MINING AND PREPROCESSING TWITTER DATA FOR DETECTING POTENTIAL EFFECTS OF REMDESIVIR

R. Komalavalli [1] & Dr. R.Vidyabanu[2]
Research Scholar[1], Assistant Professor[2],
Department of Computer Science,
L.R.G. Government Arts College for Women, Tirupur.

Abstract: *This paper discusses a new model towards opinion mining and sentiment analysis of the text reviews posted in twitter website which are mostly in unstructured format. In recent years, web forums and social media has become an excellent platform to share opinions in the form of text about any topic especially in medical terms. These opinions are used for making decisions to choose any manufactured goods. Usually, opining mining deals with analyzing and summarizing opinions about specified items however sentiment analysis classifies prejudiced text into positive/negative. The outbreak of Coronavirus, namely COVID-19, has created a calamitous situation throughout the world. The cumulative incidence of COVID-19 is rapidly increasing day by day. In the absence of any curative drug, the United States gave Emergency use authorization to the antiviral* ***'Remdesivir'*** *for people hospitalized with severe COVID‑19. This research will try to analyze and find the most relevant drugs names mentioned in CORD-19 data corpus related to the treatment of COVID-19. In this research work, a proficient pre-processing technique for opinion mining is implemented and will be utilized for investigating patients or users' comments on 'Twitter' social network about 'remdesivir'. Therefore, various text pre-processing methods have been utilized on the dataset to attain an adequate standard text.*

Keywords: COVID‑19, Remdesivir, Twitter data, Sentiment Analysis, Preprocessing.

1. Introduction

Beside with the vital development of social media, individuals as well as companies are progressively getting public opinions which support their decisions. Opinion mining is scrutinized as a sub-field of Natural Language Processing (NLP), information retrieval, and text mining. It is the route of understanding the users' opinions from their statement that have been signified as unstructured texts. Appearance of 'online social media' has led to the invention of a vast amount of user statements on websites and thus, has raised opinion mining as a valuable also interesting problem. Twitter is a micro blogging service in which people share and discuss their thoughts and views in 140 characters without being constrained by space and time. Millions of tweets are generated each day on different issues. People usually express their sentiments towards various issues [1].

The novel Coronavirus disease (COVID-19) was first reported on 31 December 2019 in the Wuhan, Hubei Province, China. It started spreading rapidly across the world [2]. The cumulative incidence of the COVID-19 is rapidly increasing and has affected 196 countries and territories with USA, Spain, Italy, U.K. and France being the most affected. World Health Organization (WHO) has declared the coronavirus outbreak a pandemic, while the virus continues to spread. As on 4 May 2020, a total of 3,581,884 confirmed positive cases have been reported leading to 248,558 deaths. The major difference between the pandemic caused by COVID-19 and related viruses, like Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), is the ability of COVID-19 to spread rapidly through human contact and leave nearly 20% infected subjects as symptom-less carriers [2]. Moreover,

various studies reported that the disease caused by COVID-19 is more dangerous for people with weak immune system. The below figure demonstrates the total confirmed COVID-19 cases across the world and India till June-14, 2020 [3].

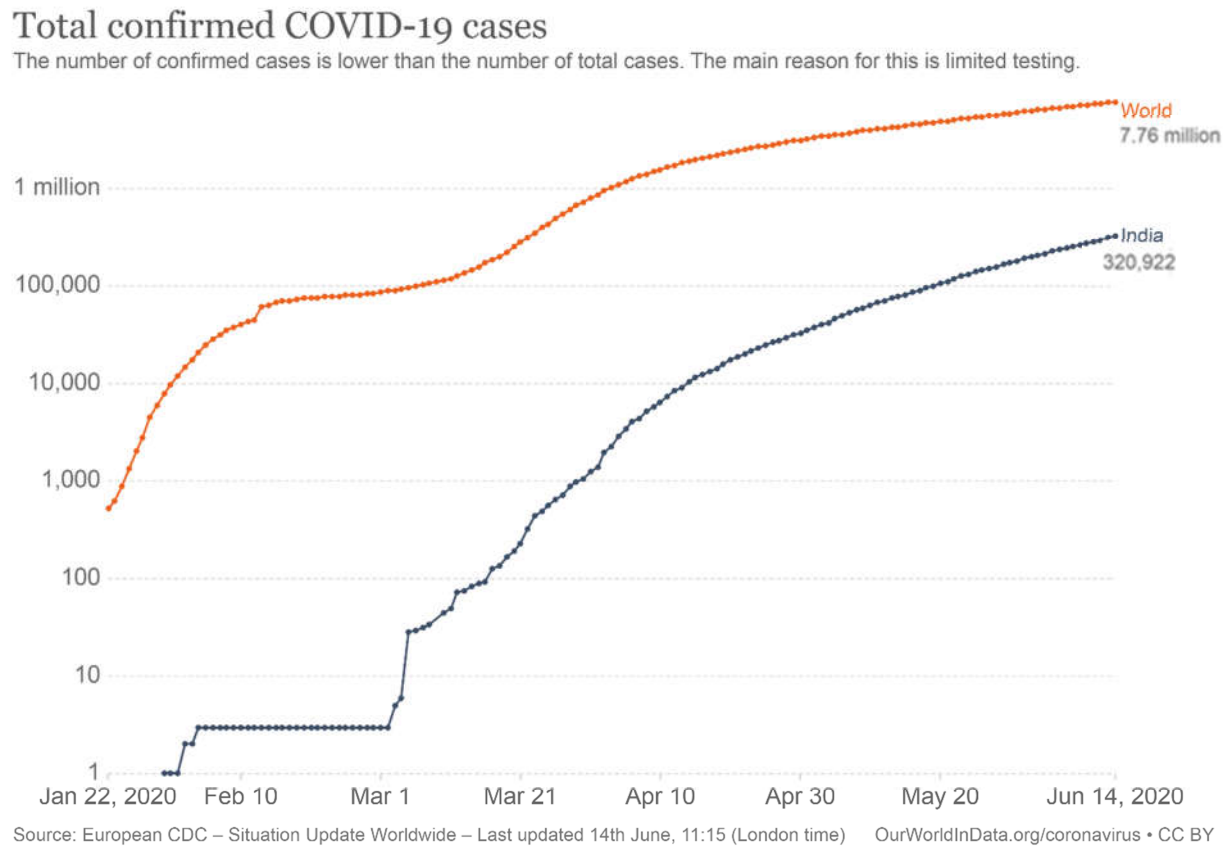

Fig1. Total Confirmed COVID-19 cases Across the World and India

Remdesivir was developed by Gilead and found to be effective against severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) in animal models [. Remdesivir was also evaluated by Gilead for COVID-19 in early 2020. It was then used by Chinese medical researchers in patients for clinical testing in late January 2020, suggesting a favorable inhibitory effect on SARS-CoV-2. Since then, several clinical trials of remdesivir have been initiated by China and the World Health Organization (WHO). Further, randomized clinical trials are ongoing or planned to determine the effect on improvements in patient recovery. Remdesivir is still considered one of the most promising drug candidates for the treatment of COVID-19 [5].

On May 1, 2020, the US Food and Drug Administration (FDA) stated that the potential benefits of remdesivir outweigh its known and potential risks for some patients with severe COVID-19, since a NIH study demonstrated better recovery times than with placebo. The FDA disturbed an alternative use approval for ‘remdesivir’ for the treatment of supposed or laboratory confirmed COVID-19 patients with severe infection. Altogether, undesirable reactions have become one of the top causes of deaths [5]. For surveillance of this events, patients have gradually become involved in reporting their experiences with medications through the use of

dedicated and structured systems. The embryonic of social networking offers a technique for patients to define their drug experiences online in less structured free text format. We developed a computational approach that collects, processes and analyzes Twitter data for drug (remdesivir) effects.

1.1 Motivation and our contributions

In our proposed system, sentiment analysis and machine learning techniques can be utilized to handle large amount of twitter data and intelligently predict the effects of the remdesivir drugs efficiently. Traditional methods of reporting drug effects include clinical trials and spontaneous reporting which has been an effective solution for post-marketing surveillance of approved medications in normal clinical use, detecting many important signals. Recently, mining twitter data has been analyzed in detecting drugs effects.

A tweet contains a lot of opinions about the numerous domain which are expressed in different ways by different users. It is often necessary to normalize the text for any NLP (Natural Language Processing) tasks. Tweets are often represented in informal/unstructured way. Hence systematic pre-processing of tweets is required to enhance the accuracy of sentiment analyzer. This paper implement the tweet extraction and the preprocessing phase. Parallel python framework is used to implement the preprocessing phase to address the subsequent phases efficiently.

1.2 Article structure

The rest of the paper is organized as follows: Section 2 presents the Literature Survey. Section 3 provides discussions on Sources for Opinion Mining. Section 4 provides the details on Opinion Mining (OM)/Sentiment analysis (SA). Section 5 demonstrated the Methodology. Finally, Section 6 concludes the work.

2. Literature Survey

Ostrowski (2013) confers exactly how solicitation of ‘semantic filtering’ on ‘twitter’ data can be utilized to find problems trending presently. Afterward a phase of realistic filtering where the tweets go through very elementary filtering, a knowledge base is established by examining the tweets by machine learning methods. Trend plots were then taken for periods of time which showed periods of spikes and drops. This records was reliable with the facts from Google Trends that is Google’s proprietary trend determination tool. By evaluating the experimental outcomes from Google Trends the authors were capable to attain the accuracy of their technique [6].

Asghar et al. (2014) also have proposed a medical opinion lexicon for mining health reviews available on different health forums. This technique works based on the incremental modal and corpus of health reviews by creating medical polarity lexicon for medical terms. In every growth, the lexis of vocabulary is improved analytically, polarity score with every word is committed. Finally, the resulting lexicon is filtered from pointless words by using word logic disambiguation methods. The proportional results demonstrate the effectiveness of the

established technique with an accuracy of 82% on training corpus also 78% on testing corpus of health reviews [7].

Akhtar (2014) utilized different social network analysis such as Networkx, Gephi, Pajek and IGraph stated comparative results on proficiency, visualization and graph features. Latterly, the authors decided that IGraph beat other tools in treating difficult also huge network [8].

Hogenboom et al. work (2015) that focuses in using rhetorical structure in sentiment analysis, and utilizes structural aspects of text as an aid to distinguish important segments from those less important, as far as contributing to the overall sentiment being communicated. As such, they put forward a hypothesis based on segments' rhetorical roles while accounting for the full hierarchical rhetorical structure in which these roles are defined [9].

The detection of the real-time abuse of the drug using the tweets has been analyzed by Phan et al. (2017). Authors use legal and illegal drugs dataset, original text with the collection of 31,478 tweets. It does not use any preprocessing and uses the J48, Random Forest, Naïve Bayes, and SVM (Support Vector Machine) classifier for training purposes. The developed classifier developed has been tested on the real-world tweet dataset with the precision of 74.8% with the J48 algorithm. The suggested work includes Term Frequency-Inverse Document Frequency (TFIDF) used to reflect the relevance of the term in the given document and to improve the accuracy and to use Mechanical Turk for the collection of vast amounts of data [10].

Bhat et al. (2018) used sentiments for the development of the system that observes the opinion by people on some product or people. It uses the Twitter API to extract 1000 latest tweets and performs text processing like stemming and stop word removal. No machine learning algorithm was used for the classifying purpose. This gave us a model that calculates the sentiment by multiplies of adverbs value instead of summing up the whole sentiment of tweets. Future work such as the development of the algorithm for identifying the offensive statements and, improving the efficiency of mapped words are suggested [11].

Table1. Exploit of existing Sentiment analysis

Author & Year	**Preprocessing method**	**Dataset**	**Sentiment analysis method used**	**Accuracy (%)**
Haddi, E et al (2013) [12]	Expanding Acronym, Removing Numbers & URL's, Removing Negation terms with both prior polarity and N-grams	STS-test STS-gold SS-Twitter SE-Twitter SemEval2014	LR, SVM, NB, RF	92.5, 91.3, 90.7, 93.2
Bao, Y .et al (2014) [13]	Denoising, slang and URL removal, feature selection	STS-test STS-gold	LibLinear	85.5
Dos Santos, F.L et al (2014) [14]	Terms standardization, stemming, lemmatization, spell check	Google Play reviews[a] and Movie reviews	NB, SVM	82.0331 79.6

		(IMDB)		
Dwi Aji Kurniawan et al (2016)	RT removal, Case converting, Website address removal, Twitter username removal and Changing abbreviations to their actual phrases.	Twitter API (Real-time traffic data)	NB SVM DT	98.02 98.31 98.41
Jianqiang, Z et al (2017) [15]	Basic cleaning, emoticon, negation, PyEnchant, stemming, stopwords	SemEval 2015 & 2016	Naïve Bayes Multinomial	80.84
María del Pilar Salas-Zárate et al (2017) [16]	Normalization, Stop words removal, Sentence splitter POS tagging, Lemmatization, Tokenization	Twitter API (Diabetes)	Aspect-level sentiment classification	81.93

3. Sources for Opinion Mining

Having already discussed the characteristic of Opinion Mining, the natural next question is related to the possible applications of this technique and subsequently, the set of possible web data sources to use as input. As the Web continues to grow, the number of possible sources for Opinion Mining grows rapidly too. Within social media, it is possible to find a variety of different platforms whose content is being increasingly used by individuals and organizations for their decision making. Social media includes web pages such as reviews, forum discussions, blogs and social networks, like Facebook, YouTube, Instagram, Tik Tok, Twitter and so on. This variety of sources suggests a heterogeneous mix of structures to work with. As a consequence of this, one needs to specify a different strategy for each source, each one oriented to the particular problem of extracting data, then processing this data and finally discovering valuable information locally within the selected source [19]. The below figure demonstrates the total number of social media users around the world,

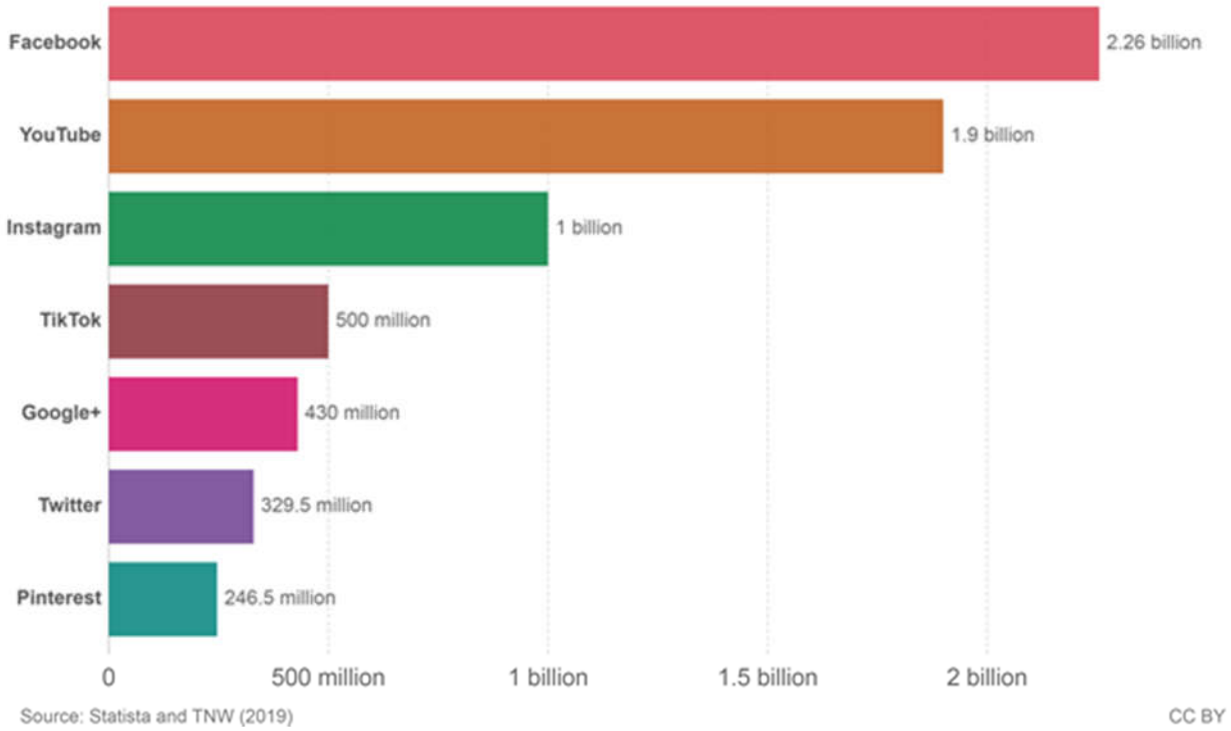

Fig2. Total Number of people using social media platforms (2004 – 2019)

3.1 Twitter

Twitter is often considered a microblogging platform, but it is also referred as a social network. Considering the features of this platform, probably it has been perform like a social network but for the point of Opinion Mining, Twitter is a highly relevant to microblog system. Microblogging is a growing popular communication channel on the Internet, where users can write short text entrances in a public or private way. Messages are extremely short, allowing users to write a maximum of 140 characters on each post, called a tweet. These tweets can be written through the Twitter web interface or through a variety of mobile devices, like smartphones, some cell phones and other devices. These short messages can be seen as being a newspaper headline and subtitle, which makes them easy to produce (write) and process (read). This feature makes microblogs unique when compared to other similar web communication media, like blogs and web pages.

As a microblogging social network, the first relevant feature of Twitter's messages are their brevity, which make users look for various special ways to add content in the messages. The most-used approaches are adding content indirectly or trying to use fewer characters to express the same ideas. Twitter framework refers to the use of Twitter's special characters, like hashtag (#) to denote a particular topic or to call a user, and RT (retweet). On the other hand, a different problem associated with tweets is that they are written in colloquial or informal language. In addition to the brevity of messages, this supposes the use of many colloquial symbols or expressions that, in order to be understood as regular text, require preprocessing.

Finally, it is interesting to see Twitter as a network of related users, who share opinions among themselves and influence each other. Twitter provides some useful tools that can be used

to analyze how a specific topic or opinion tendency is spread through the network, and discover users who influence others or are more easily influenced by another [20].

4. Opinion Mining (OM)/Sentiment analysis (SA)

Opinion Mining (OM) is a type of Natural Language Processing (NLP) for tracking the opinion of the public about a particular product, service or topic. It is also called as Sentiment Analysis (SA) which involves building a system to collect and examine the emotions, opinions about the product, service and topic made in blog posts, comments, reviews or tweets.

Opinion mining software facilitates automatic extraction of opinions, emotions and sentiments in text and also tracks attitudes and feelings. People express their views by writing blog posts, comments, reviews and tweets about different topics. Tracking products and brands on the web and then determining whether they are viewed positively or negatively can be done.

Retrieving opinions (opinion mining) from text have recently drawn much attention because of their many useful applications such as extracting customer sentiments, automated recommender systems or deducing public opinions about a certain product, topic or service for companies and organizations. There are two main goals of sentiment mining:

1. To check whether a given text contains an opinion as opposed to being factual or objective.

2. To extract the opinion of a given text by classifying it as positive, negative, or neutral with respect to the given target.

Researchers incline to ignore the "neutral" class under the hypothesis in which, neutral is less to absorb sentiment on neutral texts difference to positive or negative groups. At the fact of sentiment analysis, neutral typically indicates 'no opinion'. It is the computational learning of people's minds, opinions, sentiments, judgments, outlooks, and sensations in the way of objects then their aspects stated in text.

Textual data (i.e) data can be categorize into two types. First one is exact then second one is opinionated information. Elements are 'objective' and 'subjective' sentences enclose explicit opinions, experience, and views about specific artifact. E.g. consider two sentences 'S_1' and 'S_2'.

- *Objective Sentence (S_1)*："I took the dolopar last week for fever.

- *Subjective Sentence (S_2)*："It is a best drug for fever"

Where 'S_1'is an objective sentence that indicates the fact about the 'dolopar' tablet, whereas 'S_2'is a subjective sentence that indicates the view about that medicine. Subjective statements, classify as positive or negative division, for example sentence 'S_2' and 'S_3'hold positive and negative polarity correspondingly.

- *S_2(Positive):* "It is such a good medicine for fever"

- *S_3(negative)：*"I recovered from fever, but headache was not cured"

Sentiment analysis mostly focuses on analyzing polarity value of subjective sentences. However with some regards subjective sentences don't have any explicit opinions and objective sentence have. For example, S_1 is a subjective sentence but don't have any polarity. Whereas S_2 and S_3 refer to objective sentence that contain implicit positive and negative polarity. This situation lead to extract prime candidates having a semantic orientation that would be leaning more towards objective than to subjective and classify objective sentences according to their sentiment polarity.
Therefore, the target of sentiment analysis is to find opinions, identify the sentiments they express, and then classify their polarity using machine learning approach [20].

5. Methodology

In order to extract the opinion, first all data is selected and extracted from twitter in the form of tweets. After collecting the data set, these tweets were cleaned from emoticons, unnecessary punctuation marks etc., and then database is created to store this twitter data in a specific transformed structure. In this structure, all the transformed tweets are in lowercase alphabets and are divided into different parts of tweets in the specific field. The details about the steps adopted for the transformation raw data into cleaned dataset is defined in below figure.

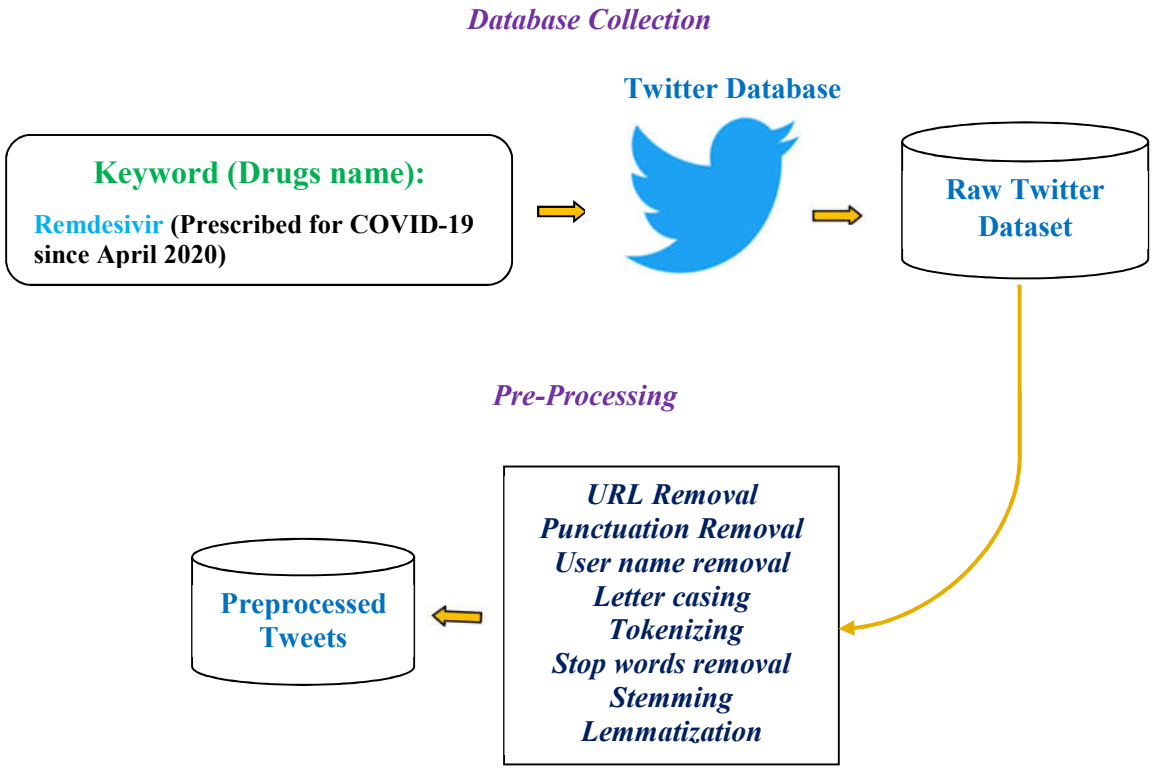

Fig3. Architecture of the proposed system

5.1 Selection of Medications

To query Twitter data, we needed to have particular drug names. In this research, we chose ***'Remdesivir'*** drug to test if any related effects could be found in the Twitter data, and they were chosen based upon the following. First, they should have been on market for a number of years so that sufficient tweets would exist for effects to be reported, and second, they should not be taken by patients whose conditions kept them from posting tweets regularly [5]. The below details shows the drugs we chose to be investigated in this research.

Drug Name: Remdesivir

Company Name: Gilead Sciences

Prescribed for: Hepatitis C Virus (HCV), RNA viruses (SARS and MERS) and Ebola virus

Remdesivir was originally developed to treat HCV and was then tested against RNA, Ebola virus disease and Marburg virus disease. Side effects may comprise liver inflammation then an distillation associated reaction with nausea, low blood pressure, and sweating. As of April 2020, remdesivir was viewed as the most promising treatment for COVID-19.

5.2 Implementation with python Libraries

Python is a high level, object-oriented, and interpreter-based multi–purpose programming language that was designed by Guido Van Rossum. Python contains a few keywords; however, it provides huge library resources for its programmers. This research utilizes the python Libraries (Natural Language Toolkit (NLTK), TextBlob, CoreNLP, Gensim, spaCy, polyglot, scikit–learn sys, numpy, pandas, matplotlib, sklearn, csv, string, os and so on) for implementing the proposed model.

Natural language processing projects that required superior knowledge of mathematics, machine learning, and linguistics. SO, Python model provide ready-made tools that simplify text preprocessing and effectively building machine learning models with given data. There are many tools and libraries created to solve NLP problems [21].

5.3 Collecting Drug-Related Tweets

Twitter provides streaming API (Application Programming Interface) to allow the users to retrieve real time data. It is a tool that makes the interaction with computer programs and web services easy. Many tools such as python, JavaScript, and R-tool services are developed to interact with twitter services and to access data in programmatic way. Here, R-Tool utilized to searching for tweets posted recently by users to extract real time tweets. Before retrieving tweets, the drug names need to be determined as the keywords in tracking and retrieved data is stored into the .csv file.

Fig4. Dataset Retrieval process

To access Twitter Streaming API, Getting Twitter API keys (API key, API secret, Access token and Access token secret) is important step to connect to Twitter Streaming API. Library files called 'Tweepy' is utilized to Twitter Streaming API and downloading the data. Primarily, drugs should have been on the market to treat substantial syndrome, so that sufficient tweets would exist for reporting their effects. The data set was constructed using a portion of drug keyword tweets and the tweets gathered across the world. The total amount of tweets in this data set is about 1500 tweets. For fetching the twitter data from the twitter API includes the following steps such as Installation of the needed software, authentication of twitters data and import the required libraries [20]. The collected Tweets were normalized to expand condensed words and phrases, abbreviations and acronyms to the normal format. SO, after dataset collection, preprocessing is significant procedure for further data analysis.

5.4 Preprocess

Applying text preprocessing steps before analyzing the tweets is very important for achieving good results. The purpose of the preprocessing is to illuminate the input data for further analysis. The proposed method exploits numerous methods of natural language preprocessing, namely URL Removal, Punctuation Removal, User name removal, Letter casing, Tokenizing, Stop word removal, Normalization, Stemming and Lemmatization [19], to make a standard dataset. The preprocessing methods are illustrated below,

Twitter Data Extraction: This phase involves creating a Twitter API and downloading the tweets as per the requirements, i.e., downloading tweets with a specified keyword (Remdesivir). Twitter API supports to extracting the tweets and the data can be retrieved in .csv format. Examples of raw tweets are shown in Table 2.

Table2. Examples of raw Tweets with patient's opinion about Drug effects

'Remdesivir' To Be Available For Coronavirus Patients This Week, Gilead CEO Says $GILD #remdisivir
https://t.co/kKlgN7w1zd

RT @abledoc: @jimcramer Gilead should go back to the drawing board & try other drugs. #remdisivir An antiviral drug that cannot even reduce…

RT @graziella2149: Huh, how about that. Fauci promoting #remdisivir which has limited studies & success vs hydroxychorloquine which has mor…

RT @aaravmeanspeace: How true is this that China holds the patent for #remdisivir drug, which is being claimed to be a cure for #Covid19, m…

RT @RonaldBruceBar3: #remdisivir
REMDISIVIR, WORKS BEST WITH *OTHER DRUGS* THAT PREVENT VIRUS RECPTORS FROM CONNECTING WITH HUMAN CELLS!…

@CBSNews Um hello, it's extremely irresponsible to call #remdisivir "coronavirus drug". While it's shown some promi… https://t.co/bYs9HTMwMj

RT @naashonomics: $GILD #remdisivir available to all #hospitals #FDA $SPY $QQQ https://t.co/yinUcw0iFH

10s of thousands of #remdisivir treatments going out this week
Media @WHO will tell us how this is somehow bad n… https://t.co/ccSjG2KeGZ

#remdisivir #Antiviral #Covid_19
I'm glad somebody said what needed to be said. The hype about Remdisivir is sort… https://t.co/woRCbIVEOD

RT @abledoc: @jimcramer Gilead should go back to the drawing board & try other drugs. #remdisivir An antiviral drug that cannot even reduce…

@LaurelOld How is the Ebola's RNA-dependent RNA polymerase different from SARS's RNA-dependent RNA polymerase?
Two… https://t.co/tlAdbqMyBu

URLs Removal: URLs have got nothing to do with emotion analysis. They sometimes mislead the emotional interpretation of the tweet. So, URLs should be removed from the tweets for effective analysis.

Input:
I have a feeling we're setting ourselves up for a classic rope-a-dope with #covid19
All governments selling hope b… https://t.co/9gbGkPmGzd

Output:
I have a feeling we're setting ourselves up for a classic rope-a-dope with #covid19 All governments selling hope b…

Converting to Lower Case: Text in the tweets will be in the combination of both upper and lower characters [6]. So, the twitter data is converted into the lower case so that it would become easy to analyze.

Input:

I have a feeling we're setting ourselves up for a classic rope-a-dope with #covid19 All governments selling hope b…

Output:
i have a feeling we're setting ourselves up for a classic rope-a-dope with #covid19 all governments selling hope b…

User name Removal: In Twitter texts, almost many sentence contains a user names. Their presence does not contain any sentiment and it is important step to remove them in pre-processing step.

Input:
@jimcramer Why spent money researching new drugs when you can make more spending on Lobbying instead. You would exp… https://t.co/LLVWCJ1poW

Output:
why spent money researching new drugs when you can make more spending on lobbying instead you would exp

Removing the Punctuations (#, @, etc,): Punctuations are just used to highlight a particular word(s) in the whole tweet, and it does not share any contribution toward analyzing the opinion of a person. Hence, they should be removed to make analysis process easy.

Input:
RT @abledoc: @jimcramer Gilead should go back to the drawing board & try other drugs. #remdisivir An antiviral drug that cannot even reduce…

Output:
gilead should go back to the drawing board amp try other drugs remdisivir an antiviral drug that cannot even reduce

Remove Blank spaces: This step is used to remove the unwanted blank space which helps for the tokenization of the tweets. Tokenization means breaking the sentence into words.

Input:
RT @abledoc: @jimcramer Gilead should go back to the drawing board & try other drugs. #remdisivir An antiviral drug that cannot even reduce…

Output:
'gilead should go back to the drawing board amp try other drugs remdisivir an antiviral drug that cannot even reduce'

Removing the RT: If a tweet is compelling and interesting enough, users might republish that tweet, commonly known as retweeting, and twitter employs “RT” to represent re-tweeting. This can be removed to make subsequent step successfully.

Input:
RT @abledoc: @jimcramer Gilead should go back to the drawing board & try other drugs. #remdisivir An antiviral drug that cannot even reduce…

Output:

'gilead should go back to the drawing board amp try other drugs remdisivir an antiviral drug that cannot even reduce'

Tokenization: Opinion mining tools usually need to analyze their input comments, lexically. In the process of lexical analysis, a sentence is converted into a sequence of tokens, which are the meaningful parts of the text of a comment. This process is called tokenization and leads to generation of a collection of independent meaningful parts, called tokens

Input:
remdisivirremdisivir works best with other drugs that prevent virus recptors from connecting with human cell

Output:
['remdisivirremdisivir', 'works', 'best', 'with', 'other', 'drugs', 'that', 'prevent', 'virus', 'recptors', 'from', 'connecting', 'with', 'human', 'cell']

Removing stop words: Despite the frequent use of stop words, they are pragmatically insignificant. Some examples of stop words in English are: “or”, “and” “this” and so on. Although it is thought that only the linking words are stop words, many verbs, auxiliary verbs, nouns, adverbs, and adjectives can be stop words, too. In most text mining operations, the processing result is significantly improved by eliminating such words. Therefore, removing the stop words reduces the computational load and increases the speed. In this phase, the proposed method loads a complete list of English stop words that is gathered in Python programming language.

Input:
['remdisivirremdisivir', 'works', 'best', 'with', 'other', 'drugs', 'that', 'prevent', 'virus', 'recptors', 'from', 'connecting', 'with', 'human', 'cell']

Output:
['remdisivirremdisivir', 'works', 'best', 'drugs', 'prevent', 'virus', 'recptors', 'connecting', 'human', 'cells']

Lemmatization: It refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

Input:
['remdisivirremdisivir', 'works', 'best', 'with', 'other', 'drugs', 'that', 'prevent', 'virus', 'recptors', 'from', 'connecting', 'with', 'human', 'cell']

Output:
['remdisivirremdisivir', 'work', 'best', 'drug', 'prevent', 'virus', 'recptor', 'connecting', 'human', 'cell']

Stemming: It refers to a basic experimental process which chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.

Input:
['remdisivirremdisivir', 'work', 'best', 'drug', 'prevent', 'virus', 'recptor', 'connecting', 'human', 'cell’]

Output:
['remdisivirremdisivir', 'work', 'best', 'drug', 'prevent', 'virus', 'recptor', 'connect', 'human', 'cell']

The outcome of data after preprocessing, i.e., after performing all the above steps is as shown in below Table.

Table3. Pre-processed Tweets

['remdesivir', 'available', 'coronavirus', 'patient', 'week', 'gilead', 'ceo', 'says, 'gild', 'remdisivir']
['gilead', 'go', 'back', 'draw', 'board', 'amp', 'try', 'drug', 'remdisivir', 'antiviral', 'drug', 'cannot', 'even', 'reduce']
['huh', 'fauci', 'promot', 'remdisivir', 'limit', 'study', 'amp', 'success', 'v', 'hydroxychorloquine', 'mor']
['true', 'china', 'hold', 'patent', 'remdisivir', 'drug', 'claim', 'cure', 'covid']
['remdisivirremdisivir', 'work', 'best', 'drug', 'prevent', 'virus', 'recptor', 'connect', 'human', 'cell']
['um', 'hello', 'extreme', 'irresponse', 'call', 'remdisivir', 'coronavirus', 'drug', 'shown', 'promise']
['gild', 'remdisivir', 'available', 'hospital', 'fda', 'spy', 'qqq']
['thousand', 'remdisivir', 'treatment', 'go', 'week', 'medium', 'tell', 'u', 'somehow', 'bad', 'n']
['remdisivir', 'antiviral', 'covid', 'im', 'glad', 'somebody', 'said', 'need', 'said', 'hype', 'remdisivir']
['gilead', 'go', 'back', 'draw', 'board', 'amp', 'try', 'drug', 'remdisivir', 'antiviral', 'drug', 'cannot', 'even', 'reduce']
['ebola', 'rnadepend', 'rna', 'polymeras', 'differ', 'sar', 'rnadepend', 'rna', 'polymerasetwo']

The next step is to create visual plots to visualize the opinions of the users. The words within a sentence can be associated with side effects of the medicine. The example of wordcloud visualization is depicted in below figure. It gives the visualization of the most used words in the tweets.

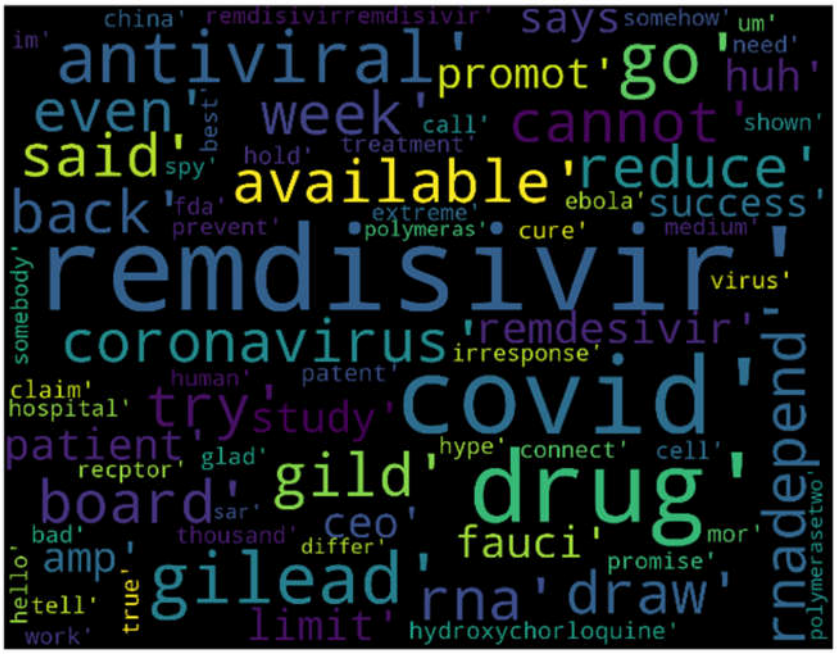

Fig5. Visualization of preprocessed data

The above figure shows visualization for data of various tweets from preprocessed data. This figure is giving us insight about a lot of words that related to specified keyword. In fact, the data originated in twitter about remdesivir are new sources for understanding the effects of users. A very important issue in this kind of COVID-19 situation, this data is normally utilized to make important decision for some professional, usually what the user is thinking and facing the effects of this drugs, However, these opinion can be necessary for analyzing the remdesivir behavior.

6. Conclusion

As the COVID-19 pandemic races across the globe, the scientific community, from academic and government laboratories to small biotechnology companies and multinational pharmaceutical corporations, has mobilized to develop and evaluate potential therapeutics and vaccines. While remdesivir represents one compound whose recent use authorization may, in part, mitigate the morbidity, mortality, and strain on global healthcare systems caused by COVID-19, additional ongoing clinical trials will provide much-needed clarity surrounding the repurposing of approved drugs and experimental agents against COVID-19. Tweets are an effective way for people to express their opinion and this fact can be exploited to analyze the different sentiments of public in various spheres of society such as the medical field. This research work developed a feasible yet effective method for extracting potential remdesivir drug effects from the Twitter data and preprocess the twitter data through the use of natural language processing tool.

References

1. J. Zhao, and G Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis." IEEE Access, Vol 5, pp 2870-2879, 2017.
2. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses
3. https://ourworldindata.org/grapher/total-cases-covid-19?yScale=log&time=2020-06-14&country=IND~OWID_WRL
4. Vassilara, F.; Spyridaki, A.; Pothitos, G.; Deliveliotou, A.; Papadopoulos, A. A Rare Case of Human Coronavirus 229E Associated with Acute Respiratory Distress Syndrome in a Healthy Adult. Case. Rep.Infect. Dis. 2018, 2018, 6796839.
5. Richard T. Eastman, Jacob S. Roth, Kyle R. Brimacombe, Anton Simeonov, Min Shen, Samarjit Patnaik, and Matthew D. Hallcorresponding author "Remdesivir: A Review of Its Discovery and Development Leading to Emergency Use Authorization for Treatment of COVID-19" ACS Cent Sci. 2020 May 27; 6(5): 672–683.
6. Ostrowski, DA. (2013). Semantic Filtering in Social Media for Trend Modeling. In 2013 IEEE Seventh International Conference on Semantic Computing (pp 399–404).
7. Muhammad Z. Asghar, A. Khan, Fazal M. Kundi, M. Qasim, F. Khan, R. Ullah, Irfan U. Nawaz. Medical opinion lexicon: an incremental model for mining health reviews. International Journal of Academic Research Part A; 2014; 6(1), 295-302.
8. Akhtar, N. (2014). Social Network Analysis Tools. In Fourth International Conference on Communication Systems and Network Technologies (pp 382–388)
9. A. Hogenboom , F. Frasincar , F. de Jong , U. Kaymak , Using rhetorical structure in sentiment analysis, Commun. ACM 58 (7) (2015) 69–77.
10. N. Phan, S. Chun, M. Bhole and J. Geller, "Enabling Real-Time Drug Abuse Detection in Tweets", IEEE 33rd International Conference on Data Engineering (ICDE), 2017.
11. S. Bhat, S. Garg and G. Poornalatha, "Assigning Sentiment Score for Twitter Tweets", 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.

12. Haddi, E., Liu, X., Shi, Y.: The role of text pre-processing in sentiment analysis. Procedia Comput. Sci. 17, 26–32 (2013)
13. Bao, Y., Quan, C., Wang, L., Ren, F.: The role of pre-processing in twitter sentiment analysis. Lecture Notes in Computuer Science (including Subser. Lecture Notes in Artificial Intelligence, Lecture Notes in Bioinformatics) 8589 LNAI, pp. 615–624 (2014)
14. Dos Santos, F.L., Ladeira, M.: The role of text pre-processing in opinion mining on a social media language dataset. In: Proceedings—2014 Brazilian Conference on Intelligent System, BRACIS 2014, pp. 50–54 (2014)
15. Jianqiang, Z., Xiaolin, G.: Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access. 5, 2870–2879 (2017)
16. María del Pilar Salas-Zárate, José Medina-Moreira, Katty Lagos-Ortiz, Harry Luna-Aveiga, Miguel Ángel Rodríguez-García, Rafael Valencia-García "Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach" Hindawi Computational and Mathematical Methods in Medicine Volume 2017.
17. Liu, B.: Web data mining: exploring hyperlinks, contents, and usage data. Springer (2011)
18. Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., Manicardi, S.: A comparison between preprocessing techniques for sentiment analysis in Twitter. In: CEUR Workshop Proceeding 1748 (2016)
19. Neetu Anand, Dhruvi Goyal and Tapas Kumar "Analyzing and Preprocessing the Twitter Data for Opinion Mining" Springer Nature Singapore Pte Ltd. 2018 B. Tiwari et al. (eds.), Proceedings of International Conference on Recent Advancement on Computer and Communication, Lecture Notes in Networks and Systems 34,
20. Bifet, A., Frank, E.: Sentiment Knowledge Discovery in Twitter Streaming Data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 1–15. Springer, Heidelberg.
21. https://sunscrapers.com/blog/8-best-python-natural-language-processing-nlp-libraries/#What_is_an_NLP_library