEEE Access, vol 8, pp 138162–138169. https://doi.org/10.1109/ACCESS.2020.3012595
22. Youness M, Mohammed E (2018) Semantic indexing of a Corpus. Int J Grid Distrib. Comput. 11(7):63–80