

**“Analysis of Different Social Media Networks to Identify Best Fit for a Task”**

**Submitted by –**

Pratishtha Gaur – 19BCE0104

Abhinav Sudhir Chawla –19BCE0343

Akshat Rastogi – 19BCE2220

**Under the guidance of –**

Prof. Jaya Subalakshmi K.

**Course**: Social and Information Network (CSE3021)

**Slot** : E1

**Date**: 28-4-2022

Table Of Contents

Abstract

Introduction

● Overview

● Motivation

● Scope

● Application

Literature Survey

System Architecture

● Existing

● Proposed

● Diagram

● Hardware requirement

● Software requirement

Modules

Implementation

Results and Analysis

Conclusion

Future Work  
References  
Appendix-A(Code snippets)

1. Abstract

We all are aware about this rush of social media platforms. People's urge to share their thoughts and learn about others gain knowledge from public forums like Twitter, Youtube, Facebook, and Instagram. Social media platforms like Twitter and YouTube have attracted a wide community opening it to many possibilities and scope like education, marketing, entertainment, business and campaigns. In this study we try to compare these social media platforms to decide which genre of works are more popular in each platform considering a wide niche of Indian Election. The study identifies which platform reaches out to a greater number of people of this specific sect thus explaining where to put in more resources for an election campaign.

2. Introduction

2.1 Overview

Politicians and political groups frequently utilize social media to promote themselves. Many people voiced their opinions on various social media platforms, including Twitter, a major micro blogging service. Each micro-blog is referred to as a "Tweet," and it can be no longer than 140 characters long. Many tweets also include a label for other Twitter users they're mentioning, as well as a "hashtag" that commonly denotes the tweet's topic. Tweets are essential because they convey the users' feelings. A hashtag allows Twitter users to express their feelings about a candidate. Because the election is dominated by two parties, a hashtag used to show positive sentiment for one candidate may be used to represent negative sentiment for the other, and vice versa. The Indian National Congress and the Bharatiya Janata Party are the two main parties we will consider. We use hashtags to retrieve tweets for our job. The hashtags will link the tweets to a specific political party or leader. Both sides' tweets will be gathered and categorized as favorable or negative based on the sentiment score. YouTube in addition to being the most used video platform in the world, has become an essential tool for the communication and advertising strategy of brands, since on YouTube they can generate really relevant audiovisual content with the possibility of being interactive, commented and shared. These comments, likes and shares signify popularity of a video. This information can be extracted using YouTube Data V3 api or using Selenium or ChromeDriver libraries.

The content posted by the users can represent their political belief or is either used to comment, be sarcastic or express negative opinion towards a political party or ideology. Twitter and YouTube, while constituting two of the most popular online social networks attracting millions of users daily, capture a very important proportion of this online discourse. Sentiment in these analyses is represented by a variable with values like Positive, Negative and Neutral. Each word in the corpus can be assigned with more than one sentiment. Most Neutral comments have to be removed since they don't help in the training process. In this study we obtain the most popular hashtags around the Indian elections and gather a dataset of 10000 tweets, for a period of three months. We extract 10 unique YouTube videos contained in this dataset as well as their metadata (likes, comments, authors etc.). Initially, we perform a volume analysis and an association of the diverse features of the YouTube videos. The models built are a Bidirectional Recurrent Neural Network and Glove word embedding model.

**2.2** **Motivation**

Different social media platforms have different impact factors. We understand that communities on YouTube and Twitter have some similar and some dissimilar interests which makes them different in its own ways. But today both these platforms are used for all kinds of information sharing and is not restricted by any niche. Still, it has been observed that some form of information sharing is more inclined towards one platform than the other. Here we compare these social media platforms as to decide which genre of works are more popular in each platform with an example of Indian Election sentiment analysis. This study might help people to understand which platform is better for a campaign like this and where to put in all the resources. This study will also help people to know which social media platform has a higher negativity rate on certain kinds of topics, this way one can decide which social media platform suits them the best. Also, both YouTube and Twitter have their own freely available datasets for people to explore its trends and analyze.

**2.3** **Scope**

During the course of the project, we obtain data specific to Indian elections and analyze them to obtain that Twitter is a better suited platform for election campaigns than YouTube is. Scope of project is not limited to Election but is far wide. It is about understanding which social network is better in a certain genre. For example, in the case of advertisements it will reach more people through YouTube but a freedom campaign will reach more people through Twitter. Distinguishing the fine line of difference between the two networks and their difference.

**2.4** **Application**

Applications of social sentiment analysis are varied. It is used to identify the audience's pain points. With sentiment analysis, you can delve deeper into all of your content's interactions. You may track the general sentiment connected with a brand using social media sentiment analysis. Having a presence on social media isn't going to cut it for your business. You must also be aware of user sentiment. The most typical application of sentiment analysis in diverse markets is brand monitoring and reputation management. It's no surprise that understanding how people view your brand/product/service is beneficial to tech companies, marketing agencies, fashion brands, media outlets, and a variety of other businesses. Text analysis can even be extended to the emojis which are used by a viewer as a content conveys a lot about how they feel about your posts. This is the best insight to help you understand audiences’ psyche. Apart from the above-mentioned data elements, the Unbox Social tool gives you deep insights on post-performance and audience. The tool is well-equipped to fulfill your sentiment analysis requirements. ·

**Literature Survey**

**1.** **Emotional community detection in social networks**

This paper addresses the need for an efficient and innovative methodology for community detection that also leverages users’ behavior on an emotional level.Thus, the inferred networks are assessed with the utilization of three distinct measurements, while the weighted version of a modularity community detection algorithm is used. There is significant proof demonstrating that the proposed procedure makes persuasive enough community.

**Pros**

**1.** The technology used is scalable

**2.** It can be further used for advertising and marketing.

**Cons**

Dependency on external tools is prominent for sentiment analysis.

**2. Talking politics on Twitter**

The paper assesses relational power granted to candidates through Twitter conversations about them and whether they change depending on the gender of their opponent. This paper discusses about the studies that are gathered in three effective classes: studies tending to the utilization of Twitter by lawmakers and missions; studie tending to the utilization of Twitter by different publics during political decision and issue missions; and remarks on Twitter during campaign occasions - - like broadcast discusses, party shows, and final voting day inclusion.

**Pros**

Dataset used is large and hence more reliable.

**Cons**

One of the important con of this paper is that the demographics such as age etc are not considered.

**3.** **Solving Community Detection in Social Networks: A comprehensive Study**

In this day and age, social media platforms, for example, Facebook ,Instagram, LinkedIn interfaces different clients framing a socialnetwork graph. In these social media graphs, recognizing networks is an extremely fundamental task as networks help us in gathering clients showing comparable ways of behaving and in this manner the social network can be partitioned into various bunches of nodes with the same behavior.

This paper conducted a survey on different methods applied to the problem of community detection.This paper summarizes and compares all the techniques by classifying them into broad domains.

**Pros**

1. Tackles the problem of overlapping communities.

2. Algorithms used are scalable and robust for real world applications

**Cons**

The methods surveyed here did not include the factor of emotion in community Detection.

**4. A Parallel graph partitioning approach to enhance community detection in social Networks.**

This paper proposes a new parallel partitioning algorithm that focuses on assisting in community detection in social networks. Managing complex organizations is many times a test because of the great computational expense in breaking down a gigantic measure of information.

Partitioning methods can diminish the intricacy of huge designs by decreasing them to more modest, less connected parts. Likewise, data splitting permits the utilization of multiprocessing to speed up the execution of information strategies with synchronization and parallelism. In this paper, we propose another equal apportioning calculation with an attention on aiding community detection in social networks.

**Pros**

This Paper proposes a distributed memory approach and hence execution is carried in parallel manner to reduce execution time.

**Cons**

Memory consumption and complexity is increased.

**5**. **Adaptive Community detection incorporating topology and content in**

**social networks.**

In social network analysis, community detection is a fundamental stage to figure out the design and capacity of organizations. Some traditional community detection strategies might have restricted execution since they simply focus around the organizations' topological structure. Other than topology, content data is one more huge part of social networks. The paper proposes integration of topology as well as content of networks and has an adaptive parameter (with two variations) to effectively control the contribution of content for understanding the network.

**Pros**

Paper Considers comprehensive effect of incorporating both node-induced and edge induced attributes.

**Cons**

The node attributes used in this paper can have one form whereas for real social networks the content may have various forms.

**6**. **Analytical Models of Information-Psychological Impact of Social**

**Information Networks on Users.**

The paper discusses the information-psychology impact of Social Media networks and threats they impose. The article proves the probabilistic model of IPI sway on the clients of the SIN

portraying a way to deal with assurance of the quantity of the most visited sites of SIN and further, the complete likelihood of successful assaults on the most visited destinations of the SIN of the total number of sites - the subjects of IPI.

**Pros**

Probabilistic model is provided to determine the probability of occurrence.

**Cons**

During the review of this paper we do not find any limitation related to the Psychological

Impact of Social Networks on users.

**7**. **A Framework for Analyzing and Detracting Negative Emotional**

**Contagion in Online Social.**

This paper expresses the use of clustering for detecting the community where the negative

emotions may spread.

Online social networks are a strong stage for the spread of negative emotion contagion which is influencing clients according to alternate points of view for example psychology, financial matters, promoting and neuroscience. Online social networks have gigantic measures of data and Knowledge that should be concentrated using data mining techniques. This paper focuses around introducing another system for analyzing and detracting negative emotional contagion using clustering for distinguishing the community where the negative emotions might spread.

**Pros**

This framework can be implemented using high performance programming languages, data mining algorithms.

**Cons**

1. Unprecedented growth and magnitude of networked data generated by online social networks**.**

2. High rate of spam is seen due to the algorithm used.

**8. Opinion Mining and Sentiment Analysis on a Twitter Data Stream**

Option mining and sentiment analysis is a quickly developing subject with different world applications, from surveys to promotion situations. Generally people assemble input from their companions or family members prior to buying a thing, yet today the pattern is to recognize the assessments of a variety of people all over the planet utilizing microblogging information. This paper talks about a methodology that is exposed. a stream of tweets from the Twitter microblogging website are preprocessed and grouped in light of their passionate substance as positive, negative and superfluous; and analyses the performance of different classifying algorithms in view of their precision and recall in such cases.

Further, the paper exemplifies the applications of this research and its limitations.

**Pros**

Limitations of algorithms are expressed explicitly for a case study.

**Cons**

The dataset used is not resampled and hence can be unbalanced to produce overfitting.

**9**. **A Compression-Based Multi-Objective Evolutionary Algorithm for**

**Community Detection in Social Networks.**

Community Detection is a critical perspective for understanding organization designs and uncovers the fundamental capacities or attributes of complex frameworks. A community typically refers to a set of nodes that are densely associated among themselves, yet inadequately associated with the leftover nodes of the organization.

Identifying the community has ended up being a NP-difficult issue. In this way, evolutionary based optimization approaches can be utilized to settle it. In any case, an essential test for them is the higher computational intricacy while managing large scale networks.

This paper basically focuses on dividing the whole network into subgraphs requiring that nodes inside the subgraphs are tightly connected.

**Pros**

1. The proposed algorithm is for both undirected and unweighted graph.

2. The proposed algorithm is an optimized version.

**Cons**

During the review of this paper we do not find any such limitation present in this paperas all the areas are perfectly covered in this paper.

**10. Sentiment Analysis on Twitter using Ordinal Regression**

Lately, research on Twitter sentiment analysis, which breaks down Twitter information (tweets) to separate client feelings about a subject, has developed quickly. Numerous specialists incline toward the utilization of AI calculations for such examinations. This study intends to play out a definite feeling investigation of tweets in light of ordinal regression using machine learning strategies. The proposed approach comprises first pre-processing tweets and utilizing a component extraction technique that makes a proficient feature.

This research aims to apply sentiment analysis on Twitter data by Ordinal Regression. It uses Random Forest (RF), Support Vector Machine (SVM), and Multinomial Logistic Regression (MLR) classifiers to classify the tweet sentiment into five categories.

**Pros**

Accuracy is very high and the model outperforms all other similar models.

**Cons**

Small dataset is used and hence sampling is not reliable.

**4. System Architecture**

**Existing**

**Proposed**

The Proposed Model involves sentiment analysis of Twitter threads and Youtube comments over the scenario of Indian elections and comparing similarities and dissimilarities between the two social networks. Data is obtained using tweepy API of Twitter by extracting comments based on hashtag and keywords, For YouTube, YouTube Data v3 API and web scrapping is used to gather the necessary information like, number of replies, comments, number of like, statistics and more. The data collected is later passed on for preprocessing inorder to reduce noise by removing unnecessary and vague data. Preprocessed data is clean and ready as a single set but for training it is important to have a balanced number of classes to give an unbiased result of the query. So we will use processed data to check for proportion of positive or negative comments or number of comments for BJP and for Congress later based on that resample and create a new balanced dataset. Data flows to exploratory data analysis stage next where trends are analyzed and preprocessing is checked for using visualization. For example, creating a word cloud to know if genuine words appear bigger or not, or plotting a count plot of polarity to understand positivity and negativity rates. For the training phase the approach proposed involves training using two different models and using the best trained model for use. The concept of voting is applied to the process of selection of the model which will later be used for prediction. For training initially data is randomly divided into training and testing groups with a ratio of 4:1. Later two selected models which are LSTM and GloVe models fit across the dataset. Parameters of training were chosen by trial and analysis. Later an epoch run of 100, activation function as softmax and loss function as categorical cross entropy gave the required result. The result needs to be checked and analyzed for both training and testing data but these parameters showed overfitting so a regularization function was added to this function basically adding more penalty at cost function hence keeping the result more general than specific. Based on positivity rate, polarity, negativity rate and subjectivity produce a conclusion as to which platform influences the most or is used to spread hate during election times, and which platform instead should be used for campaigning for election and to reach genuine people.

EDA MODELING, PREDICTING

**Diagram:**

**Hardware Specification**

1. Training of models takes a lot of computational energy that may require a higher version of the hardware system, in that case model development done using Google collab platform provides more GPU functionality
2. While working with trained models is easy for hardware resources as for command line interface required hardware is basic CPU like intel i3 very basic requirement even with no GPU configuration.
3. Secure Internet connection

**Software Specification**

1. Python Libraries:
2. NLTK
3. Sklearn
4. Keras for LSTM
5. Selenium
6. Google colab

**5. Modules**:

1. Problem Definition

Social media is a source where people put in their thoughts about varying fields, it is used by small businesses for expansion and even for advertising and marketing for products, election campaigns, cleanliness drives, people groups and more. For our analysis this time we chose ‘Indian Election Campaign’ as our topic considering that a large set of people in these social networks are involved in this sec so we cover a bigger community in the process of study of different social media platforms as per their reachability.

2. Data Collection

We obtained data for analysis using two very prominent and in light social networks twitter and youtube. Twitter has created a free API called Tweepy, it can be used to retrieve tweets under a specific hashtag, find locally trending subjects, create bots, send a tweet, edit tweet. Python has a library called tweepy which we used to extract tweet content its hashtags, author, user mentions, retweets and like count for better analysis. Similarly, YouTube has an API called Youtube Data V3 lets you fetch search results and to retrieve, insert, update, and delete resources like videos or playlists. All the extracted content is stored in the form of a csv file. The limit of tweets extracted was kept in correct proportion to the number of youtube comments extracted to ensure right balance while training. Furthermore, comments of all classes are kept as equal, for example number of positive comments with BJP is close to number of positive content to obtain a balanced dataset. Resampling approach was used for the same.

3. Data Cleaning

Cleaning extracted text is a separate branch of study and choosing which texts to leave out of data processing step to avoid false positives or negatives is necessary. Some options to consider could be removing or elongating abbreviations, removing stop words like etc, where, when, what (NTLK library contains such 180 stop words which we can handle), and remove extra spaces in case they are counted as words and add to noise. When handling large amount of data there always is an issue to manage noise. Words that are of no use in a sentence or emoticons which don’t clearly explain the context are counterproductive and affects badly during training since such words take up some weight in the process of training which should actually not be given. We during our project removed emoticons from the texts since they are vague and can even convey sarcasm at times. Later we churned punctuations, special characters and numbers since it’s difficult to identify the context of use with present resources.

4. Resampling and Labelling

During sentiment analysis it has been studied that social network has a higher number of negative instances than positive ones. This can cause discrepancy in modelling the analysis. To cater to this issue, we understand the difference in proportions of positive and negative comments and resample the data such to have a balanced dataset. This approach involves randomly dividing or multiplying the set of observations according to their labels to obtain a nearly equal sized dataset.

The tweepy libary gives out sentiment polarity in the range of -1 to +1. For our topic of election prediction the neutral tweets would be of no use as they will not provide any valuable information. Thus for simplicity purpose we have labeled tweets as only positive and negative. Tweets with polarity less than 0 will be labelled negative(0) and greater than 0 will be positive(1). On fetching polarity of BJP and Congress negativity was observed to be very high and since the ratio of the negative tweets to positive tweets is not proportional. Our data set is not balanced. This will create a bias while training the model. To avoid this we have resampled the data. In order to balance the data we needed to add only positive tweets into our original data. From the new data, so we randomly pick the positive tweets and then add them to our main data.

5. Exploratory data analysis

Exploratory Data Analysis (EDA) is the process by which the data analyst becomes acquainted with their data to drive intuition and begin to formulate testable hypotheses. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. One of the most common approaches to visualize data is word cloud that we used. Word cloud is basically a visual representation of the words used in a particular piece of text, with the size of each word indicating its relative frequency. So, while generating a word cloud for BJP comments one will observe words like, BJP, Modi whereas in Congress words like Rahul Gandhi, Amethi will be more frequent. These common words in cloud visualization is appear relatively bigger because higher frequency usage than other words.

6. Modelling

Modelling is the core of the whole process, place where machine learns to classify and analyse what is expected out of it. Choosing appropriate mathematical and computational models at this stage are important. For our problem we performed training using two models. Namely, GloVe - Word Embedding Model and Bidirectional Recurrent Neural Network(RNN) or LSTM.

GloVe is an unsupervised learning algorithm which makes use of the concept of embedding matrix to analyze text. The embedding layer maps the words present in glove dictionary to their corresponding vectors from the embedding matrix and words which didn’t occur in GloVe dictionary are assigned a zero vector. Word Embeddings are basically text converted into numbers. There are number of ways to represent the numeric forms. Types of embeddings include Frequency based and Prediction based. Frequency Based uses a Co-occurrence matrix whereas Prediction-Based uses BOW, Skip-gram model using Pre-trained word vectors Word2vec and Glove. Here word embedding is done for the experiment with the pre trained word vector Glove. Glove version used 100-dimensional GloVe embeddings of 400k words computed on a 2014 dump of English Wikipedia. Training is performed on an aggregated global word-word co-occurrence matrix, giving us a vector space with meaningful substructures

After GloVe LSTM approach is used because it handles long sequence dependencies well like comments of youtube which is basically just one long statement.Bidirectional long-short term memory (bi-lstm) is the method of allowing any neural network to store sequence information in both backwards (future to past) and forwards (ahead to future) directions (past to future) thus preserving both the future and past. The blank area in the line "boys go to..." cannot be filled. Still, when we have a future sentence like "boys come out of school," we can easily anticipate the previous blank space and have our model do the same thing, and bidirectional LSTM allows the neural network to do so. We used pretrained Bi-LSTM with the features as, 100 epochs of training, softmax activation function and categorical crossentropy loss functions.

7. Predicting

As a result of prediction of sentiment analyzer we obtain four parameters, polarity, subjectivity, positive rate and negative rate. These parameters can together give us a clear idea of the situation and thus derive some insights out of it. The TextBlob library is used for the same. It returns polarity and subjectivity of a set of statements. The range of polarity is [-1,1], with -1 indicating a negative sentiment and 1 indicating a positive sentiment. Negative words are used to change the polarity of a sentence. Semantic labels in TextBlob aid in fine-grained analysis also by including emoticons, exclamation marks, and emojis in the process but their informal use in social networks boosts the chances of false positives. Subjectivity rate lies between 0 and 1. It is the degree of personal opinion and factual information in a text. Because of the text's heightened subjectivity, it contains personal opinion rather than factual information. In a given use-case we are dealing with opinions only and not facts since we study the inclination of people by their comments so the subjectivity rate is not of much use.

8. Observation

Interpreting the result in a form from data not percievable by a common person is the function of this module. For example adding context to the numbers or adding meaning to the graph. The result like high polarity and high subjectivity will be presented in an understandable form here for the user.

**Result and Discussion**

Social networks have a very important role in online social discourse and especially during pre-elections period.

The results of comparative sentiment analysis indicates a X% express positive sentiment towards Party1 in Twitter and Y% positive sentiment towards Party2

while A% of users express positive sentiment in YouTube metadata gathered towards Party1 and B% positive sentiment towards Party2.

Our analysis fill the gap between the connection of offline events and their consequences in social media by monitoring important events in real world and measuring public volume and sentiment before and after the event in social media.

**Conclusion**

**Future Work**

For this project only, two social networks were considered but it can be expanded to include more like Facebook, Reddit, Instagram for a more generalized result. To prove that different social platforms are positively inclined towards different niche other sects and categories can be analyzed for example studying reach of advertisements, or reach of educational platforms say Byjus. We can even start by increasing the dataset that we used now since understanding election is not an easy task so estimating based on the limited amount of information won’t be justified. Apart from these we can try different models like Bidirectional RNN with attention mechanism. We can implement BERT which is currently the state of the art for solving various Natural Language Processing problems

**To learn**

**· Embedding layer, embedding matrix – glove**

**· Bidirectional RNN**

**· LSTM**

**· Word cloud**