Twitter and Reddit Sentimental Analysis

NAME- MASWOOD AHMAD

REG. NO. – 11911106

SECTION- K19HV

GITHUB LINK- https://github.com/maswoodahmad/int247project

Abstract

With the advancement of net technology and its growth, there's an enormous volume of information gift within the web for web users and loads of data is generated too. web has become a platform for on-line learning, exchanging concepts and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are quickly gaining quality as they permit folks to share and specific their views regarding topics, have discussion with completely different communities, or post messages across the world. There has been lot of labor in the field of sentiment analysis of twitter data. This survey focuses in the main on sentiment analysis of twitter knowledge that is useful to investigate the data within the tweets wherever opinions are extremely unstructured, heterogeneous and are either positive or negative, or neutral in some cases. during this paper, we offer a survey and a comparative analyses of existing techniques for opinion mining like machine learning and lexicon-based approaches, alongside analysis metrics. mistreatment varied machine learning algorithms like Naive Bayes, GHB Entropy, and Support Vector Machine, we provide research on twitter data streams. We have additionally mentioned general challenges and applications of Sentiment Analysis on Twitter.

Introduction

This project is centred around analysing sentiments of text from multiple social media networking sites. each of} the social media analyses tackled here is Twitter, that is that the world' largest micro-blogging web site and it' the place on the web wherever folks express their affairs of state and ideologies additionally to alternative things through short messages referred to as 'tweets'. Not solely do voters use this platform to overtly support political parties however conjointly to specific their opinions on every current affairs and problems happening in the country. Recently it's become a standard ground for politicians and party leaders to convey their messages to the folks of the state and hold campaigns among alternative things. So, analysing the emotions of the voters on this media can paint a image of the political sway of the country.

Literature Review

The paper [1] (Kaur, 2015) describe regarding geographic place flood data set gathered from twitter and understand the opinion of people. They used Naive Bayes method for the class of data and end result they were given 67% accuracy. They gathered numerous resolution from the people which might be beneficial for every authorities and non-authorities business

enterprise to address such state of affairs in an rather better manner. These techniques easier than lexicon-primarily based totally method.

The paper [2] (Paul, 2017) describe concerning the final fit of Indian most efficient league game occasion 2015. Objective of this paper to analyze standardity of IPL fit and that participant are famous and that group is dominate. They want used Hadoop and Map reduce back synthetic language. They were given end result like MS Dhoni is maximum talked concerning participant and town Indians group fairly dominated. This approach gave better end result.

The paper [3] (Mittal, 2016) describe the requirement and effect of the sentiment evaluation on on line platform. They want moreover bestowed a list of sentiments of emotions, interjections and feedback which might be extracted from posts and status updates. They want were given end result to knowing whether or not or now no longer {the on line the internet the net} opinions and posts are being beneficial to purchaser or now no longer and that on line web sites being maximum famous through the purchasers.
The paper [4] (Anto, 2016) describe the merchandise score mistreatment sentiment evaluation. In selling of any product the manufacturer can get the right end result from the purchaser remarks. After were given remarks they will modifications to his product consistent with the remarks. Some customers constantly fail to deliver their feedbacks. Objective of this paper is to avoid the trouble of supplying feedbacks and deliver the approach which would possibly offer
automated remarks on the basis of records amassed from twitter. They used the approach SVM and were given end result 80th accuracy. This machine offer brief and treasured remarks.

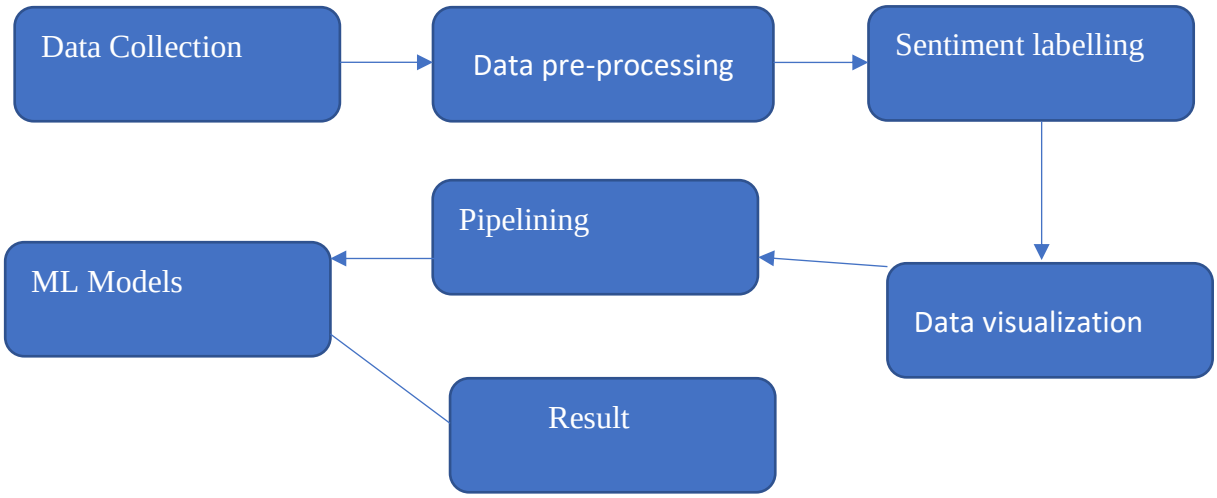
This paper [5](Mamgain, 2016) describe relating to the sentiment analysis of people's opinions relating to high schools in India. they have pictured comparison between the result obtained by the next machine learning algorithms: Naive Thomas Bayes and SVM and Artificial Neural Network model: Multilayer Perception. Naive Bayes outperforms SVM for the aim of matter polarity classification that's fascinating as a results of the model used by Naive Bayes is simple (use of freelance probabilities) and so the probability estimates created by such a model are of caliber. Yet, the classification picks created by the Naive Thomas Bayes model portray an honest accuracy as a results of whenever a decision with the higher likelihood is that being created.

Steps:
1. Scrapping data from twitter and reddit based on keyword search(which is available on https://www.kaggle.com/datasets/cosmos98/twitter-and-reddit-sentimental-analysis-dataset/code)
2. Analyse the sentiments of the texts from the data collected.
3. Gain useful knowledge after processing the collected data.
4. Compare contemporary text classification machine learning algorithms and justification

Objectives of Project

- ❖ To conduct literature survey on sentiment analysis using Spark and python.
- ❖ To arrive at requirement specification for sentiment analysis using multisource social-media data.
- ❖ To design and develop a sentiment analysis model for Indian general election 2019 using PySpark.
- ❖ To implement and compare contemporary machine learning techniques for sentiment analysis.
- ❖ To test and validate the developed sentiment analysis techniques.
- ❖ To document the report by unifying all the results and outcomes.

Flow of program



Data Pre-processing

Pre-processing the data is that the method of cleanup and getting ready the text for classification. on-line texts contain typically ample noise and uninformative components resembling hypertext markup language tags, scripts and advertisements. In addition, on words level, many words within the text don't have an effect on the overall orientation of it. Keeping those words makes the spatiality of the matter high and therefore the classification harder since every word in the text is treated jointly dimension. Pandas library permits several performs to use on pandas information frame. therefore representing in pandas data frame is economical for pre-processing. Duplicate tweets are dropped by using drop duplicates function of pandas library.

Sentiment Labelling

Before developing any model for predicting sentiment knowledge should be tagged this method is termed sentiment labelling. when this datasets is split into plaything and check set.To label the datasets python library TextBlob is used. TextBlob may be a Python library for process matter data. It provides a straightforward API for diving into common tongue processing (NLP) tasks equivalent to part-of-speech tagging, phrase extraction, sentiment analysis, classification, translation, and more.

The sentiment property returns a named tuple of the form Sentiment (polarity, subjectivity). Polarity is a float value within the range [-1.0 to 1.0] where 0 indicates neutral, +1 indicates a very positive sentiment and -1 represents a very negative sentiment. Subjectivity is a float value within the range [0.0 to 1.0] where 0.0 is very objective and 1.0 is very subjective. Subjective sentence expresses some personal feelings, views, beliefs, opinions, and speculations whereas Objective sentences are factual. TextBlob goes along finding words and

phrases it can assign polarity and subjectivity to, and it averages them all together for longer text. Data are labelled based on polarity values. All polarity values which is greater than 1 are classified into positive sentiment that is +1 and values less than 0 classified into negative sentiment that is -1 and polarity value with zero are returned as neutral sentiment that is 0.

Data-Visualization Techniques:

Diagrams, charts, graphs Python libraries used for Data visualization – Matplotlib, seaborn and folium. The extracted tweets are stored in a csv file and is read as df, 'numpy' Is used as 'np' , a package of python for general purpose array processing. 'Max' is a method of numpy that returns the maximum of array.

Pipelining

MLlib standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline, or workflow. The ML Pipelines is a High-Level API for MLlib that lives under the spark.ml package. A pipeline consists of a sequence of stages.

Pipeline working Flow

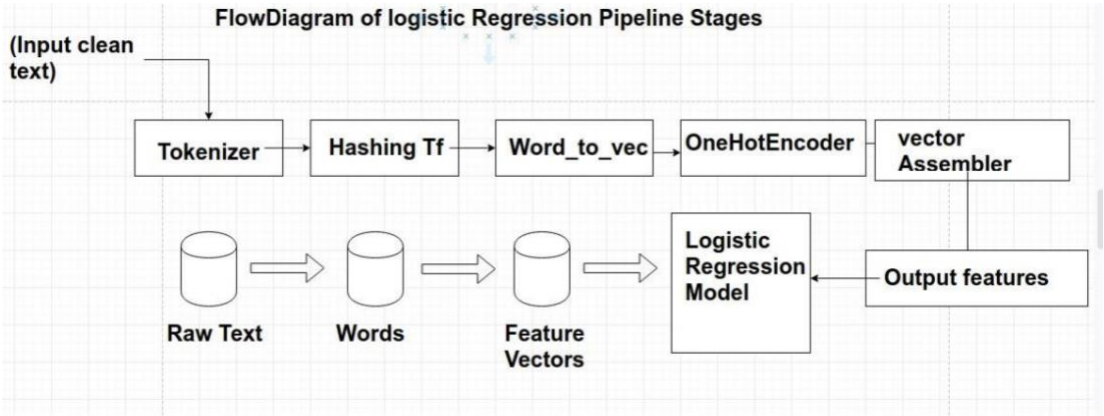
In machine learning, it is common to run a sequence of algorithms to process and learn from data. E.g., a simple text document processing workflow might include several stages:

- Split each document's text into words.
- Convert each document's words into a numerical feature vector.
- Learn a prediction model using the feature vectors and labels

MLlib represents such a workflow as a Pipeline, which consists of a sequence of Pipeline Stages (Transformers and Estimators) to be run in a specific order.

Flow Diagram

A Pipeline is specified as a sequence of stages, and each stage is either Transformer or an Estimator. These stages are run in order, and the input Data Frame is transformed as it passes through each stage. For Transformer stages, the transform () method is called on the Data Frame. For Estimator stages, the fit() method is called to produce a Transformer (which becomes part of the Pipeline Model, or fitted Pipeline), and that Transformer's transform() method is called on the Data Frame.



Machine Learning Models

1. **Logistic Regression**

Logistic regression is a popular method to predict a categorical response. It is a special case of Generalized Linear models that predicts the probability of the outcomes. In spark.ml logistic regression can be used to predict a binary outcome by using binomial logistic regression, or it can be used to predict a multiclass outcome by using multinomial logistic regression which can be changed by Using the family parameter to select between these two algorithms, or leave it unset and Spark will infer the correct variant.

Logistic Regression for Multiclass Classification: Multiclass classification is supported via multinomial logistic (SoftMax) regression. In multinomial logistic regression, the algorithm produces K sets of coefficients, or a matrix of dimension K×J where K is the number of outcome classes and J is the number of features. If the algorithm is fit with an intercept term, then a length K vector of intercepts is available. The conditional probabilities of the outcome classes k∈1,2,…,K are modelled using the SoftMax function.

$$P(Y = k|\mathbf{X}, \boldsymbol{\beta}_k, \beta_{0k}) = \frac{e^{\boldsymbol{\beta}_k \cdot \mathbf{X} + \beta_{0k}}}{\sum_{k'=0}^{K-1} e^{\boldsymbol{\beta}_{k'} \cdot \mathbf{X} + \beta_{0k'}}}$$

We minimize the weighted negative log-likelihood, using a multinomial response model, with elastic-net penalty to control for overfitting.

$$\min_{\boldsymbol{\beta}, \beta_0} - \left[\sum_{i=1}^{L} w_i \cdot \log P(Y = y_i|\mathbf{x}_i) \right] + \lambda \left[\frac{1}{2}(1-\alpha) ||\boldsymbol{\beta}||_2^2 + \alpha ||\boldsymbol{\beta}||_1 \right]$$

When using multinomial logistic regression, one category of the dependent variable is chosen as the reference category. Separate odds ratios are determined for all independent variables for each category of the dependent variable with the exception of the reference category, which is omitted from the analysis. The exponential beta coefficient represents the change in the odds of the dependent variable being in a particular category vis-a-vis the reference category, associated with a one-unit change of the corresponding independent variable.

Random Forest

Random forests are ensembles of decision trees. Random forests are one of the most successful machine learning models for classification and regression. They combine many decision trees in order to reduce the risk of overfitting. Like decision trees, random forests handle categorical features, extend to the multiclass classification setting, do not require feature scaling, and are able to capture non-linearities and feature interactions. spark.mllib supports random forests for binary and multiclass classification and for regression, using both continuous and categorical features. spark.mllib implements random forests using the existing decision tree implementation. Please see the decision tree guide for more information on trees.

The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots$$

where the final model g is the sum of simple base models fi. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensemble. In random forests, all the base models are constructed independently using a different subsample of the data.

Naïve Bayes

Naive Bayes is a simple multiclass classification algorithm with the assumption of independence between every pair of features. Naive Bayes can be trained very efficiently. Within a single pass to the training data, it computes the conditional probability distribution of each feature given label, and then it applies Bayes' theorem to compute the conditional probability distribution of label given an observation and use it for prediction. spark.mllib supports multinomial naive Bayes and Bernoulli naive Bayes. These models are typically used for document classification. Within that context, each observation is a document and each feature represent a term whose value is the frequency of the term (in multinomial naive Bayes) or a zero or one indicating whether the term was found in the document (in Bernoulli naive Bayes). Feature values must be nonnegative. The model type is selected with an optional parameter "multinomial" or "Bernoulli" with "multinomial" as the default. Additive smoothing can be used by setting the parameter λ (default to 1.0). For document classification, the input feature vectors are usually sparse, and sparse vectors should be supplied as input to take advantage of sparsity. Since the training data is only used once, it is not necessary to cache it. An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

Principle of Naive Bayes Classifier:

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

One-Vs-Rest Classifier

One-vs-the-rest (OvR) multiclass/multilabel strategy Also known as one-vs-all, this strategy consists in fitting one classifier per class. For each classifier, the class is fitted against all the other classes. In addition to its computational efficiency (only n classes classifiers are needed), one advantage of this approach is its interpretability. Since each class is represented by one and one classifier only, it is possible to gain knowledge about the class by inspecting its corresponding classifier. This is the most commonly used strategy for multiclass classification and is a fair default choice.

One-Vs-Rest is an example of a machine learning reduction for performing multiclass classification given a base classifier that can perform binary classification efficiently. It is also known as “One-vs-All.” One-Vs-Rest is implemented as an Estimator. For the base classifier it takes instances of Classifier and creates a binary classification problem for each of the k classes. The classifier for class i is trained to predict whether the label is i or not, distinguishing class i from all other classes. Predictions are done by evaluating each binary classifier and the index of the most confident classifier is output as label. In pseudocode, the training algorithm for an OvA learner constructed from a binary classification learner L is as follows:

Inputs:

- L, a learner (training algorithm for binary classifiers)
- samples X
- labels y where yi ∈ {1, … K} is the label for the sample Xi

Outputs:

- a list of classifiers fk for k ∈ {1, …, K}

Procedure:

- For each k in {1, …, K}
  1. Construct a new label vector z where zi = yi if yi = k and zi = 0 otherwise
  2. Apply L to X, z to obtain fk

Making decisions means applying all classifiers to an unseen sample x and predicting the label k for which the corresponding classifier reports the highest confidence score:

$$\hat{y} = \underset{k \in \{1 \ldots K\}}{\operatorname{argmax}} f_k(x)$$

Although this strategy is popular, it is a heuristic that suffers from several problems. Firstly, the scale of the confidence values may differ between the binary classifiers. Second, even if the class distribution is balanced in the training set, the binary classification learners see unbalanced distributions because typically the set of negatives they see is much larger than the set of positives.

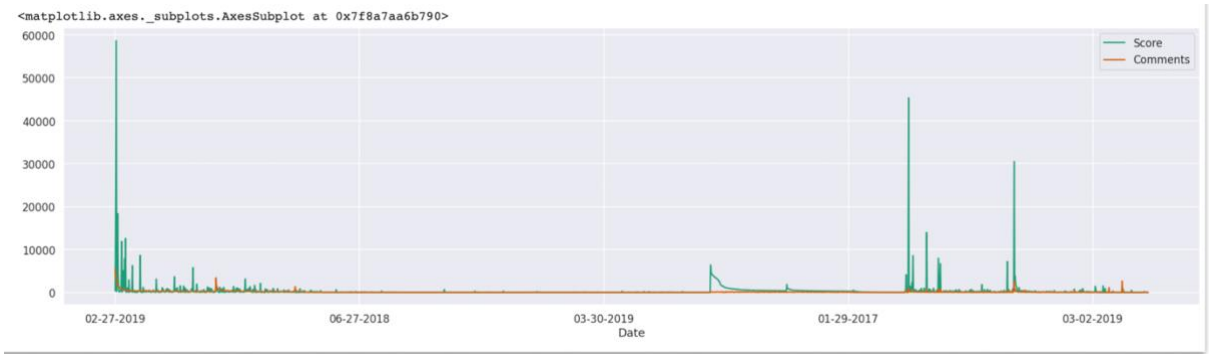
DATA VISIUALIZATION RESULT:

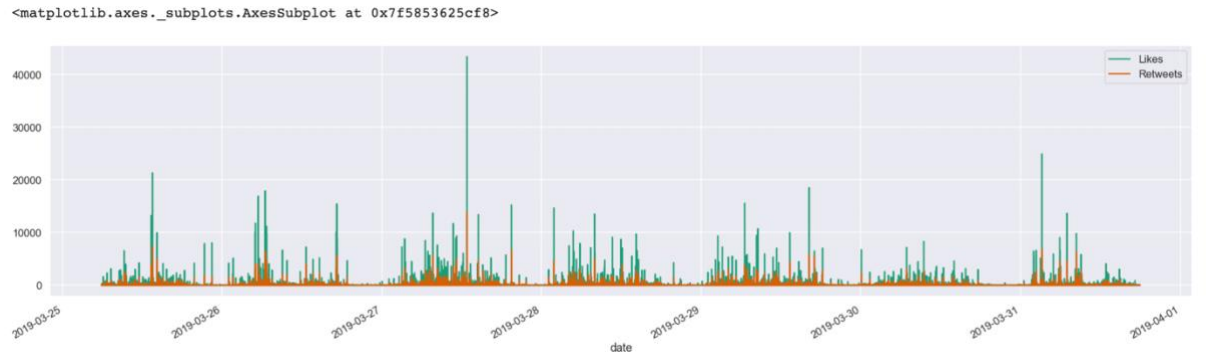
```
The tweet with more likes is:
Indian Prime Minister Narendra Modi told German Chancellor Angela Merkel in talks in Berlin on Tuesday that India would stay in the Paris climate accord even
Number of likes: 58611

The tweet with more retweets is:
[R] Megathread II: India-Pakistan Border Skirmish
Number of retweets: 5408
```

```
The comment with more score is:
I never understood how it makes sense for people, who speak only one language, to make fun of people who can spe
ak 2 or 3.


EDIT: Thanks for the engagement people. I wanted to add a note regarding jokes. Some people have pointed out tha
t it should be okay to joke about accents and I totally agree; The main issue is not taking people seriously bec
ause of an accent.
This comment by u/zh1K476tt9pq summarized it very well.
https://www.reddit.com/r/worldnews/comments/7s6k5t/trump_used_accent_to_imitate_indias_prime/dt2zmpq/

Number of likes: 10932
```

----Statistical analysis of twitter and Reddit Dataset

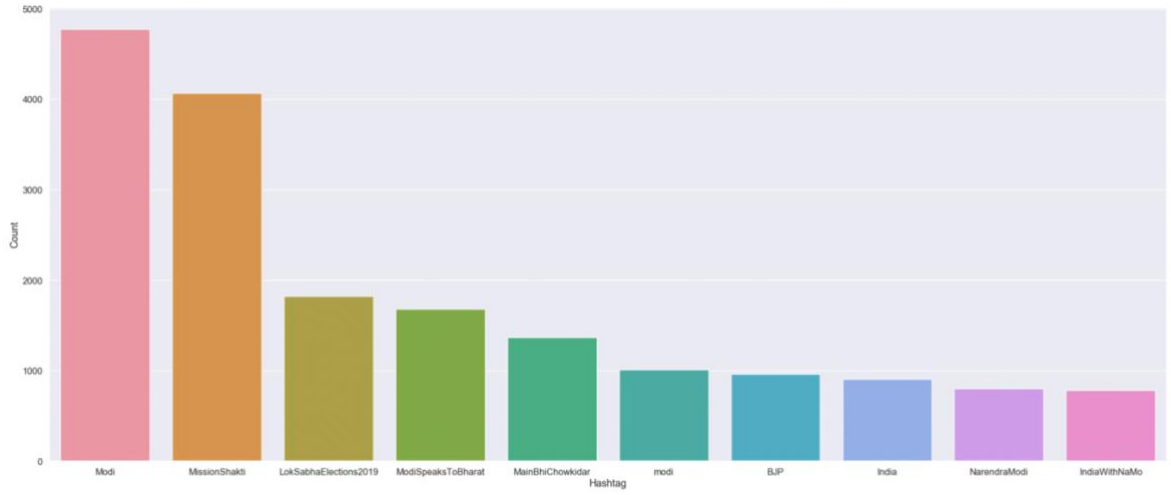
The above fig is the result obtained for a tweet with maximum number of likes and retweets on twitter and comment with maximum score on Reddit.





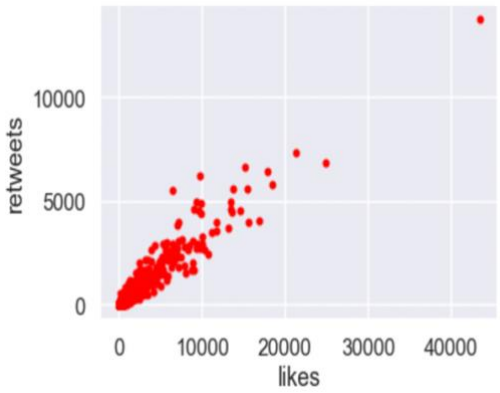
-----Time-series plot for Twitter and Reddit

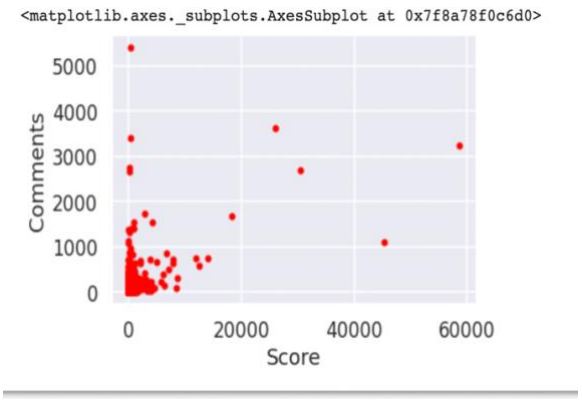
A time series plot obtained shows the correlation between likes and retweets , It is plotted from period ‘26-03-2019’ to ‘31-03-2019’.



------Bar chart that depicts the maximum usage of Hashtags.

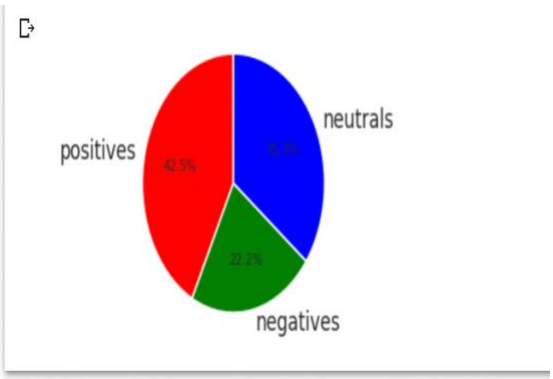
A bar chart shows the Hash tagged word along with-it frequency that are being hash tagged in tweets.

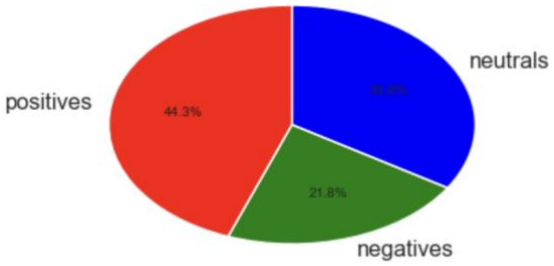




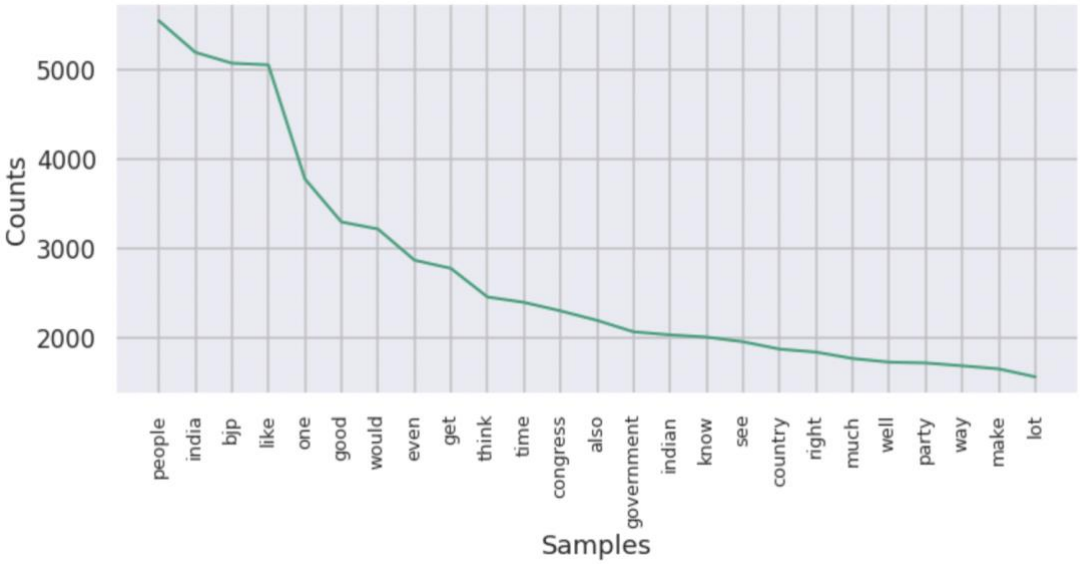
-----Scatter plot

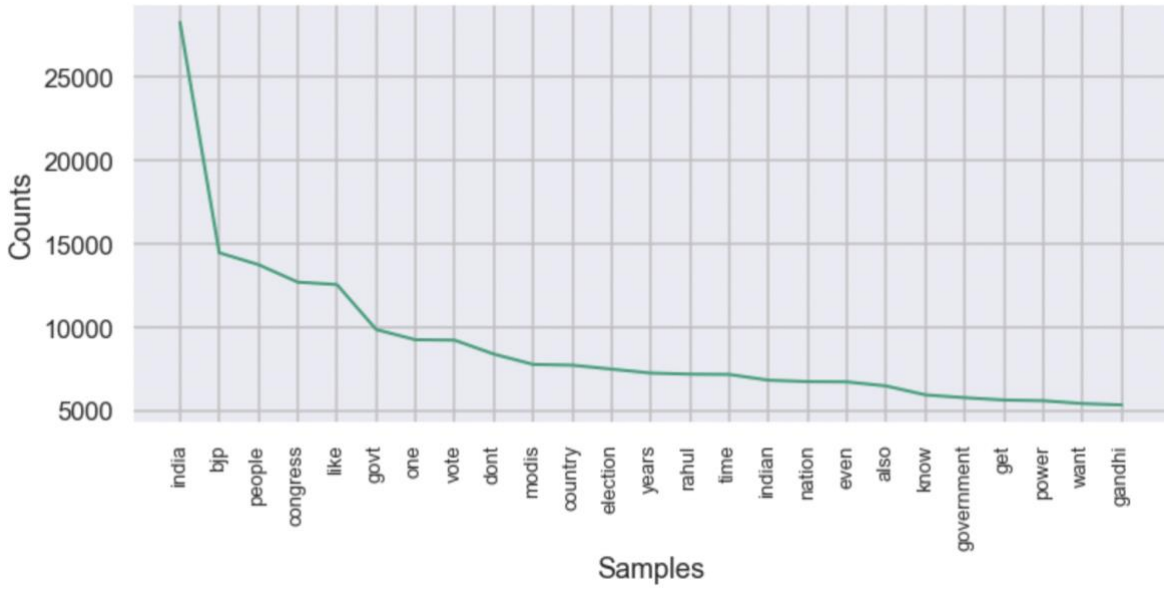
The above scatter plot shows corelation plot between likes a and retweets of the tweet. Notice that as like increases, retweets also increase.

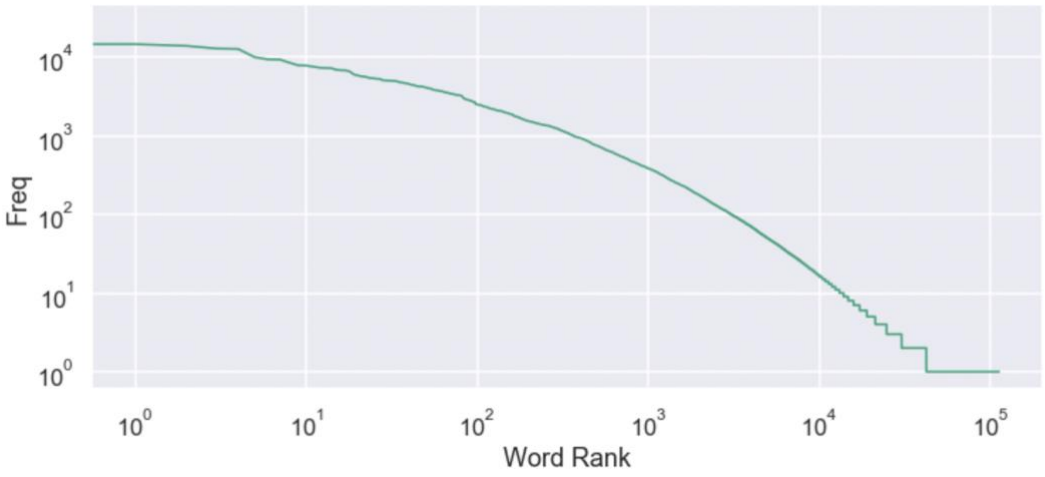


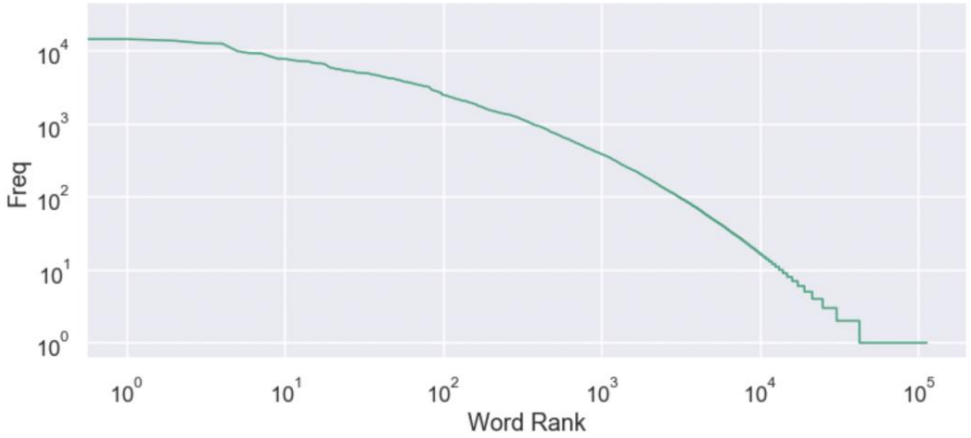


--Pie chart for Categorized data of Twitter and Reddit Dataset


Counts
5000
4000
3000
2000
people
india
bjp
like
one
good
would
even
get
think
time
congress
also
government
indian
know
see
country
right
much
well
party
way
make
lot
Samples

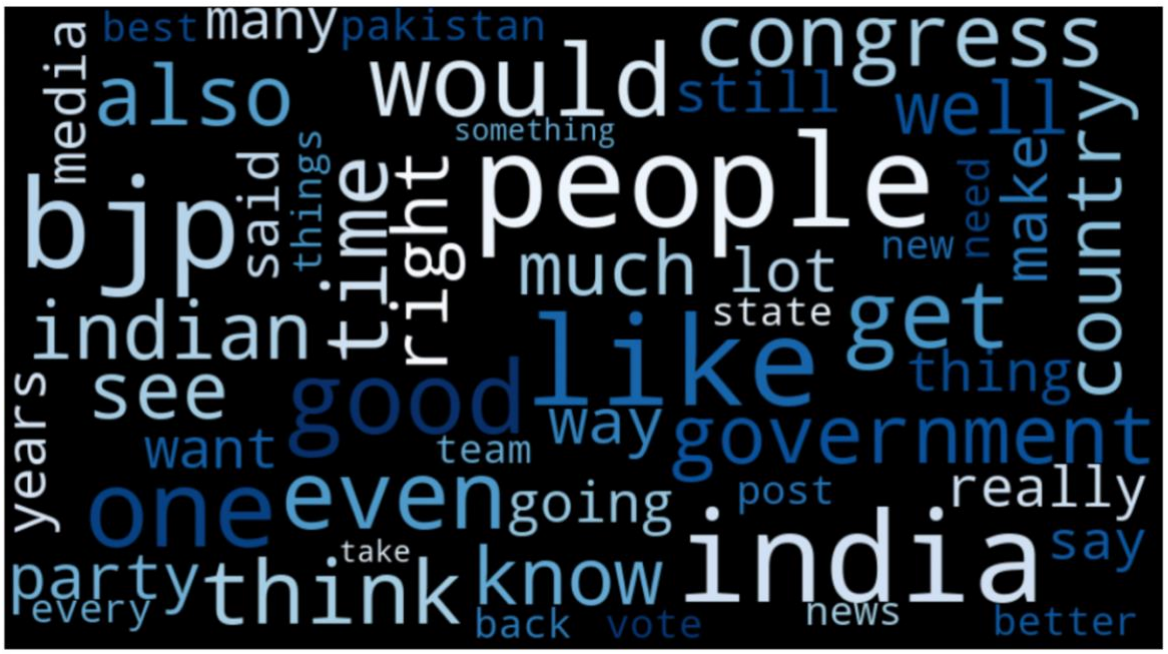

Counts
25000
20000
15000
10000
5000
india
bjp
people
congress
like
govt
one
vote
dont
modis
country
election
years
rahul
time
indian
nation
even
also
know
government
get
power
want
gandhi
Samples



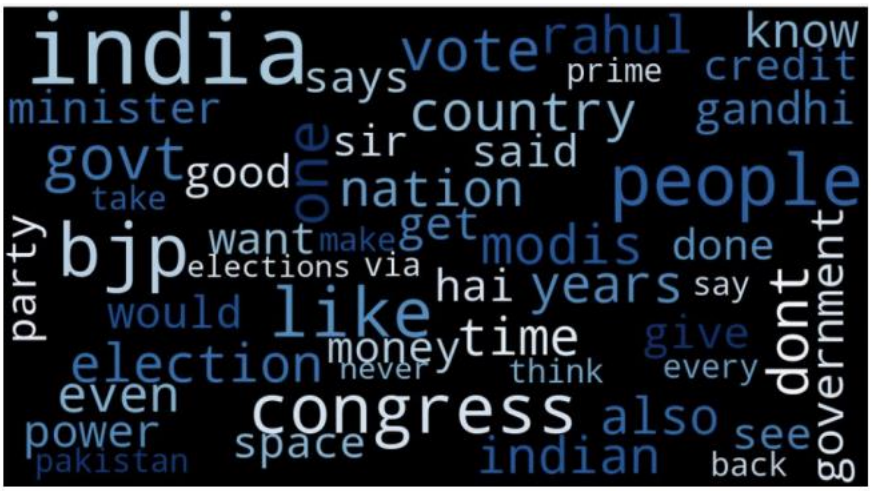


----line plot that shows Word Rank

Line plot that tells the number of times(frequency) a particular word is used in tweets. ‘India’ is used more than 25k times in tweets that are collected. In reddit, ‘People’ is used most and is seen in most of the comments.



----REDDIT



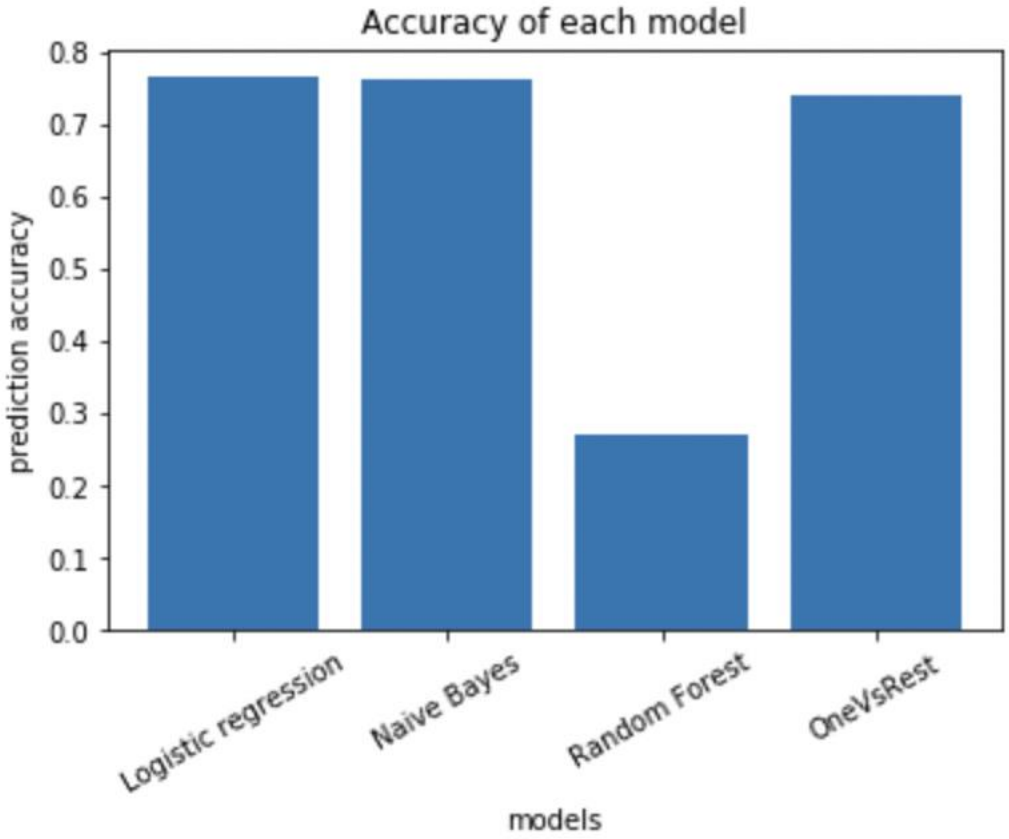
Twitter

The above Word cloud depicts , ‘India’ is used highest number of times in the tweets , and colour intensity indicates how much positively that word is used.

Conclusion :

In this project, we present a large-scale benchmarking experiment based on the datasets containing the tweets and the comments of the people on the 2019 Indian General Elections and finally we compare the prediction performance of the original version of Random Forest

with default parameters and Logistic Regression, Decision Trees and OneVs-Rest classifier as Multiclass classification tools. Comparison of accuracy of all the Machine Learning Models



--- Bar graph for accuracy of sentiment prediction of each trained model

Around 10 lakh tweets and the Reddit Comments have been mixed the use of the pandas Library and all the ones mixed tweets and the feedback turned into in addition saved withinside the Spark records frames after. Machine Learning Models have been Applied on the ones Spark records frames the use of the spark Mllib Library and the accuracy of every Model turned into Predicted as Shown withinside the above figure. Logistic Regression and Naïve Bayes Machine Learning algorithms finished higher than One-Vs-Rest and the Random Forest in line with the taken into consideration accuracy measured in about 80% of the datasets.

REFERENCES

[1]. Anto, M. P. (2016). PRODUCT RATING USING SENTIMENT ANALYSIS. IEEE, pp. 3458-3462.

[2]. Kaur, H. J. (2015). Sentiment Analysis from Social Media in Crisis Situations. IEEE, (pp. 251-256).

[3]. Mamgain, N. M. (2016). Sentiment Analysis of Top Colleges in India Using Twitter Data. IEEE, pp. 525-530.
[4]. Mittal, S. A. (2016). Sentiment Analysis of E-Commerce and Social Networking Sites. IEEE, pp. 2300-2305.
[5]. Shahare, F. F. (2017). Sentiment Analysis for the News Data Based on the social Media. IEEE, pp. 1365-1370.