[[1]](#footnote-1)

**SENTIMENT ANALYSIS OF NOVEL CORONAVIRUS BASED ON DATA FROM SOCIA MEDIA**

**Arunachalam DhakshinaMurthy**

*Abstract*— Sentiment Analysis plays a vital role in the internet era due to an extensive range of business applications and social media. The inspiration behind the sentimental analysis is that it provides people’s opinion about a product, an event, or any crisis happening all over the world. In this paper, sentiment analysis is applied to observe people’s opinions about the novel coronavirus using data from social media. Twitter is chosen as the social media platform to extract tweets. Topic modeling and machine learning techniques are performed on the dataset. Topic modeling is performed to study the semantic structure of the data whereas supervised machine learning techniques are performed to study the dataset. The sentimental scores are obtained and performance is studied for each N-gram approach. The best performing model is chosen based on the performance metrics of the model.

*Keywords*— Coronavirus, topic modeling, LDA, sentiment analysis, machine learning approach

# **INTRODUCTION**

C

coronavirus is one of the incurable viruses that has been spreading to people across continents. It has become the talk of the whole world and nobody can deny the fact.

On the other hand, increasing social media interaction among people has opened out a lot of questions regarding regional issues, national issues, and sometimes even international issues are discussed on social media. It has made people connect from any part of the world and discuss the current topics publicly.

Twitter has become one of the most important discussion platforms lately. People tweet about the current happenings around the world and it has become a huge forum to express their opinions and sentiments regarding the issue. Twitter’s API is so flexible that it allows the user to retrieve tweets from the user timeline, searching tweets, and also filtering real-time tweets as they are being uploaded on the platform.

Sentiment Analysis is an automated process that uses artificial intelligence to identify positive, negative, and neutral opinions from the text. It can be used to get widespread responses from reviews, survey questions on social media, and also can be used to make data-driven decisions. Polarity classification can be performed on the dataset to get to know about the opinions of people through any social media platform.

The paper focuses to perform sentiment analysis of how people feel about coronavirus using tweets posted by them on Twitter. The sentiments may vary from person to person because different people may exhibit different opinions about the virus. The main aim of the paper is to mine the opinions of the people on twitter and perform sentiment analysis of the dataset.

The paper is based on topic modeling and sentiment analysis of coronavirus using twitter data. It will give larger insights into how the world thinks about the virus and will be able to arrive at any conclusions depending on the data modeling and visualizations.

In this paper, the dataset from Twitter is extracted using Twitter API and basic data cleaning is performed. The cleaned data is used for feature engineering where all the feature extraction techniques are performed. This data is further used to visualize tweets based on the polarity of the sentiment. It can also be trained to perform modeling which can predict tweets based on the appropriate machine learning model.

Basic machine learning algorithms like Naïve Bayes, Logistic Regression, ensemble algorithms like Random Forest and neural networks like Multilayer Perceptron can be implemented on these feature extracted datasets.

The sentiment analysis is normally performed to analyze and understand people’s sentiment towards a particular topic. This paper can be used to understand the views of people on coronavirus and how they feel about it. It can also be modeled using machine learning techniques to know, how people would feel about it in the future. It will be great learning indeed at the end of the project to see the results of the virus which has been the talk of the year.

# **literature review**

There are many studies in sentiment analysis but almost most of those were based on modeling tweets using Naïve Bayes and Support Vector Machine algorithms. The number of characters per tweet cannot exceed 140 characters. Sentiment analysis of various algorithms has been widely studied. People’s views on any socio-economic factor can be implemented and discovered using sentiment analysis techniques. Sentiment Analysis technique can be used in various domains like healthcare, product reviews, stock market, predicting outcomes of elections, people's views on natural calamities, etc.

**A) Topic-based content and sentiment analysis of Ebola virus on twitter and in the news**

To improve the accuracy, Erin Hea-Jin Kim, Yoo Kyung Jeong, Yuyoung Kim, Keun Young Kang, and Min Song, developed a paper [1] on topic coverage and sentiment analysis of Ebola virus in two different media, Twitter and news publication. An N-gram LDA topic modeling technique is employed to extract entities and the sentimental scores based on the topics are studied. Entity extraction and Entity network modeling are two steps to extract domain-specific features from the data. One of the main drawbacks of the paper is that the datasets were collected for a relatively shorter period of time- only 3 months of the Ebola outbreak.

**B) Analyzing Twitter and web queries for flu trend prediction**

A research paper [2] by Jose Carlos Santos and Sergio Matos, based on the prediction of flu incidence using data from search engines or blogs with social media and various machine learning models are compared and studied. One of the most important novelties of the work is the combination of tweets through multiple linear regression models. But, the limited amount of available data on languages other than English and sometimes language specificities might influence the final results.

**C) Twitter as a Potential Data Source for Cardiovascular disease research**

This research paper [3] by Lauren E Sinnenberg, Christie L DiSilvestro, Christina Mancheno, and Karl Dailey, describes the volume and content of Tweets associated with cardiovascular disease as well as the characteristics of Twitter users and a statistical analysis test is performed among the population of Twitter users. The findings of the paper were related to volume, content, and sender of these tweets which would essentially give a broader perspective. The tweets were geographically restricted. The use of keywords to extract tweets were narrow and lack of insight.

**D) Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer’s Disease Stigma on Twitter**

The paper [4] developed by Oscar N, Fox PA, Croucher R, Wernick R, Keune J, Hooker K, aims at the methodologies borrowed from different areas including computer science, econometrics, statistics, data mining, and sociology which are used to analyze Facebook data to investigate the patients’ perspectives on a given medical prescription. We used social networks to analyze the perception of therapies by Crohn’s disease patients. The major drawback of the paper is that the disease is typically diagnosed in young patients whose characteristics are mostly unknown.

**E) Sentiment Analysis on Twitter to improve Time Series Contextual Anomaly detection for detecting Stock Market Manipulation**

The performance of the model studied by Koosha Golmohammadi and Osmar R Zaiane is based on the Stock market manipulation of anomaly detection [5] by improving the performance of Contextual Anomaly detection by capturing the expected behavior of stocks through sentiment analysis of tweets about stocks. More number of features are selected to improve the accuracy and precision score. The performance of the model starts decaying after the number of features reaches 10000.

**F) Tracking climate change opinions from Twitter data**

The paper [5] developed by Xiaoran An, Auroop R Ganguly, Yi Fang, Steve B Scyphers, Ann M Hunter, and Jenifer G Dy, researches on mining social media data to analyze how people’s sentiment changes according to the climatic changes. Naïve Bayes and SVM classifiers are developed to study the positive and negative sentiments of people. The paper focuses on discovering opinion mining and time series techniques for calculating Z score normalization. A major climatic change might affect the polarity of the sentiment which leads to uncertainty in the overall sentiment.

**G) Modelling public mood and emotion: Twitter Sentiment and Socio-Economic Phenomena**

The research paper [7] developed by Johan Bollen, Huina Mao, and Alberto Pepe is based on sentiment analysis of messages published on Twitter during the second half of 2008 and each tweet is classified on basis of POMS(Profile of Mood States) and then compared with the events that took place during that time. The paper address the moods of people based on a social, political, cultural, and economic sphere which helps study various dimensions of a public mood. But when dealing with machine learning modeling of large data, the minute text of some microblogs is still a challenge. POMS is a traditional approach.

**H) Sentiment Analysis of Tweets on Diabetes: An aspect level approach**

The paper [8] by María del Pilar Salas-Zárate, JoséMedina-Moreira, Katty Lagos-Ortiz,Harry Luna-Aveiga, Miguel Ángel Rodríguez-García, and Rafael Valencia-García focus on aspect level sentiment analysis based on ontologies in the diabetic domain using N-gram approach and the performance metrics are calculated. The aspect level approach is completely new and appears to perform better than the document level approach. General sentiment lexicon is not useful for capturing meaning in health text. A domain-specific approach is required.

**I) Twitter Analysis of Movie Reviews using Machine Learning technique**

The paper [9] developed by Akshay Amolik, Niketan Jivane, Mahavir Bandhari, and Dr. Venketesan, talks about the analysis of movie reviews using various machine learning algorithms like Naïve Bayes and SVM and classify them based on polarity using Twitter data. The paper gives out a basic sentiment analysis approach on how to classify tweets using the Feature Vector. The feature vector technique used is the most sought technique to extract words from a sentence. Though it uses a feature vector, it fails to perform the right amount of pre-processing of data. It doesn’t address the handling of sentences with negation. The accuracy is not as expected for even the best-performing classifier.

**J) Sentiment analysis of Polarity in Product reviews in Social media**

The model [10] developed by Marium Nafees, Hafsa Dar, Ikram Ullah Lali, and Salman Tiwana, is based on sentiment analysis of reviews of five different products using Twitter data. The polarity of the product is analyzed and modeled using three different machine learning models and the accuracy of their models are compared and studied. Sarcasm detection is also addressed as the tweets are classified into two categories emoticon-based and text-based. The text pre-processing of the twitter data is not well explained and Weka pre-processing is developed on the whole for general text pre-processing.

**K) Twitter Sentiment Analysis of Online Transportation service providers**

Sonia Anastasia and Indra Budi developed a paper [11] to measure the satisfaction of using the two most popular online transportation service providers from Indonesia by performing sentiment analysis of Twitter data. This paper uses NetSentiment score as a tool to classify tweets based on three different polarities. But, the k-value in k fold classifier has failed to produce any pattern for classification and the use of context-related keywords is missing.

**L) Emotional detection on Twitter data on the knowledge-based approach**

The paper [12] by Srinivasu Badugu and Matla Suhasini, focuses on the Rule-based approach which detects the emotion or mood of the tweet and classifies the twitter message under the appropriate emotional category. The rule-based approach is used to classify the tweets under specified class categories with the help of the knowledge base. The paper is completely based on high-level emotional detection, as the author had classified emotions into 4 classes and each class will be able to classify the extreme emotions of the user. The paper considers only the text part for emotional classification. It does not consider the emotional icons posted by the user.

**M) Twitter Text Mining for Sentiment Analysis on people’s feedback on Oman tourism**

To improve the accuracy of the model, Valikannu Ramanathan and T Meyappan developed a paper [13] on common sense knowledge created with Omar Tourism ontology based on ConceptNet. All the entities are identified from the tweets using the POS tagger and all the entities are compared with concepts in Domain-specific ontology.One of the main strengths of the paper is the usage of Conceptual sentimental analysis which is based on ontologies and sentiment networks. But, a combination of conceptual and contextual sentimental analysis could have performed better. The author did not concentrate much on the number of attributes retrieved from the twitter dataset.

**N) Sentiment analysis of Twitter data for predicting Stock Market movement**

The research paper [14] by Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, and Babita Maji, is based on sentiment analysis and supervised machine learning principle of twitter data and analysis of the correlation between stock market movements of the company and sentiments in the tweet. This would definitely encourage the number of people investing in stocks if more positive tweets are posted. The development of a sentimental analyzer which can judge the sentiment of the tweet into three categories. Twitter might not be the right platform to extract tweets about the stock market. There is a stock-specific platform called StockTwits where insights on stocks can be shared.

**O) Clustering halal food customers: A twitter sentiment analysis**

This paper [15] by Mohamad M Mustafa, is based on Clustering analysis of Halal consumers using Lexicon based approach to obtain a sentiment score for the data along with PAM clustering to distinguish them into four distinct groups. Finally, topic modeling is used to label all four groups to detect hidden topics in the text. The analysis of the research involves a set of if-then rules which are capable of understanding the dynamics of the halal food consumer. Lexicon based approach might sometimes fail to understand the subtle emotions of the consumer particularly when it becomes sarcastic or ironic.

**P) Sentiment Analysis for Aadhar for twitter data- A hybrid classification approach.**

This paper [16] by Priya Kumari and Md. Tanwir Uddin Haider is based on sentiment analysis of four aspects of Aadhar using Twitter data. A hybrid method based approach is implemented by combining a machine learning algorithm with LSTM deep neural network to improve the classification score. A hybrid based approach where a machine learning algorithm is used with LSTM is a new approach in sentimental analysis technique to improve accuracy. The features used for the classification of the tweet are limited. This will affect the desired result for the twitter analysis.

**Q) Twitter Sentiment Analysis for Brand reputation of Smartphone companies in India**

The model developed by Sudhir Kumar Sharma, Mohit Daga, and Bhawna Gemini [17], is based on sentiment analysis of smartphones to analyze the brand reputation of them in India. The smartphone from six different brands is used to analyze the Net Brand Reputation score and find out which smartphones are preferred by the consumers. The Net Brand Reputation Score will be a useful tool to compare the percentage of likelihood of all the brands. The time frame of the tweet is very short as tweets might have been biased with the launch off any new brand phone.

**R) Sentiment Analysis based on US presidential Election 2016**

The paper [18] by Ramasubbareddy Somula, K. Dinesh Kumar, S. Aravindharamanan and K. Govind, is based on sentiment analysis of twitter data for US presidential election and predict who will win the election among the two candidates listed. A Lexical analysis based approach is used to calculate the sentiment scores of positive and negative tweets. The wide range of tweets collected has helped them to predict the desired result based on tweets. Sarcasm detection is not addressed.

**S) Twitter Sentiment Analysis using Machine Learning Techniques**

This paper [19] by Bac Li and Huy Nguyen, is based on a comparative study of how twitter sentiment analysis behaves for various machine learning models with Unigram, BiGram, and Object-oriented extraction methods. Information gain and Object-Oriented feature set are completely new techniques of feature engineering applied to model Naïve Bayes and SVM algorithms. The accuracy of the Bigram and Unigram outperforms the object-oriented feature set.

**T) Stock Prediction using Twitter Sentiment Analysis**

The paper [20] developed by Anshul Mittal and Arpit Goel, is based on sentiment analysis and machine learning principles to find the correlation between public sentiment and mark sentiment. Twitter data is used to predict the public mood which is then used to predict the stock market movements using Self Organizing Fuzzy Neural Networks. The desired accuracy was not achieved as mentioned using k fold sequential cross-validation technique. The correlation between twitter users and the stock market is vague because twitter is not the platform used by the investor to communicate on the stock market.

**Tab. 1: Comparative Analysis of Research Papers**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Paper**  **Title** | **No of Tweets** | **Feature Engineering** | **Machine Learning models** | **Weakness of the paper** |
| [1] | 10,48,576 | Knowledge-based preparation | Rule-based learning | Emotional icons |
| [2] | 4432 | Domain-Specific Ontology, Lexicon based approach | Baseline model, Domain-specific model | No of attributes are less |
| [3] | 3216 | Word2Vec, N-gram | Random Forest | Not the right platform |
| [4] | 3919 | Lexicon based approach, LDA | PAM clustering with topic modeling | Lexicon based is not effective |
| [5] | 2776 | TF-TDF | Baseline models into LSTM network | Limited features |
| [6] | 7330 | Data Analysis technique | Net brand reputation score | Tweet time frame is short |
| [7] | 1,00,000 | Lexical Analysis | Sentiment score | Sarcasm detection |
| [8] | 2,00,000 | Object oriented, IG with N gram | Naïve Bayes,  SVM | Accuracy |
| [9] | 2704 | Bag of Words | SVM, Naïve Bayes, DT, RF, k-NN | Less English tweets |
| [10] | 45,000 | StringtoVec method (WEKA) | Naïve Bayes, LR, SVM | Text pre-processing |
| [11] | 9191 | Tokenisation, Manual Labelling, Sampling | SVM, Decision Tree, Naïve Bayes | K value in Kfold has failed |
| [12] | 1950 | Feature vector with N-gram | Naïve Bayes, SVM | Accuracy is not good |
| [13] | 900 | N gram approach | Baseline models | General lexicon library |
| [14] | 96,64,952 | Tokenisation, POMS scoring | Time series, Z score | POMS , micro-texts |
| [15] | NA | Daily score, Granger causality, Tweet mapping | Linear and Logistic regression, SVM, Self-Organising Fuzzy Neural Network | Twitter is not the platform for stock markets |
| [16] | 71,06,297 | Vocabulary control, N-gram, Entity extraction | Topic based Sentiment scores | Tweets- shorter period |
| [17] | 31,150 | Content Analysis, Lexicon based approach, Linguistic Inquiry and Word Count | Automatic coding and Manual coding | Characteristics of patients are unknown |
| [18] | 71,95,829 | 10 fold cross validation with chi squared selection | Support vector machine and Naïve Bayes | Uncertainty in sentiment scores |
| [19] | 57,806 | Contextual Anomaly Detection, Chi square model, Information Gain | Naïve Bayes, Maximum Entropy, Support Vector Machine | Performance of the model starts decaying after 1000 features |
| [20] | 2500 | Tweet modifiers, Tweet Analysis | Chi-squared analysis | Improper use of keywords |

**Shortcomings in Related Work**

It can be observed that works related to “Sentiment Analysis of novel coronavirus based on data from social media” are very minimal. There are very few papers related to performing sentiment analysis on Pandemics. The major drawback of sentiment analysis is the pre-processing of tweets which contain tweets coupled with hashtags, emoticons, and links, thereby creating difficulties in determining the expressed sentiment.  Another important aspect is that the tweet analysis is only suitable for the English Language hence internationalization of word dictionary for all languages should be addressed. Sarcasm detection is a major drawback in sentiment analysis techniques. This provides a clear approach on how to proceed with the sentiment analysis of social media data.

# **scientific approach**

The goal of the paper is to perform topic modeling and sentiment analysis by a collection of tweets and compare these results over time. The purpose of this paper is to identify significant results from topic modeling and sentiment analysis and discuss the potential reasons for these results. Several steps are performed which include data extraction, data preparation, data cleaning, and then the two analysis methods.

**DATA COLLECTION**

**DATA CLEANING**

**DATA PRE-PROCESSING**

**TOPIC MODELLING**

**SENTIMENT ANALYSIS**

**SENTIMENT EVALUATION**

**TOPIC EVALUATION**

Fig. 1: Basic Steps of a scientific approach

**A) EXTRACTION OF TWEETS**

The tweets are extracted from twitter after the approval of twitter to get access with twitter API credentials like Consumer Key, Consumer Secret, Access Key, and Access Secret. Tweets posted under the hashtag “#coronavirus, #COVID19, #CoronaVirusOutbreak” will be extracted from twitter. The tweets are stored in a “.csv” file. The collection of tweets are dated from 02-02-2020. A total of 20000 tweets are collected.

**B) DATA PREPROCESSING**

The first step of data pre-processing involves the normalization of data. Normalization consists of three main tasks- removal of special characters, removal of URL’s, removal of hashtags. The next step involves the correction of spelling errors and replacing abbreviations and shorthand by their expansion using appropriate dictionaries.

**Tokenization and Lemmatization**

The next step involves tokenization, which involves dividing texts into a sequence of tokens which roughly correspond to words. These tokenized texts are assembled into sentences. Subsequently, these sequences of words are assigned a lexical category to each word. In other words, various forms of sentences are assigned to each word. The next step is all about mapping words to their base form. Finally, it can be seen that each word in the tweet is tagged with its corresponding lexical category and lemma. Besides, the vocabulary control technique can be applied to minimize term variation problems.

**Bag of Words**

Before topic modeling, the tokenized and lemmatized words are converted to bag of words. It is similar to a dictionary where the key is the word and value is the number of times the word occurs in the entire corpus. The words are filtered out depending upon their frequency of occurrence.

**TD-IDF Vectorization**

Term frequency- Inverse Document Frequency, commonly called as TF-IDF is a weighted vector model used to score and rank a document’s relevance in a given query. TF-IDF can be applied to the coronavirus dataset to obtain the use of most relevant words in the form of vectors.

**Exploratory Data Analysis**

A word cloud is plotted for the twitter data set to observe the most occurring word. A frequency analysis bar plot will also be plotted based on the frequently occurring N-gram words from N=2 to 6. A geographic plot can also be plotted to observe the location of tweets from where they are posted.

**C) TOPIC MODELLING**

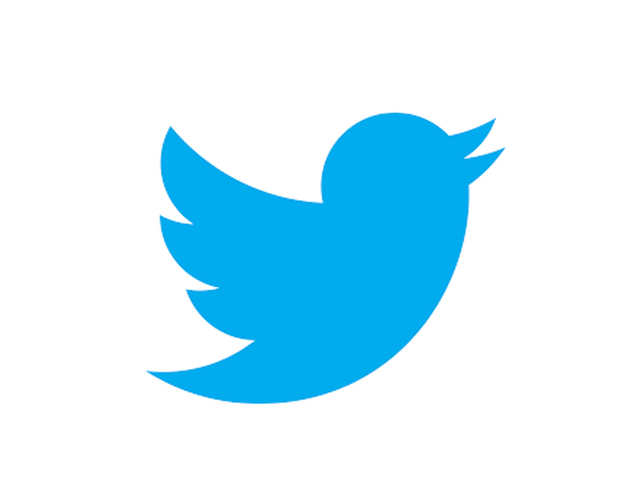
**LDA Analysis**

Latent Dirichlet Allocation (LDA) is used to classify text in a document to a particular topic. It is normally used to build a topic per document model and words per topic model. LDA is performed for the bag of words and the performance is evaluated by the vector of words. LDA is performed using TF-IDF vectors and the performance is evaluated. The method which performs the best is chosen as a topic modeling parameter.

**D) SENTIMENTAL ANALYSIS**

An aspect-based sentiment analysis technique is carried out to obtain sentiment scores and analyze people’s sentiment and views on coronavirus. N-gram based approach to the words by varying N from 2 to 6. N-gram refers to the number of words near the aspect that are considered for sentiment identification. The polarity of the closest words to the aspect identified is calculated by using SentiWordNet( SWN) to find out the positivity, objectivity, and negative scores. The sentiment scores obtained are compared with the positive, negative, and neutral scores to classify the polarity of their emotions.

In order to measure the performance of the method, precision, recall, and F-measure values are calculated. Precision represents the proportion of predicted positive cases that are real positive. On the other hand, recall calculates the proportion of actual positive cases that were correctly predicted and F-measure is the harmonic mean of precision and recall. A comparative analysis of precision, recall, and F-measure is performed with each N-gram value and the results can be interpreted.

DATA COLLECTION  Twitter API 🡪 Related Tweets 🡪 Corona Virus



DATA PREPROCESSING

Lemmatization and Tokenization

Stop Words Removal

TF-IDF VECTORIZATION

BAG OF WORDS (BOW)

Comparison of BOW AND TD-IDF Values

SENTIMENT ANALYSIS

using SentiWordNet(SWN)

TOPIC MODELLING

(LDA Method)

Fig. 2: Detailed Description of Sentimental Analysis

# **scientific approach**

Topic Modelling and Sentimental Analysis is performed on the dataset obtained from twitter. The dataset extracted from twitter is cleaned to obtain the right amount of words or tweets needed for the extraction of entities. The bag of words concept is used to exhibit the most occurring word among the tweets. Term frequency-Inverse Document frequency is used to rank words and give them scores based on their occurrence in the document. The best model among them is chosen and Latent Dirichlet Allocation (LDA) is applied at the document level of the text to obtain frequently discussed topics. The topics obtained will be visualized using Word Cluster to magnify all the important topics discussed by people on twitter about coronavirus.

Sentimental scores are calculated using the polarity measure technique. For the N-gram word, the polarity scores are calculated. The polarity of a word is divided into positive, negative, and neutral. For each N-gram word, the polarity scores are calculated. It can be seen that each word has combined scores of positive, negative, and neutral. The total sum is calculated for all the polarity measures and the word which has a higher positive value will be considered as a positive contextual word. Similarly, the word which has a higher negative value will be considered as a negative contextual word.

Performance metric parameters are taken into account to compare which N-gram model has performed better. Precision and Recall values are measured for each N-gram model and the resulting model with better precision is decided as the best performing N-gram model. F1 scores are also obtained which is the geometric mean of precision and recall values.

As a part of an experiment, the basic machine learning model (Naïve Bayes, Logistic Regression, Random Forest, Multi-layer Perceptron) is performed on the best N-gram model. This part of the research project will be considered as an out of scope experiment just to know how the model performs. Accuracy is used as a basic criterion to perform a comparative analysis of the three basic machine learning models.

**A) Data Description**

Twitter is a social networking and microblogging site that allows real-time users to post real-time messages called tweets. Tweets are short messages usually restricted to 140 characters in length. Due to the nature of this microblogging site, people use acronyms, make spelling mistakes, use emotions, and mention other characters that express special meanings.

Basic terminologies used in twitter tweets:

**Emotions:** These are facial expressions pictorially represented using punctuation and letters as they represent/express a user’s mood.

**Target:** Users of twitter use the ‘@’ symbol to refer to the other users on the microblog. Referring to other users alerts the other user by sending notification messages.

**Hashtags:** Users usually use hashtags to mark topics. This is primarily done to improve the visibility of tweets and convey the core content of the tweet in a single word.

Data in the form of raw tweets is acquired by using python library **‘tweepy’** which provides a package of simple twitter streaming API. This API allows two modes of accessing tweets: Sample Stream and General Stream. Sample Stream simply delivers a small, random sample of all tweets streaming in real-time. Filter Stream delivers tweets that match certain criteria. It can filter the delivered tweets according to three criteria,

* Specific keyword(s) to track/search for in tweets.
* Specific twitter user(s) according to their user-ids.
* Tweets originating from a specific location(s) only for geotagged tweets.

The purpose of the experiment has no special filter criteria, so it is always better to extract tweets using Sample Stream mode.

The tweet is acquired during the time frame from 31st December 2019 to 1st March 2020 because the virus outbreak expanded during that time frame. It is expected to

A tweet obtained by this method has a lot of raw information in it which may or may not find useful for a particular application. A total number of 1 million tweets have been extracted from twitter. The tweet messages are randomly generated from twitter which means that there is no specific language from which the tweets have been extracted.

It comes in the form of a python dictionary data type with various key-value pairs. Some of them are,

**Created\_at:** This attribute defines the time and data when the tweet was created.

**Tweet\_id:** This attribute refers to 64-bit unsigned integer. The full ID is composed of a timestamp, a work number, and a sequence number.

**Tweet\_text:** This attribute consists of the tweet messages posted by the user.

**Screen\_name:** This attribute refers to the user name of the twitter user.

**Name:** This attribute refers to the full name of the twitter user.

**Account\_creation\_date:** This attribute refers to the date and time during which the account holder created the account.

**URLs:** This attribute refers to the twitter URL from which the tweets have been extracted.

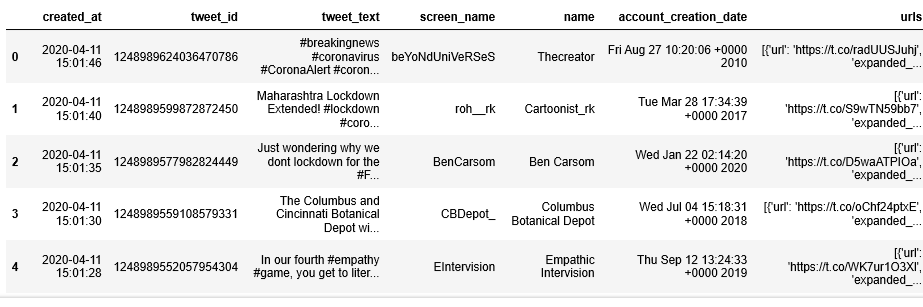
****

Fig. 3: Dataset obtained from twitter

Since there is a lot of information from the dataset, it is better to filter out some of the unnecessary attributes.

**B) The architecture of Sentimental Analysis**

Data Pre-processinginvolves cleaning the tweets by normalizing, tokenization, and lemmatization of tweets.

Normalization involves three processes,

1. Removal of special characters
2. Removal of URL’s
3. Removal of hashtags

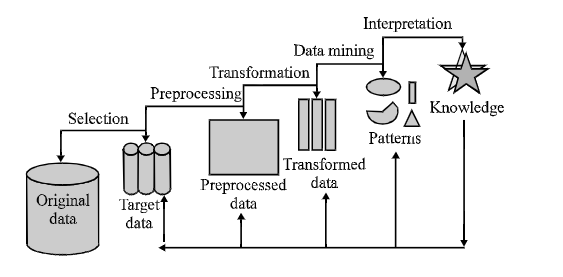


Fig. 4: Knowledge discovery steps

The abbreviation of the words are expanded and errors in their spellings are corrected using the Hunspell dictionary. The words are tokenized and divided into tokens. These words are assigned a lexical category. The words are mapped to their base. The words are tagged with their lexical category and lemma. Stanford core NLP library can be used for the lemmatization process.

**TOKENIZATION, LEMMETIZATION**

**NORMALISATION, EXPANDING ABBREVATIONS AND CORRECTING SPELLINGS**

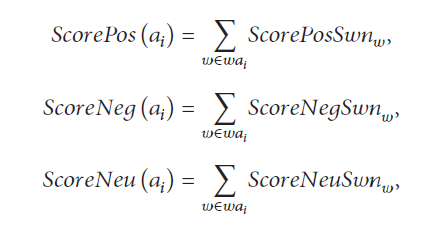
**CLEANED TEXT**

**RAW TEXT**

Fig. 5: Data preprocessing steps

In the sentiment analysis process, the main approaches are topic modeling and machine learning using sentiment scores. Bag of Words and TF-IDF factorization are important steps in topic modeling and sentiment analysis.

Bag of words is one of the easiest ways to represent text as a form of vector. It ignores the ordering between words in document representation i.e. only the count of words is important. The bag of words does not bother about the sentence structure or grammatical construction. It builds a dictionary using words available in the text and counts the frequency of words.

**Term Frequency-Inverse Document Frequency** is a weighted statistical measure to evaluate the importance of the word in a corpus. TD-IDF is not a single method but a class of techniques: term frequency and inverse document frequency. The similarity between the documents is measured by TF and term importance is measured by IDF. So, the multiplication of TF and IDF of a word produces a frequency of the word. If term frequency is TF(wi, D) and document frequency is DF(wi). Then, from DF(wi ), inverse document frequency IDF(wi ) can be calculated as follows:

1. Tf (wi, D) = (Number of time the word occurs in the text) / (Total number of words in the text)
2. Idf (wi) = (Total number of documents) / Number of documents with word t in it)
3. Tf (wi, D).Idf (wi) = (Tf \* Idf)

The TF-IDF of feature *wi*  for document D is then calculated as the product: Tf(wi, D).Idf(wi). The word with high TF-IDF scored in a document are frequently occurred in that document and deliver the most important facts about the document

**Latent Dirichlet Allocation** is used to define a number of topics that are present in the collection of documents. Training an LDA model on a document with a topic corresponds with finding the document and topic vector that best explains the data. It also represents the different topics that the document represents and how much of each of the topics is present in a document.

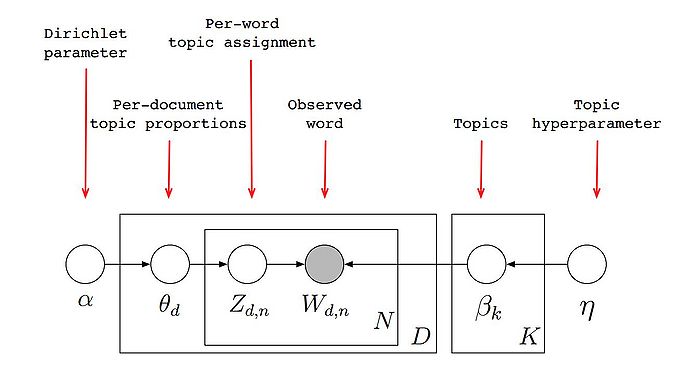


Fig. 6: Topic Modelling – LDA

The most frequently occurred topics are generated using the LDA technique.

**Sentimental scores** are calculated using the polarity of the sentence. N-gram words are obtained for three scenarios,

* **N-gram before method** uses N-gram words taken before the aspect of the tweet.
* **N-gram after the method** used N-gram words taken after the aspect of the tweet.
* **N-gram around method** used N-gram words before and after the aspect of the tweet.

The polarity of each word is calculated using the formulae,

The polarity of the closest words to the aspect identified is calculated by using **SentiWordNet (SWN).** Each entry in SWN has multiple senses. The positive, negative, and neutral scores are calculated.

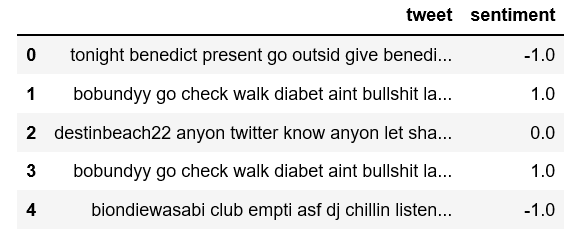
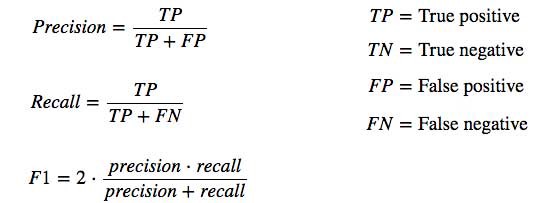


Fig. 7: Sentiment Scores of Dataset

The precision and recall values for each N-gram model are calculated and an analysis of the better performing N-gram model is obtained.



Baseline models for sentimental analysis like Naïve Bayes, Logistic Regression, and Random Forest classifiers are tested and performed to study the dataset.

**Naïve Bayes** is a classification method based on Bayes’ theorem that derives the probability of the given feature vector being associated with a label.

https://miro.medium.com/max/275/0*cZr5cDrKaIDazSbZ.gif

**Logistic regression** is a linear classification method that learns the probability of a sample belonging to a certain class.

https://miro.medium.com/max/225/0*VE_rWLv4_Gu8cBj8.gif

**Random Forest** is an ensemble machine learning model good for handling large values and improving the accuracy of the dataset. It is a combination of decision trees.



Fig. 8: Random Forest Classifier

**Multilayer Perceptron** is one of the artificial neural network algorithm used to feed forward the input with a number of hidden layers and obtain dependencies between input and output.

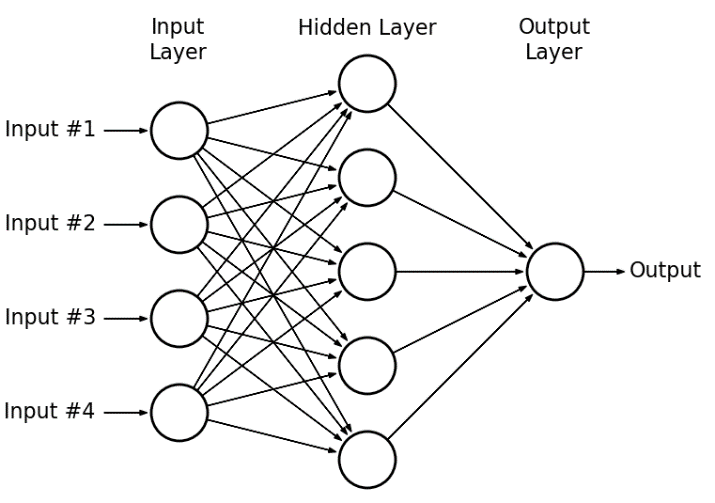


Fig. 9: Multilayer Perceptron

**C) Algorithmic Approach and Implementation**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **INPUT**  **Tweets- 20K**   |  |  |  | | --- | --- | --- | | Created\_at | UTC time when the tweet was created | DateTime64 | | Tweet\_id | Unique identifier of the tweet | Int64 | | Tweet\_text | UTF -8 text of the status update | Object(String) | | Screen\_name | Username of the twitter user- | Object(String) | | Name | Full name of the twitter user | Object(String) | | Account\_  creation\_date | Twitter join date | Object(String) | | URLs | A URL provided by the user in association with their profile | Object(String) | |
| **ALGORITHM**   1. Corona Virus – Tweets are extracted from twitter 2. Data Pre-processing  * Data Cleaning ( Normalisation, Sentence Splitter, Stop Words removal) * Data Preparation( Tokenisation, POS tagging)  1. Feature Engineering  * Bag of Words (BOW) * TF-IDF Vectorization * Comparison and analyzing the better technique  1. Topic Modelling  * Latent Dirichlet Analysis  1. Sentiment Scores  * N-gram method ( before, after, around) * Polarity Check * Performance metrics of N-gram models ( Accuracy, Precision, Recall)  1. Machine Learning  * Naïve Bayes – Accuracy * Logistic Regression – Accuracy * Random Forest – Accuracy * Multilayer Perceptron - Accuracy * Comparison of models with better accuracy |
| **OUTPUT**  Accuracy, Precision, Recall of N-gram methods are compared. The best N-gram model with better accuracy is chosen.  Accuracy of Naïve Bayes, Logistic Regression, and Random Forest are compared and the best model is chosen. |

Tab. 2: Algorithmic Approach

**D) Discussion details**

Data extraction from social media is very helpful to mine opinions of people. Twitter is chosen as the data extraction platform because twitter has a separate API that allows us to read and write data. It helps to read tweets, create user profiles, data of followers since it identifies various twitter applications and the users who register using OuAuth authentication and authorization. Tweets can be extracted in JSON format. It is really helpful to read and write data.

Topic modeling provides us a method to organize, understand, and summarize a large collection of textual information. It helps in discovering hidden topic patterns that are present across the collection.

The topic modeling of coronavirus enables us to find abstract topics related to coronavirus discussed on twitter. It helps to find the hidden semantic structures in the text body.

Classification algorithms are chosen because they are simple to implement, robust to noisy data, and helpful while training large data.

Naïve Bayes algorithm can be used because it is fast, requires less training data, and can make highly probabilistic predictions. It is one of the baseline algorithms to perform a classification task.

The logistic Regression algorithm helps the data to give relevant predictions and its direction of the association. As it is easy to implement, interpret, and very efficient to train.

Random forest algorithm is one of the best supervising algorithms and has a reliable feature importance estimate.

Multilayer Perceptron solves the complex relationship problems between inputs and outputs by adding one or more hidden layers especially while solving non-linear problems.

Each algorithm has its own use and it can be used to model and train twitter data. These baseline models in classification tasks pave the way to choose neural networks in the future which might help to get better accuracy.

# **implementation**

1. **DATA VISUALISATION**

Visualizing data enables to get more insights about the data. Tweets especially can be visualized based on the following criteria, number of tweets/re-tweets, number of followers, the word count of tweets, the positive/negative sentiment of tweets, and their respective word clouds, geospatial representation of tweets, etc.

In this paper, the visualization of data will be more focused on top words in the dataset, word cloud of the total number of tweets, their respective positive and negative word cloud, the total number of words in each tweet, POS tagged words, the ratio of positive, negative and neutral tweets.

Before data visualization, preprocessing steps have removed the most common and most rare words in order to obtain accurate results. The top words visualized below don’t have the most common words like coronavirus but it will have most repeatedly used words in the tweets. They are plotted in their decreasing order to showcase their frequency of occurrences.

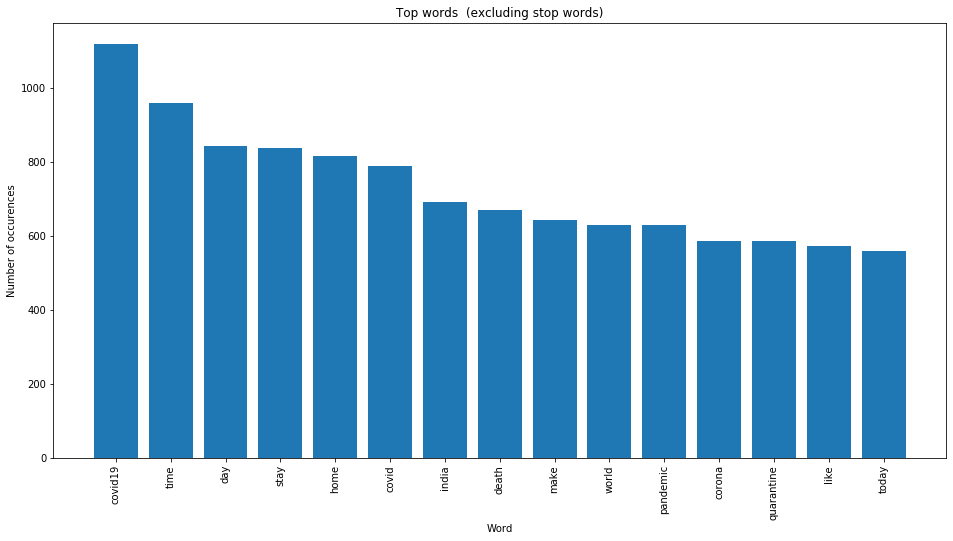


Fig. 10: Top words (excluding stop words)

It is evident that the top words include words like time, COVID-19, home, pandemic, quarantine, etc., which were included in most of the tweets.

The graph below represents the headline word lengths .i.e., the number of average words in each of the tweets posted.

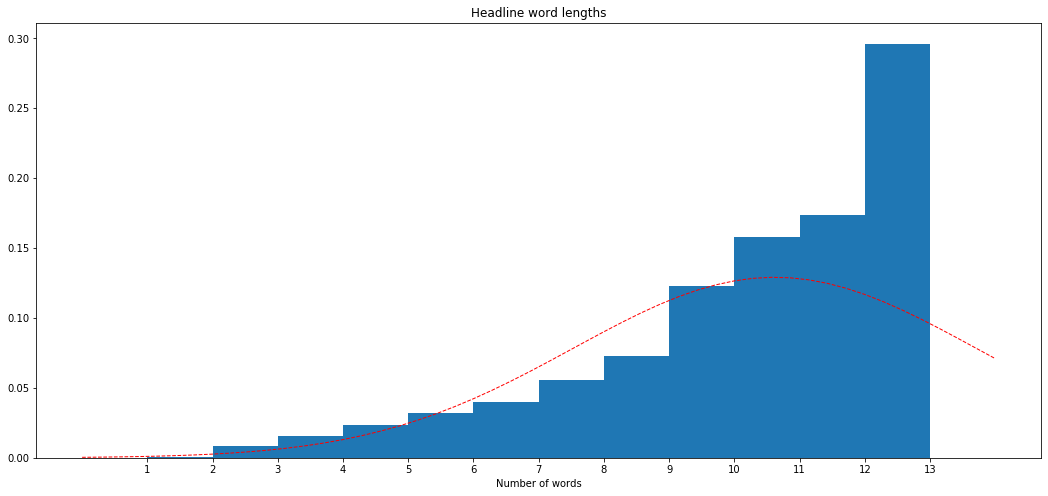


Fig. 11: Headline word lengths

It can be observed that users have used the total tweet limit of 13 words in their tweets. An average of 10-14 words was found to be the tweet length posted by twitter users. This shows the interest of people to share their views on the recent outbreak.

Parts of Speech tagged words are also represented in the analysis of data. The graph below represents the number of words that are tagged to a particular POS.

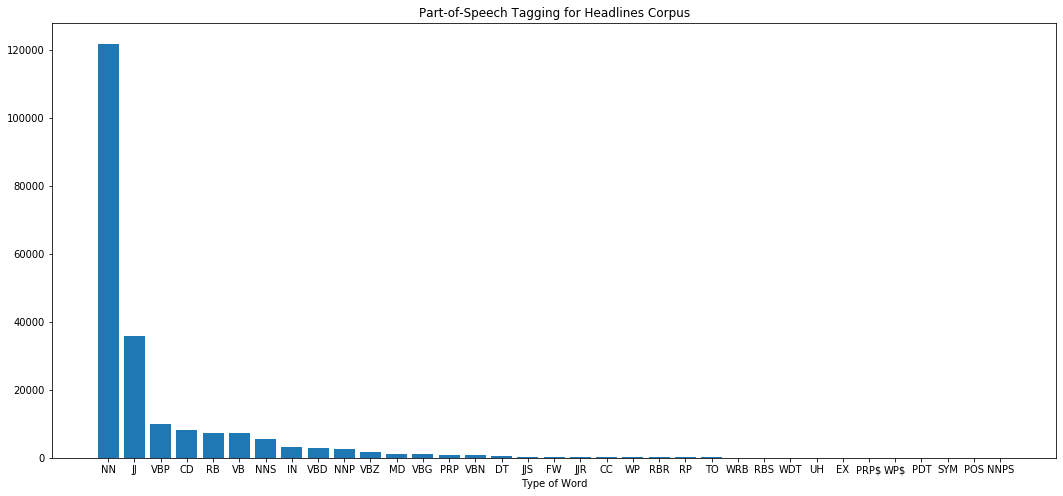


Fig. 12: Part of Speech Tagging for Headline Corpus

It can be seen that most of the words in the tweets are nouns followed by adjectives and many more.

Word clouds visually represent the keywords present in the data. Positive word clouds represent the positive keywords whereas negative word clouds represent negative keywords.

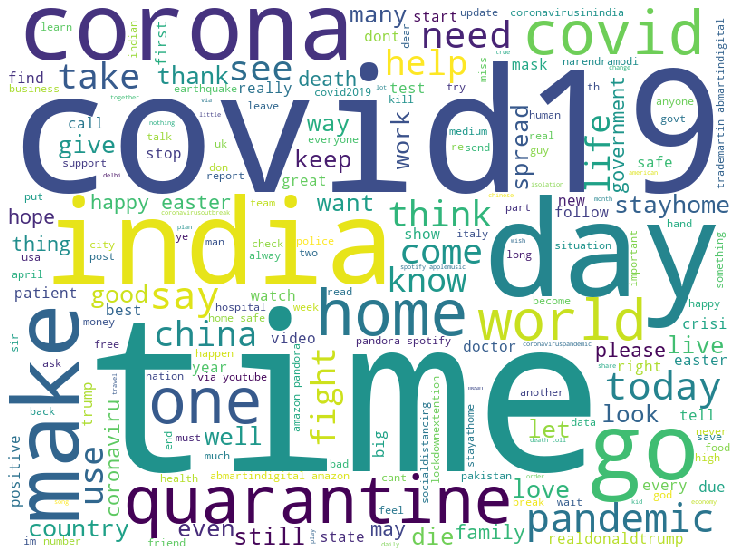


Fig. 13: Word cloud of all tweets

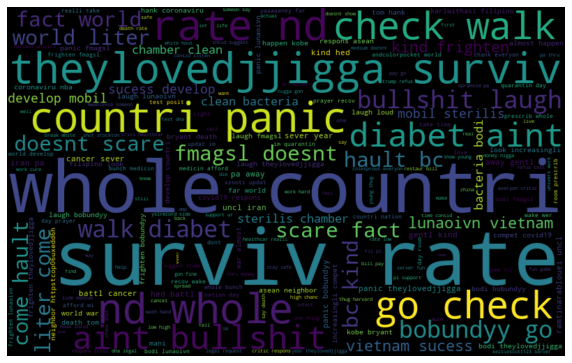


Fig. 14: Word cloud of Positive tweets

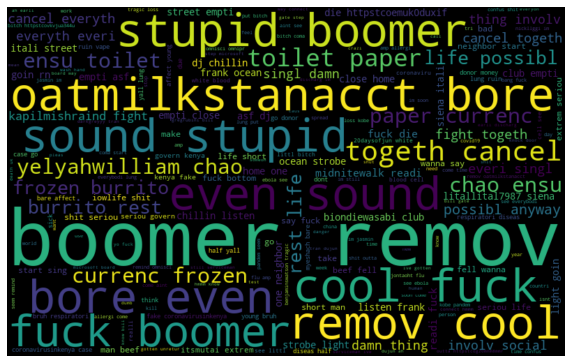


Fig. 15: Word cloud of Negative Tweets

The polarity of tweets are represented in a bar graph below,

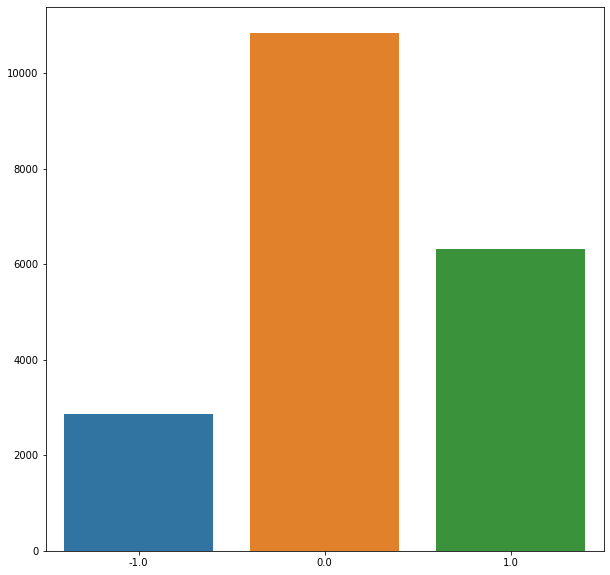


Fig. 16: Polarity of Tweets

It is seen that the number of tweets that are neutral and positive is more because people tend to post awareness measures and post tweets by spreading positivity in twitter which is reflected in the ratio of the polarity of the tweets.

1. **TOPIC MODELLING**

Topic modeling using LDA is the most sought technique to obtain distinct topics in the dataset. Topic modeling can be more effective when researchers want to obtain the semantic structure in the data set. It is very much helpful to understand the underlying topics widely discussed in the streaming platform.

There are two approaches to perform topic modeling, LDA (Latent Dirichlet Analysis) and LSA (Latent Semantic Analysis).

The paper focuses primarily on the LDA technique because it gives better results for tweets compared to LSA in general. In order to exhibit the better performance of the LDA technique, the results of LDA and LSA are compared and interpreted.

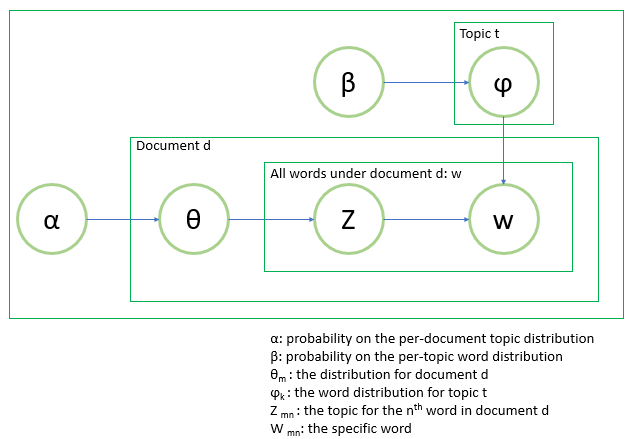


Fig. 17: Workflow of Topic modeling using LDA

The number of topics to be chosen can be a variable and the total number of n\_topics chosen is eight in this scenario.

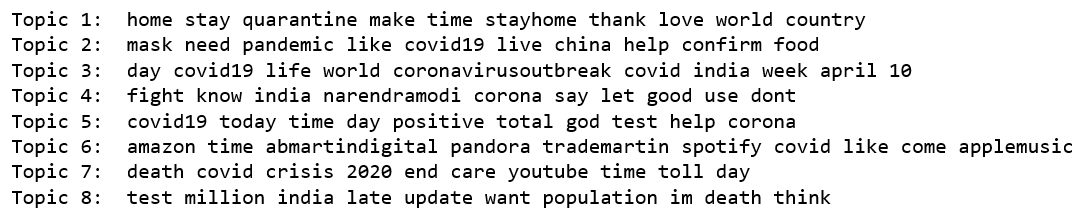


Fig 18:Topics modeled

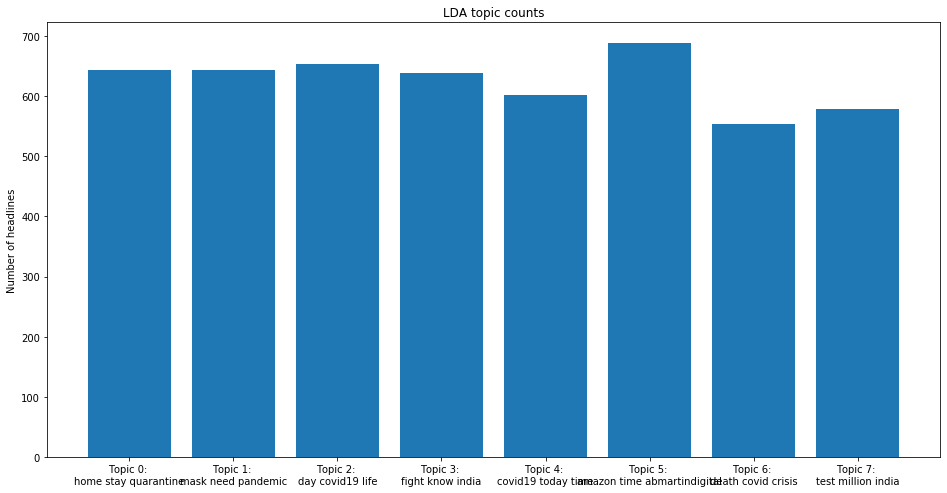


Fig. 19: Graph showing LDA topic counts

The frequency of the top 8 topics in tweets is represented in the bar graph above. The most talked topic in twitter is home quarantine, pandemic, Narendra Modi's view on coronavirus, etc.

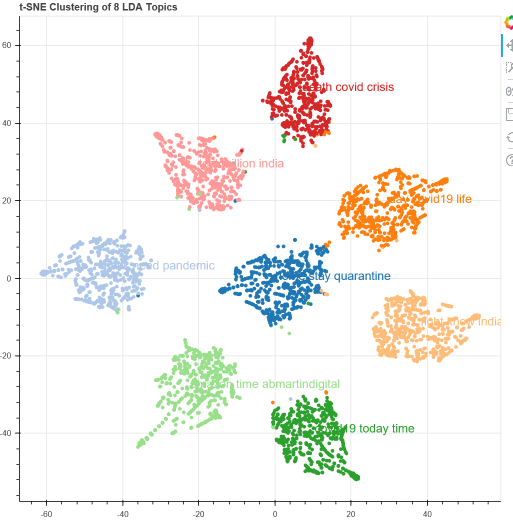


Fig. 20: LDA modeling

From the above figure, it is evident that the above clusters are distinct and clear as compared to Latent Semantic Analysis.

LSA model clustering is clumsy and the clusters are seen to be unclear and vague as it focuses more on reducing the dimensions of the matrix.

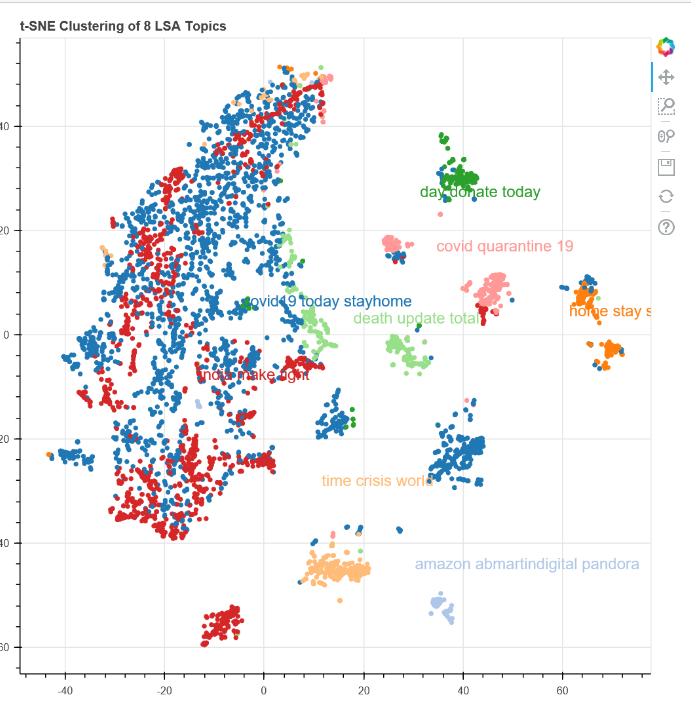


Fig. 21: LSA Modelling

1. **MACHINE LEARNING**

The sentiment scores are obtained using SentiWordNet for about 1 million twitter dataset. Considering the efficiency of processing, only 20000 tweets were considered for modeling.

The sentiment scores are obtained as continuous values which are then converted into a ‘0’ for neutral, ‘1’ for positive, ‘-1’ for negative.

The pre-processed dataset is now modeled with three machine learning models, Logistic Regression, Multinomial Naïve Bayes, and Random Forest.

Logistic Regression is modeled with basic parameters and selecting C value as 0.01.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted**  **Actual** | **Class 1** | **Class 2** | **Class 3** |
| **Class 1** | **63** | **543** | **0** |
| **Class 2** | **0** | **2145** | **2** |
| **Class 3** | **0** | **603** | **604** |

Tab. 3: Confusion Matrix – Logistic Regression

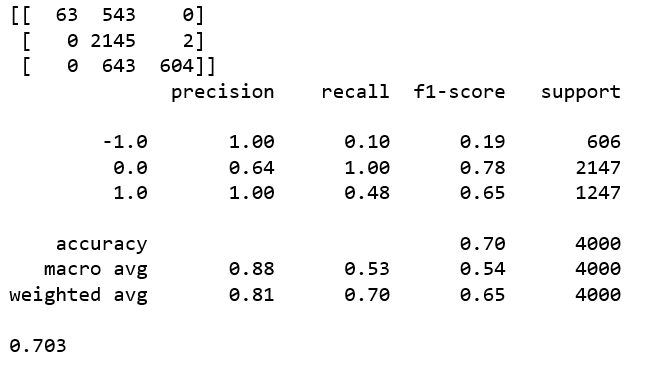
****

Fig 22: Classification Report of Logistic Regression

Multinomial Naïve Bayes model is implemented with basic parameters and the confusion matrix of the model is found to be,

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted**  **Actual** | **Class 1** | **Class 2** | **Class 3** |
| **Class 1** | **492** | **50** | **64** |
| **Class 2** | **98** | **1898** | **151** |
| **Class 3** | **94** | **49** | **1104** |

Tab. 4: Confusion Matrix- Naïve Bayes

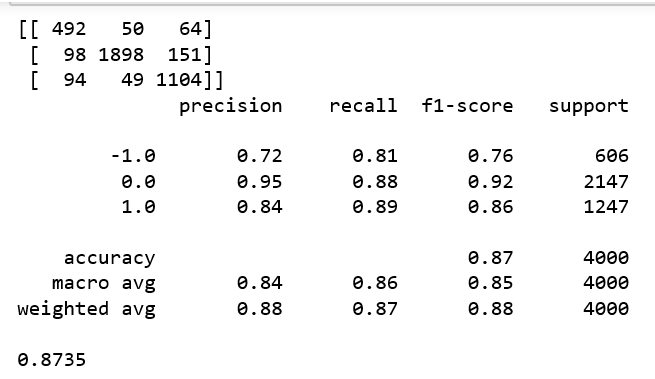


Fig 23: Classification Report of Multinomial Logistic Regression

In order to improve the accuracy of the data, an ensemble technique (Random Forest) was implemented. Random Forest classifiers are modeled with basic parameters keeping n\_estimators as 100.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted**  **Actual** | **Class 1** | **Class 2** | **Class 3** |
| **Class 1** | **483** | **78** | **45** |
| **Class 2** | **8** | **2119** | **20** |
| **Class 3** | **24** | **69** | **1154** |

Tab. 5: Confusion Matrix – Random Forest

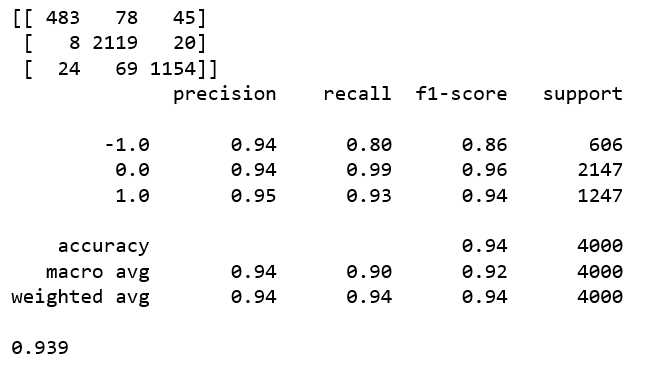


Fig 24: Classification Report of Random Forest

The Multilayer Perceptron model is implemented by fixing the hidden layer size as 60 and keeping their learning rate as constant. The confusion matrix is found to be,

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted**  **Actual** | **Class 1** | **Class 2** | **Class 3** |
| **Class 1** | **497** | **48** | **61** |
| **Class 2** | **50** | **2015** | **82** |
| **Class 3** | **44** | **55** | **1148** |

Tab. 6: Confusion Matrix – MLP

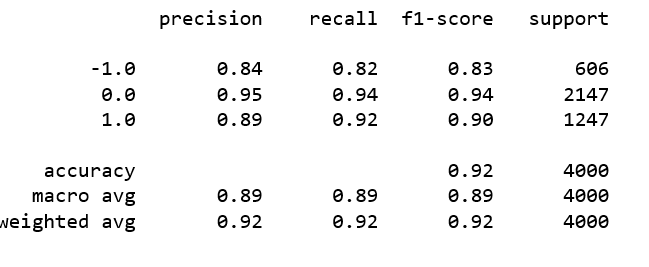


Fig 25: Classification Report of Multi-Layer Perceptron

# **evaluation details**

1. **Operational Analysis**

|  |  |
| --- | --- |
| **MODELS** | **ACCURACY in %** |
| LOGISTIC REGRESSION | 70.3 |
| NAÏVE BAYES | 87.35 |
| **RANDOM FOREST** | **93.9** |
| MULTILAYER PERCEPTRON | 91.50 |

Tab. 7: Comparison of Accuracy values

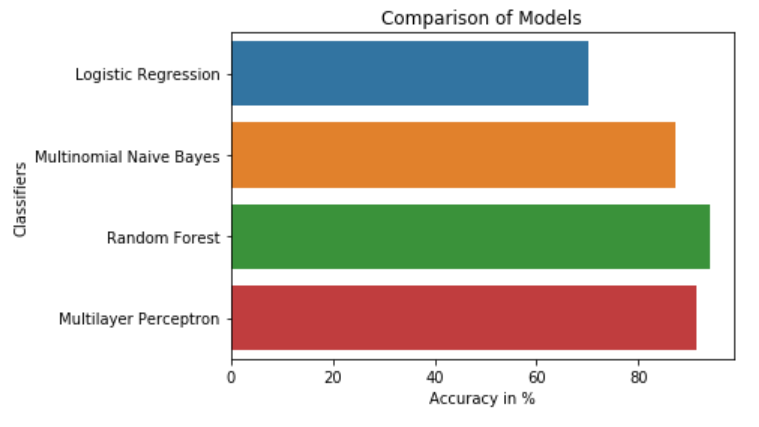


Fig 26: Accuracy of Classifiers

The accuracy of the random forest is high compared to logistic regression and Naïve Bayes models.

The precision, recall, and F-1 scores are calculated for the models using a confusion matrix.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Precision in %** | **Recall in %** | **F-1 Scores in %** |
| Logistic Regression | 88.2 | 52.91 | 54.11 |
| Naïve Bayes | 83.5 | 85.04 | 84.64 |
| **Random Forest** | **93.98** | **90.32** | **91.93** |
| Multilayer Perceptron | 89 | 92 | 90 |

Tab. 8: Table representing Comparison of Models

Random Forest classifier is the best performing model when compared with Logistic Regression, Naïve Bayes Model, MLP with their precision, recall, and F-1 score values.

The ensemble model with a number of decision trees has performed better than basic classification algorithms and surprisingly performed better than a neural network model.

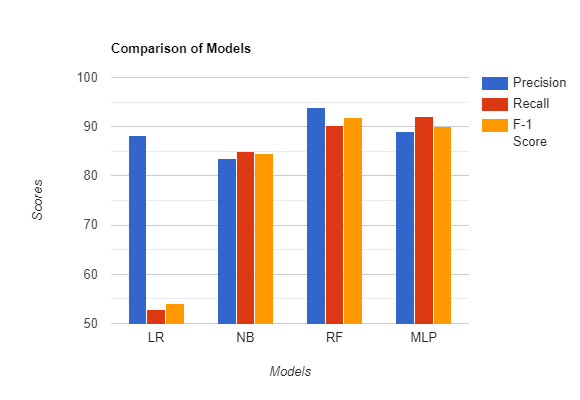


Fig 27: Comparison of models

1. **Complexity Analysis**

The complexity analysis predicts how fast/slow a model performs. It is usually a numerical function that relates the time function with the size of the input.

Data preprocessing involves TF-IDF vectorization which takes O(D\*Lavg) time to clean and process data. D represents the number of documents and Lavg represents the average length of the document.

Logistic Regression takes around O(N) time to model the dataset. N represents the number of tweets collected.

Naïve Bayes Model takes around O(N\*d) time to model the dataset. N usually represents the number of features of the dataset. In this case, N is used for the number of tweets, and d represents the dimensionality of the features.

Random Forest takes O(V\* Nlog(N)) time for unpruned decision trees. N represents the number of tweets and V for the number of features. While building a random forest, the number of trees to be built and the number of features has to be specified.

Multilayer Perceptron model takes around O(T\*N( ab+bc+..). T represents the number of training examples, N represents the number of epochs and a,b,c represents the number of hidden layers. MLP, in this case, was modeled with basic parameters by keeping learning size as 60.

1. **Efficiency Analysis**

The efficiency analysis describes the total time taken by the CPU to run the model and memory usage of the dataset.

The memory usage occupied by the dataset to store 20K tweets is 1.1MB.

Each task takes separate time for processing the data. The detailed analysis of total time taken by CPU is given below,

|  |  |
| --- | --- |
| **STEPS** | **CPU TIME(s)** |
| Data Collection | 10s |
| Removing Punctuation | 0.001s |
| Removing Stop Words | 0.0075s |
| Lemmatization | 0.0052s |
| Removing urls | 0.0046s |
| Removing html | 0.0035s |
| Removing chats | 0.0078s |
| Removing emojis | 0.0057s |
| Removing emoticons | 0.00529 |
| Count Vectorizer  (To get top words) | 0.0045s |
| POS tagging | 0.002s |
| LDA modeling | 0.008s |
| TF-IDF vectorization | 0.0046s |
| Bag of Words | 0.004s |
| Logistic Regression  (Training) | 0.0042s |
| Logistic Regression  (Testing) | 0.049s |
| Naïve Bayes  (Training) | 0.048s |
| Naïve Bayes  (Testing) | 0.039 |
| Random Forest  (Training) | 0.046s |
| Random Forest  (Testing) | 0.042s |
| Multilayer Perceptron  (Training) | 0.014s |
| Multiplayer Perceptron  (Testing) | 0.011s |

Tab. 9: Efficiency Analysis

# **comparison with related works**

There is not any paper currently related to sentiment analysis of coronavirus using social media data. There are only a few papers related to Ebola Virus and Flu trend prediction.

|  |  |  |
| --- | --- | --- |
| **Proposal** | **Model Implemented** | **Accuracy achieved** |
| Ebola Virus | Sentiment Scores | - |
| Flu Trend Prediction | Support Vector Machine | 75% |
| Diabetes | Aspect level approach | 85.5% |
| Aadhar | Hybrid approach  (NB+LSTM) | 85.3% |
| Stock Prediction | SOFNN | 73.33% |
| Emotional detection | K-Fold | 85.1% |
| Corona Virus | Random Forest | 93.9% |

Tab. 10: Comparison with Related Works

The accuracy obtained by the random forest algorithm in this paper seems to be better compared to other related papers. The accuracy is definitely improved in the case of models.

# **conclusion**

Corona Virus is becoming one of the most vastly discussed topics nowadays and performing sentiment analysis on this topic can open many doors to the researchers and analysts to make good decisions based on the views of the people. The tweets collected from twitter based on coronavirus are used for topic modeling and sentiment analysis.

The topic modeling performed on the dataset appears to be more intuitive and represents the frequent topic being discussed on the twitter platform. Latent Dirichlet Analysis technique used in this paper appears to be the right method for obtaining the semantic structure of the tweets. LDA method is found to be better than the LSA technique. Sentiment scores of the tweets are found more to be neutral rather being positive or negative. It shows people are more aware of spreading good opinions on twitter than panicking people by posting death counts or any other devastating news. Random forest classifier performs better than of the above-implemented models with an accuracy of 93.9% which is very much high for classification analysis when compared with other related papers on viruses or pandemics. The main reason for random forest classifier is that it reduces over-fitting especially when there is huge data and the variance is very less due to multiple decision trees used in the model

Sentiment analysis was performed on a dataset with no language restriction whereas topic modeling was performed on a dataset with English tweets as they represent distinct topics discussed.

Despite the accuracy of the model seemed to be high, one of the drawbacks of the paper is that it couldn’t remove bat emoji. Since the tweets are based on coronavirus, one of the carriers of coronavirus is a bat, emoji’s representing bat were more often used by the users. Filtering out of those emojis were not addressed in this paper.

Sentiment Analysis of the novel coronavirus is still in its infant stage as there are not any related papers at present. As this seems to be a vastly growing field, people will be working on sentiment analysis of coronavirus tweets in the near future and greater accuracy can be achieved with baseline models.

References

1. Srinivasu Badugu and Matla Suhasini,” Emotional Detection of Twitter data using Knowledge base approach”
2. Valikannu Ramanathan and T Meyappan ,”Twitter Text Mining for sentimental analysis on people’s feedback on Omar’s tourism”
3. Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda and Babita Maji ,”Sentiment analysis of Twitter data for predicting stock market movement”
4. Mohamad M Mustafa,”Clustering halal food customers : A twitter sentiment analysis”
5. Priya Kumari and Md. Tanwir Uddin Haider ,”Sentiment Analysis for Aadhar for twitter data- A hybrid classification approach
6. Sudhir Kumar Sharma, Mohit Daga, and Bhawna Gemini ,”Twitter sentiment analysis for brand reputation of smart phone companies in India”
7. Ramasubbareddy Somula, K. Dinesh Kumar, S. Aravindharamanan and K. Govind ,”Sentiment analysis based on US presidential election 2016”
8. Bac Li and Huy Nguyen ,”Twitter sentiment analysis using machine learning techniques”
9. Jose Carlos Santos and Sergio Matos ,”Analyzing twitter and web queries for flu prediction”
10. Marium Nafees, Hafsa Dar, Ikram Ullah Lali and Salman Tiwana ,”Sentiment analysis of polarity in product reviews in social media”
11. Sonia Anastasia and Indra Budi ,”Twitter sentiment analysis of online transportation providers”
12. Akshay Amolik, Niketan Jivane, Mahavir Bandhari and Dr. Venketesan ,”Twitter analysis of movie reviews using machine learning techniques”
13. María del Pilar Salas-Zárate, JoséMedina-Moreira, Katty Lagos-, Harry Luna-Aveiga, Miguel Ángel Rodríguez-García and Rafael Valencia-García Ortiz,”Sentiment analysis of tweets on diabetes : An aspect level approach”
14. Johan Bollen, Huina Mao and Alberto Pepe ,”Modelling public mood and emotion: Twitter sentiment and socio-economic phenomena”
15. Anshul Mittal and Arpit Goel ,”Stock prediction using twitter sentiment analysis”
16. Erin Hea-Jin Kim, Yoo Kyung Jeong ,Yuyoung Kim, Keun Young Kang and Min Song,” Topic based content and sentiment analysis of Ebola virus of twitter and news”
17. Oscar N, Fox PA, Croucher R, Wernick R, Keune J, Hooker K ,”Machine learning, sentiment analysis and tweets: An Examination of Alzheimer’s Disease Stigma on Twitter”
18. Xiaoran An, Auroop R Ganguly, Yi Fang, Steve B Scyphers, Ann M Hunter and Jenifer G Dy ,”Tracking climate change opinions from twitter data”
19. Koosha Golmohammadi and Osmar R Zaiane ,”Sentiment analysis on twitter to improve time series contextual anomaly detection for detecting stock market manipulation”
20. Lauren E Sinnenberg, Christie L DiSilvestro, Christina Mancheno and Karl Dailey,”Twitter as a potential data source for cardio vascular disease research”.
21. Ho, Tin Kam (1995),” Random Decision Forest (PDF). Proceeding of 3rd International Conference on Document analysis and recognition, Montreal, QC, 14-16 August 1995. Pp. 278-282. Archived from the original (PDF) on 17th April 2016. Retrieved 5 June 2016 “.
22. David A Freedman (2009),”Statistical Models: Theory and Practice

Cambridge University Press”.

1. Haykin Simon (1998),” Neural networks: A comprehensive foundation ( 2 ed.) Prentice Hall”.
2. Cramer J. S (2002), The origins of logistic regression. 119. Tinbergen Institute
3. McCallum Andrew, “Graphical Models, lecture2: Bayesian Network Representation”.
4. T.Mitchell,” Generative and Discriminative Classifier: Naïve Bayes and Logistic Regression”.
5. Blei D.M, Laffert J.D,” Topic Models”.

**APPENDIX**

**A1. IMPLEMENTATION DETAILS**

1. **Platform details**

**Anaconda Navigator**

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS, and Linux.

Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages, and update them – all inside Navigator.

1. **Language Specifications**

**Python**

Python is a language of programming that is interpreted, of high level, of general use. Python's design philosophy, developed by Guido van Rossum and first published in 1991, emphasizes the readability of code with its prominent use of large whitespace. The language constructs and object-oriented methodology help programmers write basic logical code for large and small projects. The language is extensively used for garbage value collection. It is a combination of object level programming and functional programming.

**Python 3**

Python 3.0 final was released in December 3rd 2008. Python 3.0 is a new version of the language that is incompatible with the 2.x line of release. The language is mostly the same, but many details especially how built in object like dictionaries and strings work have changed considerably, and a lot of deprecated features have been removed. The standard library has been reorganized in a few prominent places.

Python 3.7

This new Python version has been in development since September 2016, and now we all get to enjoy the results of the core developers’ hard work.

While the documentation gives a good overview of the new features, this article will take a deep dive into some of the biggest pieces of news. These include:

* Easier access to debuggers through a new breakpoint() built-in
* Simple class creation using data classes
* Customized access to module attributes
* Improved support for type hinting
* Higher precision timing functions

More importantly, Python 3.7 is fast.

1. **3rd Party Libraries**

**Pandas**

Pandas is the most powerful, flexible and easy to use open source data analysis tool built on python programming language.

**Numpy**

NumPy is the fundamental package for scientific computing with Python. It contains among other things:a powerful N-dimensional array object

* sophisticated (broadcasting) functions
* tools for integrating C/C++ and Fortran code
* useful linear algebra, Fourier transform, and random number capabilities

**Sckit learn**

Scikit-learn is a free software machine learning library for the Python programming language.[[3]](https://en.wikipedia.org/wiki/Scikit-learn#cite_note-jmlr-3) It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy

**Tweepy**

Tweepy is an open source Python package that gives you a very convenient way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter’s models and API endpoints, and it transparently handles various implementation details, such as:

* Data encoding and decoding
* HTTP requests
* Results pagination
* OAuth authentication
* Rate limits
* Streams

If you weren’t using Tweepy, then you would have to deal with low-level details having to do with HTTP requests, data serialization, authentication, and rate limits. This could be time consuming and prone to error. Instead, thanks to Tweepy, you can focus on the functionality you want to build. Almost all the functionality provided by Twitter API can be used through Tweepy. The only current limitation, as of version 3.7.0, is that Direct Messages don’t work properly due to some recent changes in the Twitter API.

**Python Spell Checker**

Pyspellchecker supports Python3 and Python 2.7 but as always, Python 3 is preferred version. Pyspellchecker allows for the setting of the Levenshtein Distance to check. For longer words, it is highly recommended to use a distance of 1 and not the default 2.

**Natural Language Processing Tool Kit**

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries. Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analysing linguistic structure, and more. The online version of the book has been been updated for Python 3 and NLTK 3.

**Seaborn**

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**Matplot**

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

**A2. ONE CLASS DIAGRAM**

Data Collection

Tweets

Tweet API

Tweet IDs

POS Tagging

Latent Dirichlet Allocation

Latent Semantic Analysis

Sentiment Scores

Bag of Words

TF-IDF

Term Frequency

Feature Engineering

Stop words Removal

Lemmatization and Tokenization

Normalisation

Sentence Splitter

F-1 Score

Recall

Precision

Accuracy

Multilayer Perceptron

Random Forest

Naïve Bayes

Logistic Regression

Inverse Document Frequency

Data Pre-processing

Sentiment Analysis

Topic Modelling

Positive Tweets

Neutral Tweets

Negative Tweets

Machine Learning Approach

Negative Scores

Neutral Scores

Top 8 topics

N\_topics =8

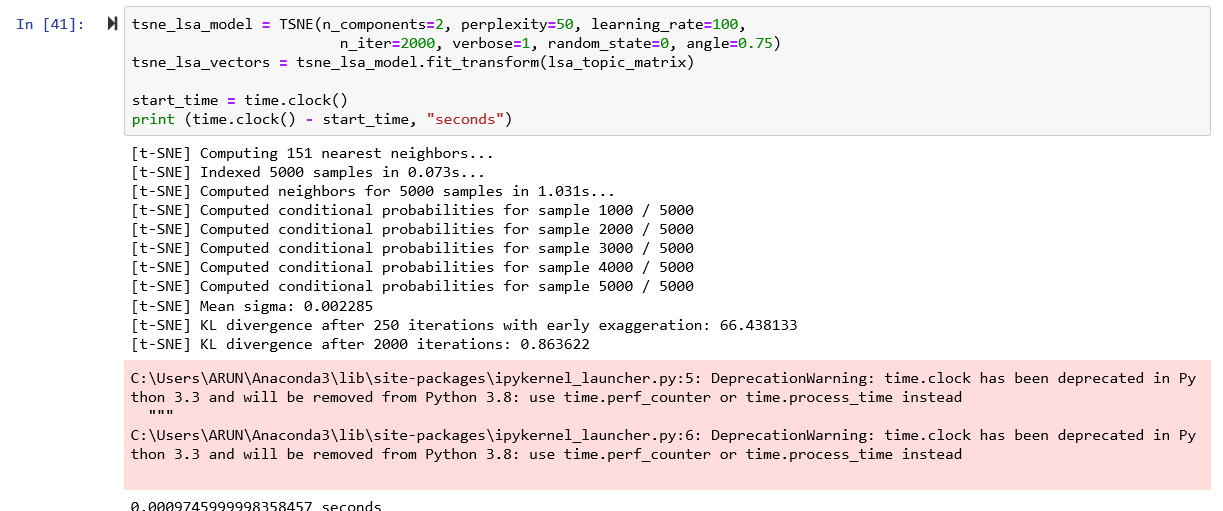
Positive Scores

**A3. SCREENSHOTS OF ALGORITHMS**

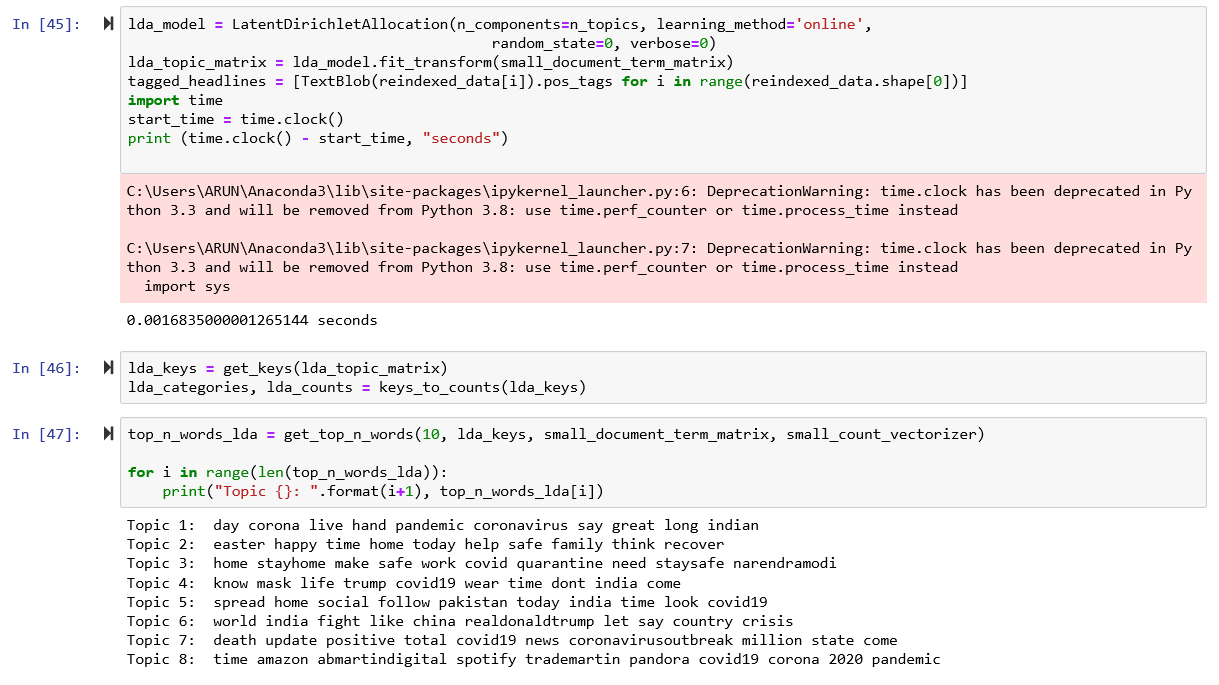
* 1. Screenshot of Data Collection



* 1. Screenshot of LSA Modelling



* 1. Screenshot of LDA Modelling



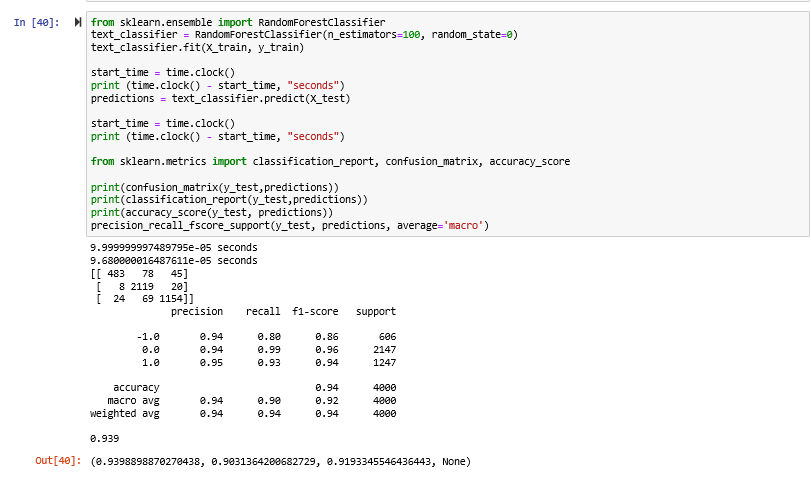
* 1. Screenshot of general Word Cloud



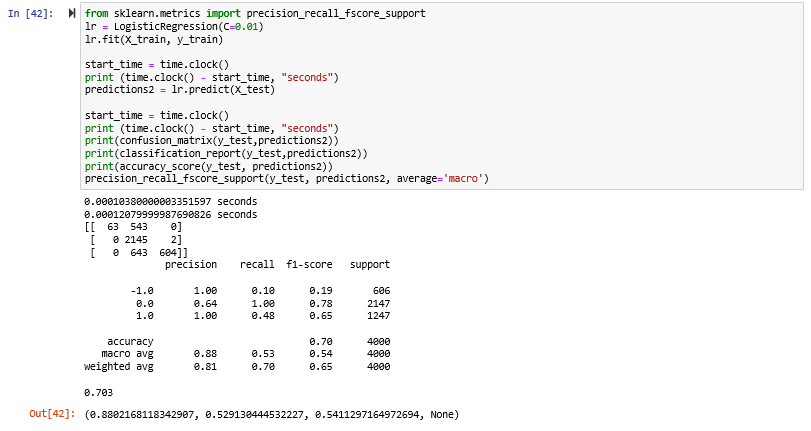
* 1. Screenshot of Positive and Negative Cloud



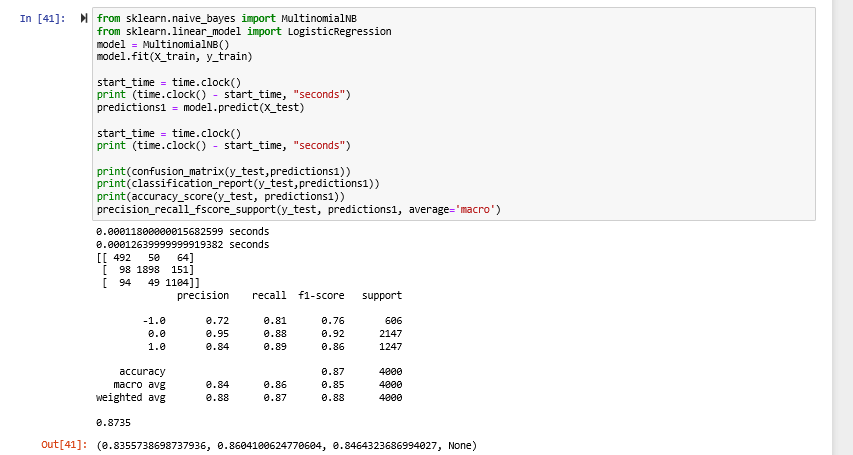
* 1. Screenshot of Machine Learning Model – Random Forest ( High Performing Model)



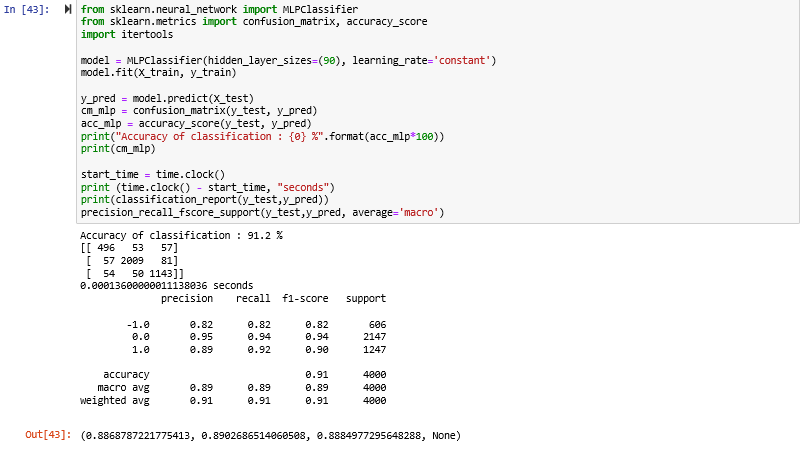
* 1. Screenshot of Machine Learning Model – Logistic Regression



* 1. Screenshot of Machine Learning Model – Naïve Bayes



* 1. Screenshot of Machine Learning Model – Multilayer Perceptron



**A4. LIST OF TABLES AND FIGURES**

1. **LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| S. No. | Description | Page No. |
| 1 | Comparative Analysis of Research papers | 4-5 |
| 2 | Algorithmic Approach | 9 |
| 3 | Confusion Matrix – Logistic Regression | 13 |
| 4 | Confusion Matrix – Naïve Bayes | 13 |
| 5 | Confusion Matrix – Random Forest | 13 |
| 6 | Confusion Matrix – MLP | 13 |
| 7 | Comparison of Accuracy Values | 13-14 |
| 8 | Comparison of Models | 14 |
| 9 | Efficiency Analysis | 15 |
| 10 | Comparison with Related Works | 15 |

1. **LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| S. No. | Description | Page No. |
| 1 | Basic Steps of scientific approach | 5 |
| 2 | Detailed Description of Sentimental Analysis | 6 |
| 3 | Dataset obtained from twitter | 7 |
| 4 | Knowledge discovery steps | 7 |
| 5 | Data pre-processing steps | 7 |
| 6 | Topic Modelling – LDA | 8 |
| 7 | Sentiment Scores of Dataset | 8 |
| 8 | Random Forest Classifier | 9 |
| 9 | Multilayer Perceptron | 9 |
| 10 | Top words ( excluding stop words) | 10 |
| 11 | Headline word lengths | 10 |
| 12 | Part of Speech Tagging for headline corpus | 11 |
| 13 | Word cloud of all tweets | 11 |
| 14 | Word cloud of positive tweets | 11 |
| 15 | Word cloud of negative tweets | 11 |
| 16 | Polarity of tweets | 11 |
| 17 | Work flow of Topic Modelling using LDA | 12 |
| 18 | Topics Modelled | 12 |
| 19 | Graph showing LDA topic counts | 12 |
| 20 | LDA Modelling | 12 |
| 21 | LSA Modelling | 12 |
| 22 | Classification report of Logistic Regression | 13 |
| 23 | Classification report of Multinomial Naïve Bayes | 13 |
| 24 | Classification report of Random Forest | 13 |
| 25 | Classification report of Multilayer Perceptron | 13 |
| 26 | Comparison of Accuracies | 14 |
| 27 | Comparison of Models | 14 |

1. [↑](#footnote-ref-1)