Efficient Twitter Data Cleansing Model for Data Analysis of the Pandemic Tweets

Chapter · March 2021

DOI: 10.1007/978-3-030-67716-9_7

CITATIONS
7

READS
2,375

4 authors:

Belal Abdullah Hezam Murshed
University of Mysore
16 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Suresha Mallappa
University of Mysore
16 PUBLICATIONS **94** CITATIONS
SEE PROFILE

Osamah Ali Mohammed Ghaleb
Mustaqbal University
8 PUBLICATIONS **38** CITATIONS
SEE PROFILE

Hasib Daowd esmail Al-ariki
Sana'a Community Collage
12 PUBLICATIONS **122** CITATIONS
SEE PROFILE

Efficient Twitter Data Cleansing Model for Data Analysis of the Pandemic Tweets

Belal Abdullah Hezam Murshed, Suresha Mallappa, Osamah A. M. Ghaleb, and Hasib Daowd Esmail Al-ariki

Abstract Twitter data generally tends to be unstructured and often very noisy, cluttered/disorganized, and clothed in informal language. In this paper, we propose an intelligent Twitter data cleansing model that can solve data quality problems associated with twitter text. This model can correct a wide variety of anomalies from slangs, typos, Elongated (repeated Characters), transposition, Concatenated words, complex spelling mistakes as unorthodox use of acronyms, manifold forms of abbreviations of same words, and word boundary errors. The effects of whole range of tasks of Twitter Data Cleansing Model (TDCM) on the performance of sentiment classification utilizing feature models and three common classifiers have been investigated and evaluated. We conducted our experiments on two sets of pandemics twitter datasets: COVID-19 and Dengue datasets. The primary objective of this paper is to both increase the accuracy and the quality of twitter data and to purify and cleanse twitter data for further analysis. The experiment results seem to indicate that the accuracy of sentiment classification increases once the data quality problems associated with the Twitter text are solved. In COVID-19 twitter dataset, the best performance obtained using Random forest classifier after cleansing the data in terms of accuracy, recall, and f1-score are found to be at 84.7%, 88.5%, and 86.3% respectively. However, the best performance in terms of precision at 84.5% was observed using SVM classifier when compared to that obtained with other

B. A. H. Murshed (✉) · S. Mallappa
Department of Studies in Computer Science, University of Mysore, Mysore, Karnataka, India
e-mail: belal.a.hezam@gmail.com

S. Mallappa
e-mail: sureshasuvi@gmail.com

O. A. M. Ghaleb
College of Engineering and Computer Science, Almustaqbal University, Qassim, Saudi Arabia
e-mail: oaghaleb-t@uom.edu.sa

H. D. E. Al-ariki
Al Saeed Faculty for Engineering and Information Technology, Taiz University - Taiz-Technical Community College, Taiz, Yemen
e-mail: hasibalariki@gmail.com

I. Arpaci et al. (eds.), *Emerging Technologies During the Era of COVID-19 Pandemic*, Studies in Systems, Decision and Control 348,
https://doi.org/10.1007/978-3-030-67716-9_7

classifiers. Further, in the Dengue twitter dataset, the best performance for cleansing data in terms of accuracy, precision and f1-score are observed to be 81.7%, 83.7% and 88.6% respectively using Random forest classifier. The best performance in terms of recall, however, is 94.9% and was obtained using SVM classifier when compared with those obtained with other classifiers.

Keywords Data cleansing · Data quality problem · Twitter · Natural language processing · Slang · Elongation · Data transformation · Informal data · COVID-19

1 Introduction

The increasing popularity of social media network and user-created web content has been generating massive quantities of data that are advantageous for a variety of applications such as sentiment analysis, information sharing, topic modeling, knowledge about the current affair. Most applications that use twitter data anticipate the tweets to be high-quality texts which are accurate, readable, and effective. For instance, high-quality data is desirable for training purposes, especially for topic modeling, and classification primarily because since high-quality dataset (training) enables training the classifier model to obtain a higher level of classification performance [1].

Training high-quality tweet classification requires needs high quality of data. Unfortunately, Twitter data tends to be noisy, extremely unstructured, and presents several challenges to traditional natural language processing technologies [2]. Twitter content is characterized by heterogeneity, which means that each tweet contains different kinds of entities, e.g., text, location, user, link, and hash-tag. Besides, twitter is replete with informal language with many words appearing in their abbreviated forms, e.g. “plz” short for “please”. Since conventional text analysis technologies primarily concentrate on structured and formal texts, it is unrealistic to expect the same performance in tweet data. Thus, from the perspective of data analytics, Twitter data poses serious challenges and requires data cleansing tasks prior to building quality Twitter data set [3, 4]. In many domain-specific applications, text documents are typically unstructured and are casually worded with a lot of spelling grammar and mistakes [1]. In such cases, automatic detection of typographical errors correction of these error continues to be a challenging issue, especially in cases of unstructured text documents that include abbreviations, symbols, and acronyms related to application domains [1].

Prior to using Twitter data, it is necessary to cleanse and convert the data into a more formal and a more usable structured dataset to ensure that the subsequent stages of the process are easy, efficient, smooth, and effective. As a modern type of social media network, Twitter has many special characteristics, which renders it difficult to be dealt with using conventional text process technologies. In data analytics such as sentiment analysis, semantic analysis, tweet classification, topic discovery on tweets, feature extraction of the efficiency and efficacy are important

to ensure both the performance and quality of the model. Research shows that Twitter data preprocessing can substantially impact model output [5, 6]. For instance, in order to analyses public discourse trends and patterns, a particular event or news story related to an individual can be defined such that she/he can determine and assess the manner in which it is being publicly debated. Nonetheless, owning to the dynamic nature of user-generated content, the user should consider the abbreviations, hashtags, and slang that could be related to the subject of interest [3].

Data of low quality could be dangerous because it can lead to incorrect or missing decisions, operations and strategies. Further, it can also slow down, the innovation processes. The losses for organizations caused by low data quality are estimated to be over billions of dollars per year [7]. The issue of bad data is a huge problem troubling close to 60% of enterprises suffering from it [8, 9]. Even though the process of data cleansing and methods have been research topics of interest for quite some time, the features of Twitter tend to reduce it to a less cumbersome it a non-trivial work. The key to leveraging the potential value presented by effective data analytics is the very process and methodology of data cleansing based on which analysis can be carried out. This, good data quality is fundamental to obtain good insights. The chief contribution of this work is the development of an intelligent Twitter data cleansing technique which can solve data quality problems associated with twitter data. As a means of justification for a need for a novel Twitter data purification model, those challenges typical of Twitter datasets but not very common in traditional datasets will be demonstrated.

The problems of Twitter Text Quality

Recent research has identified the quality of data to be a serious challenge commonly encountered in sentiment analysis and topic discovery [10]. Twitter posts tend to be inherently poor in terms of data quality which poses serious challenges to dealing and processing those data prominently featuring in big Twitter dataset. The statement of the problem can be described as follows:

How can we solve the data quality problems associated with Twitter text?

Twitter texts fail to adhere to standard rules of vocabulary, spelling and syntax. The twitter tweets are very short and usually written hurriedly in very short period of time, often in seconds or minutes, by people from varying education background and interests. As tweet writers fail to pursue the rules of grammar, they tend to include misspelt words, abbreviations, slangs, and domain specific terminologies that are not found in Standard English dictionaries Such, and their texts are barely comprehensible to people outside their fields of application. Moreover, Twitter data is very rich with a wide range of linguistic innovations including abbreviations, lengthening, concatenated words, and emoticons.

To address this problem, this study proposes an intelligent Twitter data cleansing model that can solve the data quality problems associated with twitter text. This model can also correct a wide variety of anomalies such as slangs, emoticons, typos, Elongated (repeated Characters), transposition, Concatenated Words, complex spelling mistakes, unorthodox use of acronyms, and manifold forms of

abbreviations of the same words. It makes use of the generated knowledge to recognize unorthodox acronyms, and to string together similar words (correctly spelled and misspelled). This study also investigates and evaluates the effects of all the tasks associated with the Twitter Data Cleansing Model (TDCM) on the performance of sentiment classification utilizing three common classifiers. The experiment results seem to indicate that the performance of sentiment classification tends to increases after solving the data quality problems associated with the Twitter text.

The rest of this article is structured as follows: In Sect. 2, the literature survey related to preprocessing area is provided. Section 3 describes the proposed model (TDCM) which increases the data quality and cleanses twitter dataset. The details of the data collection, performance evaluation, and the experimental results are presented in Sect. 4. In Sect. 5, a concise conclusion is given.

2 Related Work

Some of the earlier works in literature have proposed different methods of pre-processing texts. Hemalatha et al. [11] proposed a method to perform certain common pre-processing tasks such as removal of special characters and URL, tokenization, and stemming in twitter to achieve sentiment analysis. In [12], Sun et al. suggested a preprocessing method on online financial text which can perform removal of numbers, URL, and punctuation, tokenization, extensions of contractions, removal of stop word, and lemmatization. Several pre-processing tasks such as feature correlation, n-gram models, and stemming in Arabic texts were suggested by Duwairi and El-Orfali [13]. All these tasks were carried out to reconnoiter the effects of the performance of sentiment analysis in Arabic. In [14], Rushdi-Saleh et al. proposed various Pre-processing steps such as stop word removal, n-gram generation, and stemming on movie reviews collected from various Arabic blogs and web pages. Jianqiang [15] debated the impact of various tasks of Preprocessing such as URL, negation, stop word, and repeated letters on the performance of sentiment classification conducted on twitter dataset. Several steps were presented by Lizaet et al. [16]. These tweets, after various stages of preprocessing, were transformed into a vector set of features utilizing Bag of Words (BOW). Several semantic similarity measures proposed by Murshed [17]. A pre-processing approach based on gathering of words of slang with coexisting words to determine the importance and the strength of sentiment of slang words used in tweets was suggested by Tajinder Singh [18]. Itisha et al. [19] presented a two-phased preprocessing approach, where in the first stage involves noise removals (URLs, punctuation marks, usernames, and elimination of stop words). The second stage involves normalization which includes the transformation of Non-standard words into formal words. Hussein et al. [20] proposed an approach to clean twitter data

which includes several tasks such as segmentation of English tweets, elimination of stop words, cleaning, and stemming. The aim of [21] is to find out the advantages of various, twitter-specific, preprocessing techniques to preprocess the tweets efficiently rather than utilizing baseline text preprocessing methods. A preprocessing method, which consists of data collection, data cleaning, tokenization, stemming, and removal of stop words, was proposed by Naresh [22] for unstructured healthcare text data. Arpaci et al. [23] proposed a method utilizing an evolutionary clustering analysis aimed to analyze the COVID-19 Twitter dataset collected in a period of time and to describe the trend of public attention given to topics related to the COVID-19 pandemic. Further, Arpaci et al. [24] suggested an new type of the specific phobia diagnosis and initial tests for the psychometric properties of the COVID-19 Phobia Scale (C19P-S).

3 Proposed Model

3.1 Twitter Data Cleansing Model

The proposed twitter data Cleansing model starts with data collection obtained from Sect. 4.2. Our Twitter data cleansing model consists of four Phases (as shown in Fig. 1), namely, (1) Extraction and Filtering Tweets, (2) Removal of Noise, (3) Out of Vocabulary Cleansing (cleansing irrelevant vocabulary), (4) Transformation of tweets. These phases facilitate the process of saving storage space and analyzing time, (especially in cases of an enormous dataset), and also reducing data. Apart from enhancing the accuracy of the results of data analysis, these stages also ensure developing topic coherence. In addition of these advantages, these four phases transform the informal parts of tweets to formal ones. The complete details of each of these phases are described in the following sections.

3.1.1 Extraction and Filtering Tweets

Though the collected streaming tweets could be in various languages, our model, takes into account only the English tweets for analysis. Tweepy package permits specific language tweets to be fetched. The language parameter value such as ‘en’ is then assigned for those tweets with full texts collected in English language. We also observed the occurrence of many duplicate tweets when the tweeters tend to re-tweets the same text of tweets for other tweeters (called Re-tweet). Hence, to eliminate duplication, we used regular expression techniques based on NLP to seek hyperlinks in the text of tweets and to eliminate the same. Following this, the duplication tweets were dropped since they seemed to add no new knowledge to the dataset, thus proving to be inefficient in computational terms. Excluding these

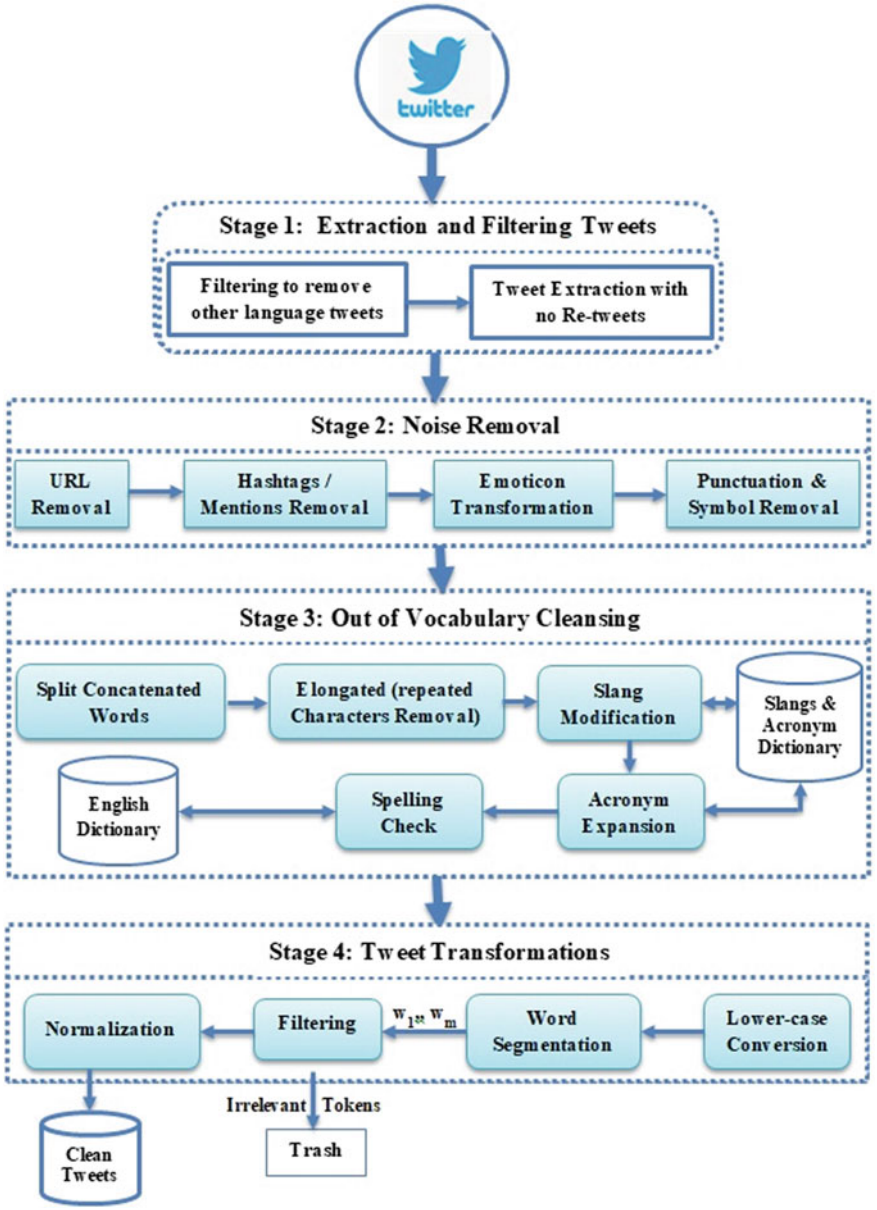

Fig. 1 Twitter data cleansing and transformation model

duplicate tweets tend to renders the twitter dataset to become all the more meaningful.

3.1.2 Noise Removal

- **URL removal**

Twitter tweets usually contain a lot of noise and unavailable data such as Uniform Resource Locator (URLs) for analysis. The embedded URL is commonly utilized to provide the source with the elaborated depiction of the content stated in the tweet—for instance, "Five lectures the Spanish flu can educate us about COVID-19 https://t.co/ANuMMUlPy7". However, the model in this study fails to take the URL for analyzing and extracting topics from tweets into account. Hence all these unnecessary data are deleted from tweets by applying preprocessing steps using a regular expression based on NLP. We only focused on meaningful English words containing in the tweets.

- **Hashtags/mentions removal**

Analysis tweets (system) and extracting topics from tweets need meaningful words that refer to a topic. Therefore, it is necessary to delete all unnecessary terms and signs from tweets. Twitter contains some of these terms/words which start with the signs '@' and '#'. User mentions are generally used in tweets to refer to someone else and these user mention entities start with a '@' sign, followed by the user's name e.g.: "Stay at home @Belal". Likewise, a hashtag is another entity associated with a tweet and refers to a specific tag or topic and usually begins with a '#' sign. This need not be a complete, meaningful word. At this stage of hashtags removal, these signs are eliminated using a regular expression in python language. However, our model restricts itself to deleting the hashtag sign, thus leaving the term intact. This is because the hashtag sign is typically followed by phrases or words describing the topic being discussed.

- **Emoticon and Emoji Transformation**

The text of tweet can include emojis which are encoded in Unicode and contains emoticons. These noises such as emoticons and emojis can affect the process of data purification and feature extraction. In the following example of a tweet, "Plz, #stayAtHome don't go outside. COVID-19 is verrrry dangerous u'\U0001f61f" contains Emoticon and ends with a Unicode encoded "worried face" icon. Both these emoticons and the emojis will be transformed into their respective word

format. We have used two wonderful emoji and Emoticons data dictionaries developed by NeelShah[1] in order to convert emoticons and emoji to word format and to utilize a regular expression based on the NLTK package in python platform.

- **Punctuation and symbol removal**

Generally Twitter tweets also include numbers, punctuations marks and symbols. Utilizing regular expression will eliminate all such inclusions and supply only the important data. This step further reduces the storage of the data set and holds only the effective data used for topic modeling and classification methods.

3.1.3 Out of Vocabulary Cleansing

The stage of anomaly detection determines the Out-of-vocabulary words/terms (i.e., those terms/words do not exist in the English dictionary) such as elongated words (e.g., stayyyyyyyyyy), slangs (e.g., plz, luv), concatenated words (e.g., StayAtHome). This phase involved many tasks, and each task is described in a concise manner.

a. **Concatenated Words**

In Twitter data, the tweet size has a character limit of 140 and tweets tend to contain very limited information typically needed for reliable extraction. Often times, two or more words (such as 'StayAtHome') are commonly concatenated to reduce the size of tweet. Such words/terms should be split into their components forms before utilizing them in data analytics. At this stage, we use regular expression techniques that split those words into their components for other analysis

b. **Contraction Replacement (Apostrophes)**

Text normalization is a much required step in data cleansing in order to correct errors in words or texts, Apostrophes give rise to significant word sense ambiguity since it may exemplify contraction for a word. For example, contractions such as "hasn't", "did't", and "he'll" commonly occur in tweets. Therefore, the Contractions in Twitter data should be transformed into regular lexicons before processing and this data is also very important for sentiment analysis. In this step, public lexicon that contains all contractions is used. Therefore contractions should be substituted with the corresponding extended expressions of the language. The first step to be considered is identifying the pattern, followed by substitution. Each instance of the pattern is substituted by an equivalent replacement pattern.

c. **Elongated Words**

Most users write their tweets and status updates informally resulting in the occurrences of repeated characters and Elongation in Tweets. At this step, those

[1] https://github.com/NeelShah18/emot/blob/master/emot/emo_unicode.py.

terms that are lengthened by an unneeded repetition of characters are shortened. These elongated terms are used to articulate emotions such as 'looooooove' and so on. Eliminating these unnecessary letters to emerge with a standard meaningful word is an important stage in preprocessing of tweets. Some of these methods have reduced the repeated characters to 2 by eliminating surplus letters. This step introduces an approach to convert these irritating Elongated (repeating) characters to a meaningful English word using a regular expression module known as back-references. A backreference is a popular method to refer to a previously matched group in a regular expression. This approach matches and excludes repeated characters. Before eliminating repeated characters recursively, a WordNet lookup is which guarantees the creation of removal of unnecessary letters. A class consisting of a replace () method which takes an individual word and returns a more proper and accurate version of that word is constructed, thus eliminating all the suspicious, recurring characters.

d. **Slangs Modifications**

The tweets text has certain anomalies that are not typically found in traditional text cleaning issues. Slang is one such anomaly, which is quite prevalent in Twitter posts and includes the use of abbreviations, acronyms or internet slang. These slangs usually tend to reduce the number of characters during writing tweets. For instance, incorrect/nonstandard words such as 'Plz', 'Gr8', and 'luv' are not found in English language and it is necessary to convert such slang words into correct lexicons in English, before rendering them beneficial for data analysis. To achieve this, we built a dictionary consists of 2864 slang which furnishes a set of all possible slangs as lookup dictionaries for transformation objectives. The slang lookup is carried out by contrasting it with the internet slang.

e. **Spelling Correction**

Yet another extremely significant step in the cleansing of twitter data is spelling correction. Typos are commonly prevalent in twitter posts and correcting these spelling errors is important before starting analysis. Pyspellchecker, a package in python, is used at this stage to correct spelling.

3.1.4 Tweet Transformations

a. **Lower-case conversion**

Lower case conversion is the process of converting words of the tweets to the same form. At this stage, all the tweets are transformed into lowercase in order to supply a constant format for all tweets utilizing python lower string function.

b. **Word segmentation(Tokenization)**

The second step of the Tweet transformations phase is the word segmentation process (Tokenization), which means dividing a tweets, sentences, and phrases into a collection of lexical units (Words) known as tokens that are both methodologically advantageous and linguistically significant. Tokenization is

typically the most significant and primary task in the plurality of text processing applications. We have used NLTK library [25] to tokenize the tweets in python language. For instance, consider the tweet "COVID-19 assaults immune system as HIV doctors afraid" When this tweet passes through tokenization, the outcome of this task would be the following tokens ['COVID-19', 'assaults', 'immune', 'system, 'as, 'HIV', 'doctors', 'afraid'].

c. **Filtering irrelevant words**

Tweets can include common and frequent words that might not contribute much to the meaning and importance of tweets. Such words/terms are known as stop words (some of these words, like 'is', 'he', 'an', 'she', 'a', etc.), which add more noise to NLP. Hence, it is important filtering these stop words from tweets during the Twitter data cleansing model. The removal of stop words is carried out by python, and all the stop words are eliminated from tweets utilizing the corpus of NLTK stop words, except for those that refers to positive or negative feeling.

d. **Normalization**

The normalization phase aims to decrease the variation in tweets data by transforming words/terms to their base form stem. In addition to that, the purpose of this stage is to reduce redundancy, eliminate disparate suffixes from tweets to economize time and storage space. For example, the words 'converter', 'converts', 'converted', 'converting', 'conversion' can be converted and stemmed to the word 'Convert'. In the tweets mining analysis, the stemming phase is the most popular phase because it assists in focusing the analysis on the basic form of the terms/words, instead of distinguishing between various terms/words which can introduce ambiguity in text/tweets mining methods. The Porter Stemmer algorithm [26], the most widely used approach in English, has been utilized in the proposed work.

3.2 Feature Extraction of the Model

The feature extraction model, which indicates a method that determines how certain features have been used to classify new data into a particular class, plays a significant role in proposing classification. Throughout the classification of the texts or tweets, various feature selection models were presented for sentiment analysis of the twitter data. Most investigators present state-of-the-art results on twitter data for sentiment analysis utilizing a unigram feature model [27, 28]. In this research, we have utilized word N-gram features models and Sklearn library to extract the features.

3.3 Twitter Sentiment Classification Process

To evaluate the quality and performance of proposed Twitter Data Cleansing model, we have utilized three common supervised techniques namely Multinomial Naïve Bayes (MNB), Random Forest (RF) Classifier, and Support Vector Machine (SVM) for performing sentiment analysis. Further, the performance of each of these classifiers has been assessed based on the following three metrics: Accuracy, precision, Recall, and F1-score. All classifiers have been implemented using a machine learning sklearn package in Python. For the process of classification, each twitter dataset has been split into two portions - one portion for training (80%) and 20% the other for testing. We have used the Sklearn library for classification purposes.

4 Experiments and Analysis

4.1 Experimental Setup

The experiments have been performed on an Intel Core i7-3210 M CPU @ 2.5 GHz machine with 16 GB RAM and Windows 10. These experiments have been executed using Python3.7 and Pycharm IDE. All graphics have been generated using Origin pro 8. We have performed our experiments on the Pandemics twitter datasets (as described in Sect. 4.2). The proposed data cleansing model has been incorporated with several tools such as a slang and acronym dictionary which supply a set of all versions of slangs, abbreviations as lookup dictionaries for transformation purposes and an English dictionary. Moreover, tools such as Twitter streaming API and "Tweepy" provide access to twitter to extract tweets. While Natural Language Toolki (NLTK) was used for common preprocessing, we have used packages such as SpellChecker that offer support for methods such as a 'spell', and 'correction' methods to check and correct the spelling in the words of tweets. Likewise, while the Sklearn library was used for feature extraction and classification, we have utilized sklearn.feature_extraction module to extract features form our pandemics twitter datasets.

4.2 Data Collection

To display the differences and effects of the proposed twitter data cleansing model, the following datasets were collected to verify the model. The Twitter Streaming Application Programming Interface (API) has been used to extract tweets from twitter depending on trending events for a period of three months starting from 13th Jan, 2020 to 30th April, 2020 with the use of language filter 'en' set to English using the relevant specific filter words provided as keywords. Tweepy is an

Table 1 Details of pandemic datasets

Description of datasets	Dataset name	No. of tweets
Corona Virus	COVID-19	501,231
Cholera	Cholera	26,068
Swine flu	Swine Flu	34,901
Dengue fever	Dengue fever	1,967
Malaria	Malaria	117,386
Ebola virus disease (EVD)	EVD	444,3
Chikungunya	Chikungunya	685
Total number of tweets		**686,681**

open-sourced Python library that enables python to access Twitter and uses its API. We just placed filter terms specifically for pandemic-related tweets such as COVID-19, Cholera, swine flu, dengue fever, malaria, Ebola Virus Disease (EVD), and Chikungunya. The complete details of the tweets collected in Pandemics dataset are given in Table 1.

4.3 Performance Evaluation

It is necessary to evaluate the Twitter data cleansing model of the Pandemic tweets to measure both the performance and efficiency of the proposed model. Appropriate performance metrics must be selected to evaluate the accuracy and quality of the data both before and after data cleansing. Therefore, the key metrics taken into account in this research are as follows: Accuracy, precision (P), recall (R), and also F1-score. These measures are calculated on the basis of the values of True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP) assigned classes. Accuracy can be formulated, as given in Eq. 1.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision metric quantifies as the number of true positive out of all positively assigned Tweets and this metric is expressed as in Eq. 2.

$$P = \frac{TP}{TP + FP} \tag{2}$$

Recall metric quantifies as the number of true positive out of the actual positive tweets. This metric can be defined as given in Eq. 3.

$$R = \frac{TP}{TP + FN} \tag{3}$$

Finally, f1-score provides a single score that combine both precision and recall into a single metric. F1-score is calculated as expressed in Eq. 3.

$$F1 - \text{score} = \frac{2 * P * R}{P + R} \tag{4}$$

where the values of its range are between 0 and 1 and the closer it is to 1, which means better results.

4.4 Experimental Results

The experimental results obtained have been compared in this section, and presented in two parts. The first part presents details of the empirical case study of the proposed TDCM and the second part presents details related to the evaluation performance and effectiveness of the proposed twitter data cleansing model on sentiment classification. Firstly, in our model, only English tweets are taken into account for analysis. We filtered all the tweets with language filter 'en' in the First Phase of our model to remove all non-English tweets from the Corpus Twitter dataset, as shown in Fig. 3 (1.1). Further, we have also removed all Re-tweets and duplication of tweets as shown in the sample of duplication in Fig. 3 (1.2). Thus, the number of tweets after removing duplication stand at 166,482, 7,047, 11,341, 1,078, 35,989, 1,676, and 261 tweets for each of these datasets COVID-19, Cholera, Swine Flu, Dengue fever, and Malaria, Ebola Virus Disease (EVD), and Chikungunya respectively, as shown in Table 2. Figure 2 shows the comparison of the number of tweets both before and after removal of duplication.

The second phase of our model is noise removal, which removes URLs, Hashtags "#", user mentions "@", emoticon transformation, punctuation marks

Table 2 Details of datasets after removing duplicate

Dataset name	No. of tweets	No. of tweets without duplicate
COVID-19	501,231	166,482
Cholera	26,068	7,047
Swine Flu	34,901	11,341
Dengue fever	1,967	1,078
Malaria	117,386	35,989
EVD	444,3	1,676
Chikungunya	685	261
Total # of Tweets	**686,681**	**223,874**

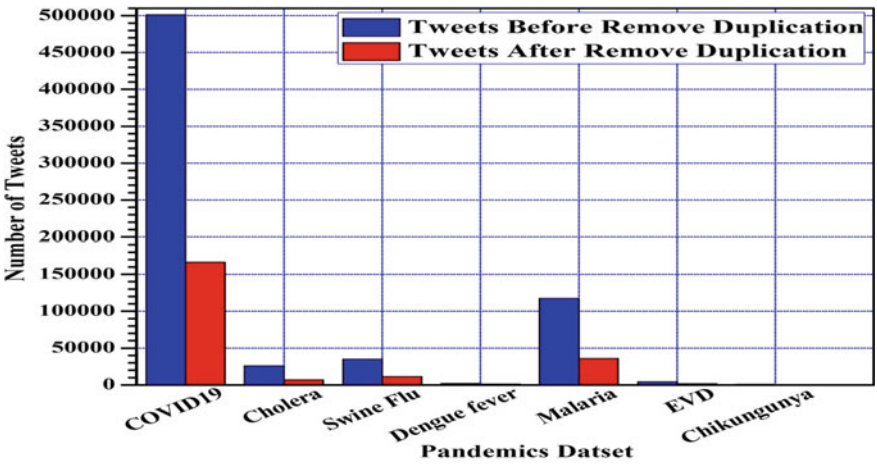

Fig. 2 Number of tweets before and after removing duplicates and re-tweets in pandemic datasets

and special characters as shown in Fig. 3 (2.1), (2.2), (2.3), (2.3), (2.4) respectively. The third phase of our model is cleansing of Out of Vocabulary words such as concatenated words. All the concatenated words have been split into the their component words, as shown in Fig. 3 (3.1) For example, the word "StayAtHooooooooommmme" is divided into three words "Stay", "At", and "Hooooooooommmme". Besides, all the characters that are repeated in elongated words have also been removed as shown in Fig. 3 (3.2), the word "Hooooooooommmme" converted to "home". The next step in this phase is modification of slang words during which informal words are converted to their formal equivalents words. For example abbreviations "ASAP" and "COVID" are converted to their respective formal equivalents namely "As soon as possible" and "coronavirus", as shown in Fig. 3 (3.3). The last step in this phase is correction of wrongly spelt words—all the errors in words such as "tronsmited" and "petient" are corrected to "Transmitted" and "patient", as shown in Fig. 3 (3.4). The final last phase of our model is Tweet Transformation which consists of Lower case conversion, word segmentation, and filtering irrelevant words and normalization, as shown in Fig. 3 (4).

Table 3 shows the number of tokens before cleansing tweets and the detailed results of the number of tokens in each phase of the twitter data cleansing framework for all the gathered datasets. Besides, Fig. 4 compares the token number in each task of the data cleansing model for all the datasets collected.

In the second part of experimental result, we have utilized three most commonly used supervised techniques namely MNB, RF and SVM classifiers to evaluate the performance and effectiveness of the proposed twitter data cleansing model on sentiment classification. We have also assessed the performance of each classifier based on the following metrics: Accuracy, precision, Recall, and f1-score. The

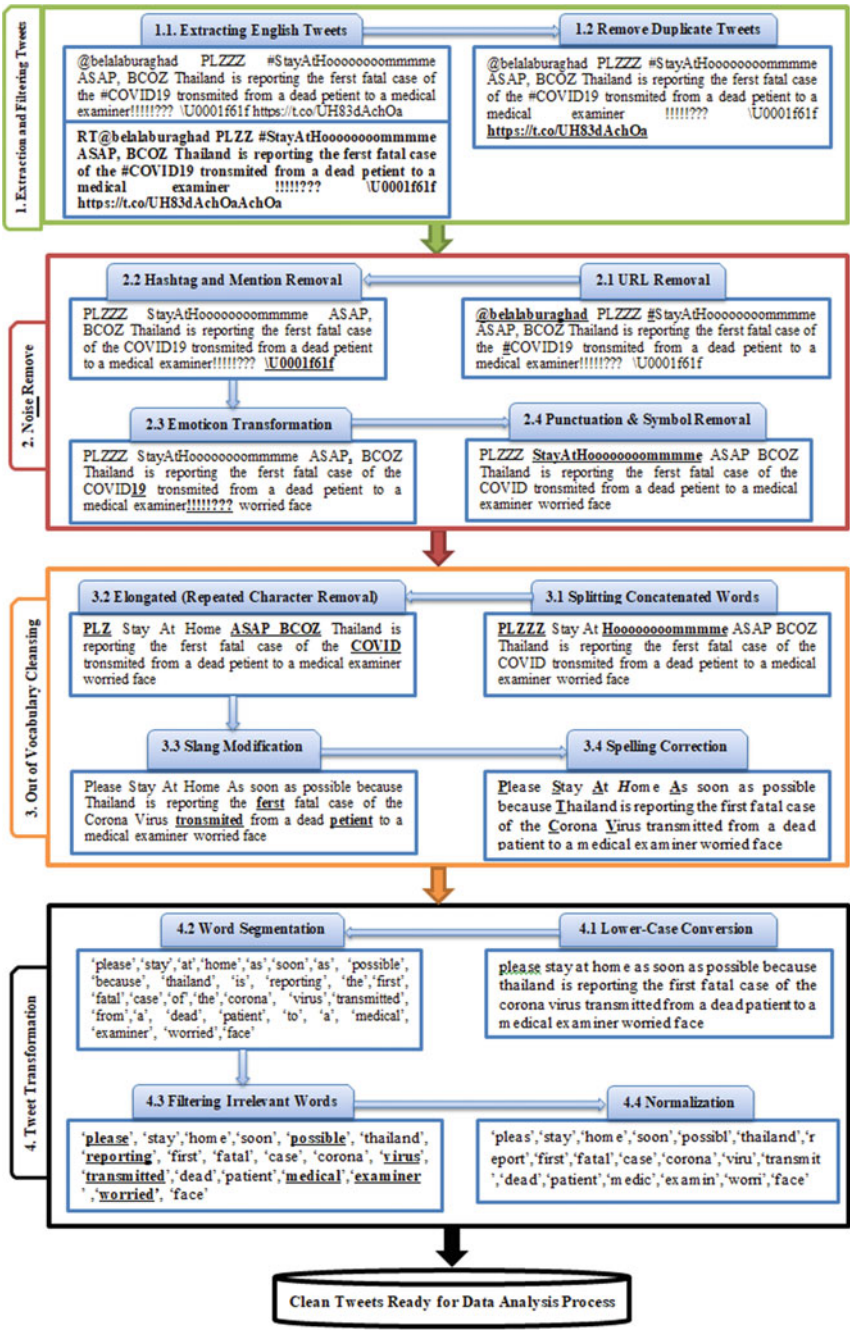

Fig. 3 Tasks of the twitter data cleansing model for a tweet of COVID-19 with tweet ID 1250049823614490000

Table 3 Number of tokens before and after each task of Twitter data cleansing model

Dataset name	COVID-19	Cholera	Swine Flu	Dengue	Malaria	EVD	Chikungunya
# of Tweets without duplicate	**166,482**	**7,047**	**11,341**	**1,078**	**35,989**	**1,676**	**261**
# of Token after remove duplicate tweets	4,281,084	202,690	345,741	31,685	973,411	53,538	7,995
# of Token after remove URL	4,053,410	191,689	325,392	30,508	916,545	51,366	7,642
# of Token after replace apostrophe	4,306,820	202,891	348,457	32,004	981,074	53,899	8,036
# of Token after splitting concatenate	4,069,551	188,870	312,690	30,271	930,839	50,676	7,627
# of Token after remove repeated letters	4,069,551	188,870	312,690	30,271	930,839	50,676	7,627
# of Token after slang modification	4,419,308	205,072	342,093	32,367	1,001,803	55,010	8,086
# of Token after spelling correction	4,419,396	205,080	342,104	32,369	1,001,855	55,013	8,078
# of Token after remove stop words	2,748,922	119,524	192,441	19,661	603,885	31,659	5,156

experiments have been conducted on 29,062 tweets chosen randomly from COVID-19 twitter dataset, and 1,078 tweets, chosen from the dengue twitter dataset.

All the tasks of the twitter data cleansing model on the three selected classifiers namely: MNB, SVM, and RF classifiers have been tested to show the performance of the proposed model. Table 4 shows a significant increase in the performance and quality of the classification after cleansing the data of COVID-19 twitter dataset in n-gram feature model, with the highest accuracy of 84.7% performed in the RF Classifier after cleansing the data, followed by 81.1% in SVM classifier and 74.5% in MNB classifier. Similarly, in the case of Dengue twitter dataset, it can be observed from Table 5 that the best levels of accuracy of this dataset after cleansing the data are at 81.7% performed in the RF Classifier, followed by 81.2% in SVM classifier and 79.7% in MNB classifier.

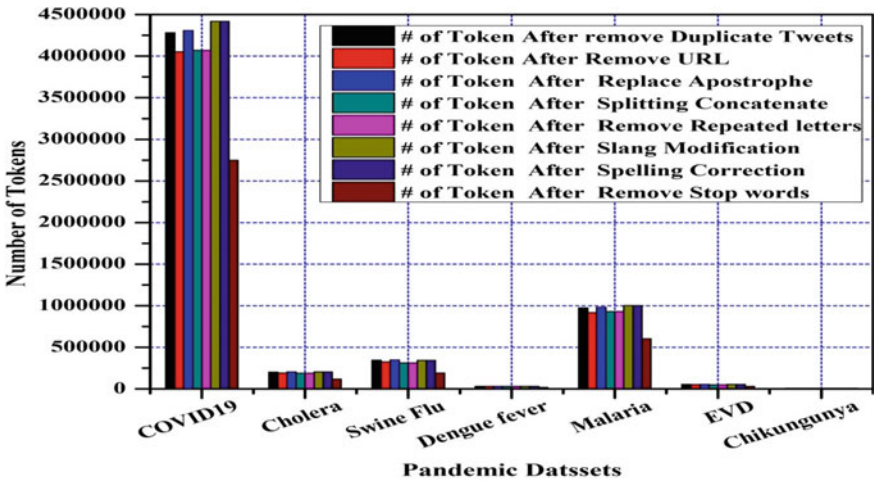

Fig. 4 Comparison of the tokens number in each task of the data cleansing model for each pandemic

Table 4 Classification results of the twitter data cleansing model for COVID-19 Twitter dataset using MNB, SVM, and RF classifiers

Methods metrics	MNB classifier		SVM classifier		RF classifier	
	Before date cleansing	After date cleansing	Before date cleansing	After date cleansing	Before date cleansing	After date cleansing
Accuracy	0.723	0.745	0.771	0.811	0.816	**0.847**
Precision (P)	0.733	0.748	0.801	**0.845**	0.817	0.842
Recall (R)	0.775	0.812	0.779	0.828	0.855	**0.885**
F1-score	0.754	0.778	0.790	0.836	0.836	**0.863**

Table 5 Classification results of the data cleansing model for dengue Twitter dataset using MNB, SVM, and RF classifiers

Methods metrics	MNB classifier		SVM classifier		RF classifier	
	Before date cleansing	After date cleansing	Before date cleansing	After date cleansing	Before date cleansing	After date cleansing
Accuracy	0.717	0.797	0.651	0.812	0.667	**0.817**
Precision (P)	0.738	0.829	0.649	0.822	0.673	**0.837**
Recall (R)	0.725	0.795	0.925	**0.949**	0.904	0.942
F1-score	0.731	0.815	0.763	0.881	0.772	**0.886**

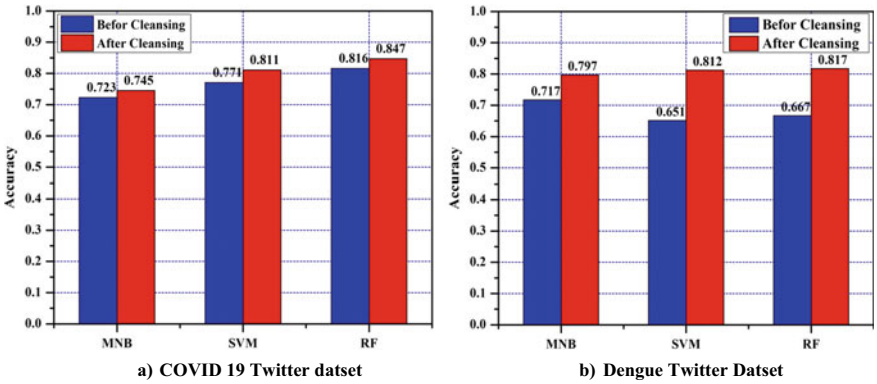

Fig. 5 Classification accuracies results using NBM, SVM, and random forest (RF) classifiers, **a** COVID-19 twitter datset, **b** dengue twitter datset

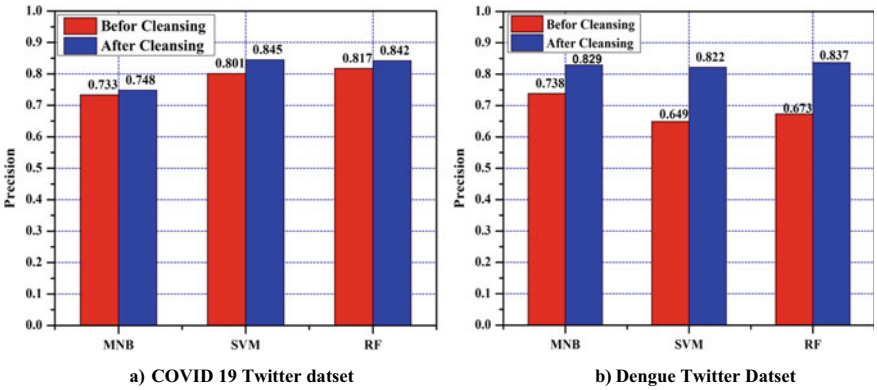

Fig. 6 Classification precision results using NBM, SVM, and random forest (RF) classifiers, **a** COVID-19 twitter datset, **b** dengue twitter datset

Figure 5 (a) and (b) show a significant increase in accuracy in both datasets. It can also be seen that quality in terms of accuracy in COVID-19 is higher when compared to the dengue twitter dataset. Likewise, the Recall performance of Dengue dataset using SVM classifiers is close to 1 around 94.9% that indicate a high performance of the classification achieved using SVM classifier, while the best recall result of the COVID-19 is 88.5% obtained using the RF classifier.

Similarly, Fig. 6 (a) and (b) show that COVID-19 and Dengue datasets has the highest precision after cleansing the data when compared to the methods before cleansing. The precision values obtained using the SVM at 0.845 seems to indicate that this classifier is the best classifier on COVID-19. Precision on the dengue dataset using the RF classifier is at 0.837, which seems to indicate that the RF classifier is the best classifier on the Dengue dataset.

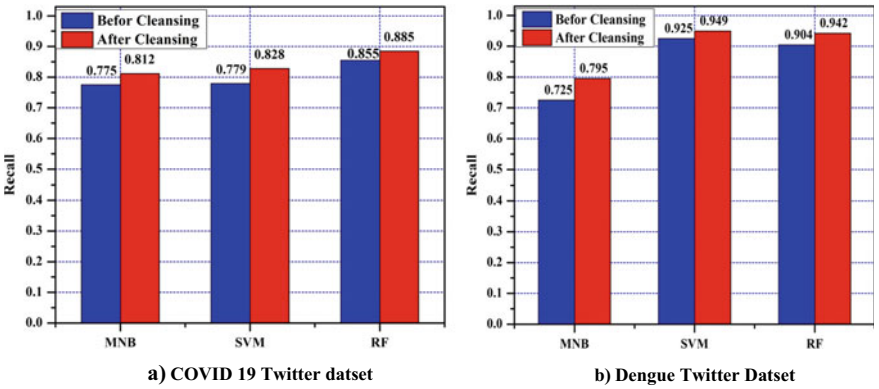

Fig. 7 Classification recall results using NBM, SVM, and random forest (RF) classifiers, **a** COVID-19 twitter datset, **b** dengue twitter datset

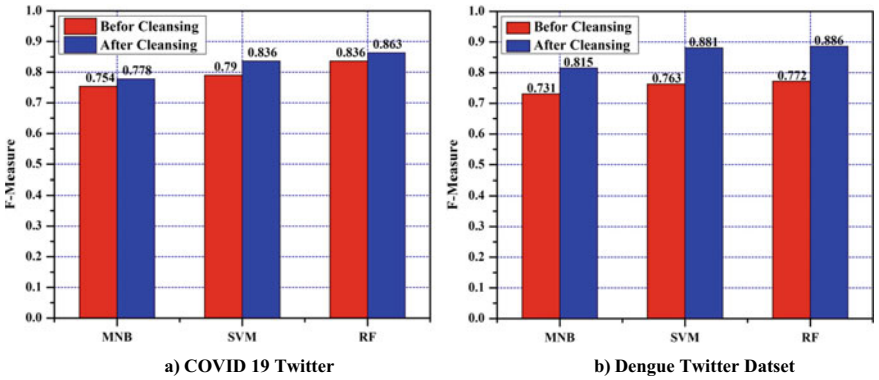

Fig. 8 Classification F1-score results using NBM, SVM, and random forest (RF) classifiers, **a** COVID-19 twitter, **b** dengue twitter datset

Table 6 The percentage of increasing accuracy, precision, recall and f1-score after cleansing to before cleaning dataset for classification

Datasets	COVID-19 Twitter dataset			Dengue Twitter dataset		
Classifiers metrics	MNB	SVM	RF	MNB	SVM	RF
Accuracy	2.2	**4.0**	3.1	8.0	**16.1**	15.0
Precision (P)	1.5	**4.4**	2.5	9.1	**17.3**	16.4
Recall (R)	3.7	**4.9**	3.0	7.0	2.40	**3.80**
F1-score	2.4	**4.6**	2.7	8.4	**11.8**	11.4

From Fig. 7 (a), it is evident that after cleansing the data, RF classifier has higher values of recall on COVID-19 Twitter dataset, while SVM classifier has higher values of recall on Dengue dataset after cleansing when compared to others classifiers, as shown in Fig. 7 (b). With respect to the parameter of recall, the recall of Dengue dataset shows higher values than that of COVID-19 dataset.

Finally, the performance of F1-score of both datasets COVID-19 and Dengue are shown in Fig. 8 (a) and (b). Hence, the use of Twitter Data Cleansing Model followed by filtering the texts data reduces noise and solves the problem of data quality problem associated with twitter texts data and also improves the performance of classification.

Table 6 shows that the percentage of increasing performance in terms of accuracy, precision, recall, and F1-score of all classifiers after the use of Twitter Data Cleansing Model in the n-grams feature model on COVID-19 and Dengue twitter datasets. At the end of all tasks in the proposed model on COVID-19 twitter dataset, the maximum performance with respect to accuracy was found to be at 4.0%, obtained using SVM classifier, when compared to those obtained with MNB and RF classifiers. In dengue dataset, the maximum improvement of accuracy using SVM was at 16.1%, when compared to those obtained with other classifiers. Further, cleansing the data on the COVID-19 dataset seems to increase Precision, Recall, and F1-score of MNB, SVM, and RF and the desired levels of improvements in terms of Precision, Recall, and F1-score using the SVM were found to be at 4.4%, 4.9%, and 4.6%. In addition, the performance of MNB, SVM, and RF in terms of Accuracy, precision, recall, and F1-score appear to increase on the Dengue dataset. However, using SVM classifier on the dengue twitter data set seems to increase Accuracy, precision and F1-score increase at 16.1%, 17.3% and 11.8% respectively. In terms of Recall, the best performance using RF classifier can be seen to be at 3.8%.

5 Conclusion

This study presented an efficient Twitter Data Cleansing Model (TDCM) which can solve the data quality problems associated with twitter data and extract both qualitative and quantitative data from the Twitter social media. The data was in the form of tweets chosen on trending topics such as COVID-19, cholera, etc. The presence of slang, acronyms, bad grammar, typos, duplications, concatenated words, repeated characters in a word, compounded with the undesirable content such as URLs, expressions, stop words, etc. present in these datasets make it incredibly difficult to obtain meaningful insight from these data sets. The challenges associated with twitter data were solved using the proposed TDCM Model, which consists of four phases namely Extraction and Filtering Tweets, noise removal, Out of vocabulary cleansing and tweets transformation. Each part of a tweet was converted into an enhanced form before being utilized as an input to the data analytics and mining systems. Comprehensive experimental results that show that the

appropriate TDCM can significantly improve the quality and efficiency of data have also been presented. In COVID-19 twitter dataset, the best performance in terms of accuracy, recall, and f1-score using Random forest classifier was observed to be at 84.7%, 88.5%, and 86.3% respectively. However, the best performance in terms of precision using SVM classifier was found to be at 84.5%. In the case of Dengue twitter dataset, the best performance in terms of accuracy, precision and f1-score using Random forest classifier were found to be at 81.7%, 83.7% and 88.6% respectively. However, the best performance in terms of recall using SVM classifier was found to be at 94.9% when compared to those obtained using other classifiers. Future studies could focus on analyzing the effect of Twitter data cleansing model on the performance of topic modeling. The proposed method has not considered data quality in terms of data streaming. In the future, we could address this problem.

References

1. Huang, Y., Murphey, Y.L., Ge, Y.: Intelligent typo correction for text mining through machine learning. Int. J. Knowl. Eng. Data Min. **3**(2), 115 (2015)
2. Kireyev, K., Palen, L., Anderson, K.M.: Applications of topics models to analysis of disaster-related twitter data. NIPS Work. Appl. Top. Model. Text Beyond, Canada, Whistler **1** (2009)
3. Kim, A.E., Hansen, H.M., Murphy, J., Richards, A.K., Duke, J., Allen, J.A.: Methodological considerations in analyzing twitter data. J. Natl. Cancer Inst. Monogr. **2013**(47), 140–146 (2013)
4. Torunoglu, D., Cakirman, E., Ganiz, M.C., Akyokus, S., Gurbuz, M.Z.: Analysis of preprocessing methods on classification of Turkish texts. Int. Symp. Innovations Intell. Syst. Appl. IEEE **2011**, 112–117 (2011)
5. Denny, M.J., Spirling, A.: Assessing the consequences of text preprocessing decisions. Available SSRN 2849145 (2016)
6. Boyd-Graber, J., Mimno, D., Newman, D.: Care and feeding of topic models: problems, diagnostics, and improvements. In: Airoldi, E.M., Blei, D., Erosheva, E.A., Fienberg, S.E., (eds.) Handbook of Mixed Membership Models and Their Applications, pp. 225–254. Chapman and Hall/CRC (2014)
7. Dey, D., Kumar, S.: Reassessing data quality for information products. Manage. Sci. **56**(12), 2316–2322 (2010). https://doi.org/10.1287/mnsc.1100.1261
8. Han, J., Chen, K., Wang, J.: Web article quality ranking based on web community knowledge. Computing **97**(5), 509–537 (2015)
9. Nurse, J.R., Rahman, S.S., Creese, S., Goldsmith, M., Lamberts, K.: Information quality and trustworthiness: a topical state-of-the-art review. Int. Conf. Comput. Appl. Netw. Secur. (ICCANS 2011) (2011)
10. Chinnov, A., Kerschke, P., Meske, C., Stieglitz, S., Trautmann, H.: An overview of topic discovery in twitter communication through social media analytics. Twenty-first Am. Conf. Inf. Syst, Puerto Rico (2015)
11. Hemalatha, I., Varma, G.P.S., Govardhan, A.: Preprocessing the informal text for efficient sentiment analysis. Int. J. Emerg. Trends Technol. Comput. Sci. **1**(2), 58–61 (2012)
12. Sun, F., Belatreche, A., Coleman, S., McGinnity, T.M., Li, Y.: Pre-processing online financial text for sentiment classification: a natural language processing approach. In: IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr), London, IEEE, pp. 122–129 (2014)

13. Duwairi, R., El-Orfali, M.: A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. J. Inf. Sci. **40**(4), 501–513 (2014). https://doi.org/10.1177/0165551514534143
14. Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., Perea-Ortega, J.M.: OCA: opinion corpus for Arabic. J. Am. Soc. Inf. Sci. Technol. **62**(10), 2045–2054 (2011)
15. Jianqiang, Z.: Pre-processing boosting twitter sentiment analysis? In: IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), IEEE, pp. 748–753 (2015)
16. Indra, S.T., Wikarsa, L., Turang, R.: Using logistic regression method to classify tweets into the selected topics. In: 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, pp. 385–390 (2016)
17. Murshed, B.A.H., Mallappa, S., Ahmed, F.A.M., Al-Ariki, H.D.E.: Semantic analysis on big twitter dataset for automatic topic modeling. Test Eng. Manag. **83**, pp. 14657–14684 (2020)
18. Singh, T., Kumari, M.: Role of text pre-processing in twitter sentiment analysis. Procedia Comput. Sci. **89**, 549–554 (2016)
19. Gupta, I., Joshi, N.: Tweet normalization: a knowledge based approach. In: International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), IEEE, pp. 157–162. Dubai, United Arab Emirates (2017)
20. Al-Khafaji, D.H.K., Habeeb, A.T.: Efficient algorithms for preprocessing and stemming of tweets in a sentiment analysis system. IOSR J. Comput. Eng. **19**(3), 44–50 (2017)
21. Ramachandran, D., Parvathi, R.: Analysis of twitter specific preprocessing technique for tweets. Procedia Comput. Sci. **165**, 245–251 (2019)
22. N. P. K M., K. P, Preprocessing methods for unstructured healthcare text data. Int. J. Innov. Technol. Explor. Eng. **9**(2), 715–719 (2019)
23. Arpaci, I., et al.: Analysis of twitter data using evolutionary clustering during the COVID-19 pandemic. Comput. Mater. Contin. **65**(1), 193–204 (2020)
24. Arpaci, I., Karataş, K., Baloğlu, M.: The development and initial tests for the psychometric properties of the COVID-19 Phobia Scale (C19P-S). Pers. Individ. Dif. **164**, 110108 (2020). https://doi.org/10.1016/j.paid.2020.110108
25. Joakim, C.: Explore python, machine learning, and the NLTK library. IBM Dev. Work. (2012)
26. Porter, M.F.: An algorithm for suffix stripping. Program **40**(3), 211–218 (2006)
27. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. In: Processing, CS224N Project Report, pp. 1–6 (2009)
28. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. pp. 1320–1326 (2010)