Detection of Offensive Tweets: A Comparative Study

Niyati Aggrawal

Jaypee Institute of Information Technology, Noida, India

Abstract

With the growing popularity, Twitter has become a major platform for posting views via tweets. Tweets contain useful, relevant and offensive content as well. More than a decade of research has resulted in numerous techniques and models to detect offensive content. However, little is known about lexically offensive and contextual offensive content. In this research paper, lexical offensive contents have been identified using two techniques- Rule-Based Naive Bayes (RNB) and a collaborative model of LDA with Naïve Bayes (LDANB). LDANB provides better results as compared to RNB for lexical offensive tweet detection. Further, contextually offensive contents are detected using newly devised Adjective Based approach. Contextual offensive content results prove to be better with Adjective based approach than Cosine similarity based results. To validate results of applied offensive tweet detection techniques three performance metrics-precision, Accuracy and recall are used.

Keywords: OSN, Machine Learning, Naive Bayes, LDA, Cosine Similarity, Information Filtering, and Adjective based filtering

I. INTRODUCTION

Online Social Network has evolved and gained much of popularity in the past few years [27]. It serves as a medium which has a large reach and can be used by any person residing in any part of the world. It spans across barriers of country, religion, region, race and language. Currently, people are spending more and more time on social media to connect with others, to share a wide variety of information, and to pursue common interests.

One of the major online social micro blogging sites is 'Twitter' which is founded in March 2006 by Jack Dorsey. Since then it has grown exponentially and revolutionized the way people access information and news about current events. Unlike traditional news media, OSN such as Twitter is a bidirectional media, in which common people also have a direct platform to share information and their opinions about the news events. It helps users to connect with other Twitter users around the globe. The messages exchanged via Twitter are referred to as micro-blogs because there is a 140 character limit imposed by Twitter for every tweet. This lets the users present any information with only a few words, optionally followed with a link to a more detailed source of information. Therefore, Twitter messages, called as "tweets" are usually briefed and focused. In such a situation where twitter has become an indispensable part of every common person's life, but it is necessary to filter out indecent or abusive content from the tweets that are being posted on twitter as such tweets that can negatively affect the users especially adolescents.

Furthermore, new press and studies found that children and adolescents were engaged in producing online hate speech (Tynes et al., 2004), 3% of adolescents participated in cyber solicitation in 2008 (Finkelhor et al., 2008), and 13% of adolescents cyber-bullied others in 2010 (Hinduja & Patchin, 2008).

This research work aims to compare various machine learning and natural language processing approaches that can be employed to detect offensive content on twitter on the basis of content as well as context.

To overcome the problems of the existing offensive tweet detection techniques, we have introduced a new technique - adjective based approach for detecting contextually offensive tweets and generated a LDANB model which is a Collaborative Model of LDA and Naive Bayes for lexically offensive content. This work aims at building a comparative analysis model in order to compare different lexical and contextual classifiers. It is intended to propose an optimum approach for classification of twitter data on the basis of contextual and lexical aspect of the tweets. There is a dire need to have optimum strategies for offensive content detection for social media as social media has become one of the most primary modes of communication. Thus, any kind of offensive content transmitted or passed through it may harness its benefits and give rise to various cyber crimes such as cyber bullying. The work has been made in a manner to support scalability and offers new solutions to the limitations present in already existing techniques.

In summary, the contribution of this paper is as follows:

a. The paper presents the concept of offensive tweets and proposes a generic framework to detect offensive tweets based on lexical and contextual contents.

b. Two techniques- Rule Based Naive Bayes (RNB) and LDA with Naive Bayes (LDANB) have been applied for lexical content based offensive tweet detection. A comparative result has been presented which shows performance of LDANB over RNB.

c. One Adjective based approach is proposed for contextual based offensive tweet detection and compared results with basic cosine similarity measure. Adjective based approach able to improve accuracy of contextual based offensive tweet detection.

The paper is organized as follows: Section 2 provides the summary of related work. Section 3 gives the details about the implementation phases and the description about the components of the Offensive Tweet Detection Model. Data set’s detail depiction is provided in section 4.

Lexical feature based offensive tweet detection procedure is described in section 5 and Contextual feature based offensive tweet detection procedure is detailed in section 6. Results of both the detection techniques are depicted and compared in section 7 and details about the outcome are explained in section 7. The whole work has been concluded in section 8.

I. RELATED WORK

In this section, we briefly summarize the existing offensive content detection research approaches and their outcome. Researchers worked on varied directions and used varied tweet attributes while detection of offensive tweets such as tweet length, URL embedded in tweet, tweet content, retweet pattern, tweet sentiment, etc. Research Radius revolves around these tweet attributes while doing research for offensive tweets detection.

In 2012, Ying Chen [1] retrieved dataset from YouTube comment board. The offensiveness of the post was based on lexical syntactic architecture which was based on sentence offensiveness calculation and user offensiveness evaluation. The proposed architecture was able to detect offensive content to some extent but while carrying out the lexical analysis; it didn’t take in consideration the dependence of offensiveness on the topic about what the tweet was.

Sibel Adal [3] analyzed the various features that can be used to assess credibility of a tweet. The evaluation is carried across 8 diverse crawls of data from twitter. It was concluded that the best indicators of credibility include URLs, tweet length and retweet chaining. However, it doesn’t take in consideration the lexical aspect of the tweet. Besides the above mentioned features, various other features such as the adjectives present in the text often play a major role in deciding credibility of a tweet.

Later, Eric Cambria [2] in 2015 proposed a framework named concept-level sentiment analysis model, which takes into account various natural language processing tasks for extracting opinionated information from tweets. The model classifies the tweet as neutral, sarcastic, positive or negative by exploiting anaphoric adjectives and pronoun. Though the proposed model pays attention to only the important components of text of tweet i.e. adjective and pronoun, but the classification will not be accurate if the set of positive and negative adjectives are topic specific.

S. B. Madankar [4] proposes an automated system called filtering wall that is able to filter unwanted messages from OSN user walls. It uses machine learning text categorization techniques to automatically assign with each short text message a set of categories based on its content. The overall short text classification strategy is on Radial Basis Function Networks (RBFN) as they are efficient in acting as soft classifiers, in managing noisy data and intrinsically vague classes. Jinju Joby P. [5] work was centered around using Machine Learning Techniques for filtering messages into neutral and non-neutral. Preprocessing of data is done like removal of slang words, stop words etc. and later Naive Bayes and SVM [25] were applied. A sentence is declared as non-neutral if that contains hatred, sexual, offensive, pun-intended content. The research concluded that Naive Bayes is better for classification of textual data, but more techniques can be applied in order to find the best technique.

G. Aghila [6] took Naive Bayes as base algorithm and gave several types of input in order to understand the behaviour of Naive Bayes like noun phrase approach, Naive Bayes with SVM, Naive Bayes with Active Learning Boosting method etc. With so many types of combinations, they concluded that Naive Bayes can be used for text classification and may work for large dataset [28] as well but they did not take into consideration the fact that classification can be dependent on topic on which the dataset is created.

J. Ratkiewicz [7] created a system application for classification of meme posted on twitter as Truthy, Legitimate and Remove. The tweets posted are classified into these categories and the proposed algorithm is compared with SVM, Decision Stamp and AdaBoost. Preprocessing of data is an important aspect for classification which was not taken consideration till the required level for any classifier.

Zhe Zhao [8] worked on rumor detection, hence considered certain phrases into consideration using keyword selection on unpredictable real world events. Experiments were made on a 72 core Hadoop clusters (version 0.20.2) For faster results even on large datasets using apache pig. Ranking of candidate rumors was done in different variants, such as- popularity (no. of tweets), trending topics – larger no. of tweets (MapReduce), hashtag tracking – different layers, corrections only - patterns and statements etc. Later, applied svm and decision tree ranking.

Dongxu Huang [9] figured Out that probabilistic model has difficulty in achieving high precision because the characters of a tweet are no more than 140. Hence c-lpa algorithm was proposed. Labeled the tweets by canopy cluster algorithm and used label propagation algorithm for labeling the uncertain tweets in the overlapping of different clusters to achieve high precision and recall rate. R. Kishore Kumar [10] Used spam dataset from uci machine learning repository for feature construction and feature selection and applied 15 different classification algorithms based on relevant features. After performance evaluation, the rnd tree classification is considered as a best classifier, as it produced 99% accuracy through fisher filtering feature selection.

Andrea Zielinski [11] perform classification on tweets by regular expressions (via preserving the sequence of words) and by naïve bayes with a list of multilingual keywords related to "earthquake" in English, Greek, Romanian and Turkish. Thus, plan to use twitter analysis for crisis management. Dimitris Gavrilis et al [12] classified the spam emails by using the genetic algorithm feature selection (entropy and fine tuning) and classification and achieved the 96-97 percent of accuracy. A comprehensive study of social networks has been done by Charalampos Chelmis et al [13]; they present the state of art of content analysis and semantic analysis impact on social media. Bonchi et al [14], give the broad overview of the social networks key problems and techniques to deal with these problems from the business applications prospective with the use of business processes classification frameworks. Salton G. et al. [15] give the broad view of automatic term weighting, and insights about the best document weighting and query weighting and provides performance based analysis and comparison among various techniques term weighting. Yu-Hwan Kim et al [16] proposed a boosing based learning method for text filtering using naïve Bayes classifier to acquire more considerable LF1, LF2, F1 and F3 measures for TREC document entries. Yerazunis, W.S [17] compare various training methods such as TEFT, TOE, and TUNE, as well as pure Bayesian, token-bag, token sequence, SBPH, and Markovian discriminators with experiments and find that the best method for training is TUNE and Bayesian is less precise than Markovian discrimination. Detection of Rumors is done by Zhe Zhao [18] on tweets and find that one third of the top 50 clusters (similar / related posts) are rumors. Classification of the tweets on the basis of their theme and affairs is done by Huang [19]. Author propsed the novel approach to classify the messages by combining the cluster algorithm with label propagation algorithm with the aid of LDA (Latent Dirichlet Allocation and Single-Pass cluster algorithm. R. Kishore Kumar, [20] apply different classification algorithms to classify the email spam dataset and find out the best classifier with its validation and accuracy measures. Ulrich Bügel et al [21] work on the multilingual tweets related to ten earthquake events and translate the main keywords into English language to detect the earthquakes related tweets.

It was observed that the existing solutions solved the problem of detection of offensive content to some extent but still there are various questions which remain unanswered. For example, the methods proposed above provide solution to detect general offensive content but in the real life scenario the offensive content [26] is sometimes specific to a particular topic and hence contains some key offensive words which can't be detected by the existing solutions.

Also, most of the previously proposed solutions are subjected to constraint that the dataset being tested has relatively equal number of offensive and non - offensive tweets but this not always true. With our proposed system we have tried to solve such problems by taking in consideration the context and the subject of tweet during analysis.

II. EXPERIMENTAL SETUP

A. DEFINITIONS:

a. Defining 'Offensive':

As there is no universal agreement as to what is "offensive," for this study, we employ Jay and Janschewitz (2008) definition of offensive language as vulgar, pornographic, and hateful language [22]. Vulgar language refers to coarse and rude expressions, which includes explicit and offensive reference to sex or bodily functions; hateful language includes any communication outside the law that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, and religion.

b. The Computational Problem:

Based on the theoretical definition, we can formally define offensive content in two categories:

LEXICALLY OFFENSIVE CONTENT:

Lexically offensive content refers to the content containing general abusive and offensive words which will be perceived as an insult by an individual or group of individuals in any situation.

CONTEXTUALLY OFFENSIVE CONTENT:

Contextually offensive content refers to the content that may not contain any abusive or offensive content but can be classified as offensive on the basis of context. For example - "you are such a cry baby" doesn't contain any offensive word but still can be classified as offensive on the basis of context.

B. OFFENSIVE TWEET DETECTION Model

A multi-component model of our proposed offensive tweet detection is presented in this section. This model consists of 3 boundaries (Data Acquisition, Data Cleaning Unit, and Analyzer unit) which divide into 10 subcomponents as shown in figure 1. Component those are depicted in figure are required to achieve a satisfactory outcome. Subcomponents are as follows:

Data Acquisition : It is done by using the twitter REST API tweepy, the system communicates with twitter in order to extract tweets.

Data Cleaning unit component of the architecture, the raw tweets extracted using the Tweepy REST API are pre-processed so that they could be further classified by the feature extraction unit.

Lexical feature generation unit has created two types of bag of words (BoW), First contains the highly offensive words and second contains the words that when seen individually are not highly offensive but when used with other offensive words, nouns and proverbs, they tend to be highly offensive. This will help in classifying the tweets as lexically offensive and lexically clean. Contextual feature generation unit also generates two types of lists those are used as training data for contextual analysis. First is a list of tweets extracted using the tweepy API. The tweets extracted contain the hashtags which are anti to the topic on which our testing dataset is based. The second one is of list of offensive adjectives from the most popular tweets which are anti to the topic of testing dataset.

Analyzer Unit: In Feature extraction unit, the tweets that are retrieved from the data cleaning unit are lexically classified using the highly offensive bag of words created in lexical Feature Generation Unit. Lexically clean unit depicts or stores the lexically clean tweets i.e. the tweets that doesn't contain any of the offensive words but contains phrases that can be contextually offensive. This unit further sends the tweets to the context based classifier for further classification.

Lexically offensive unit stores the lexically offensive tweets i.e. the tweets containing the offensive words. This unit further sends the tweets for further classification to the lexical and context based classifier. Context based classifier unit, when coupled with the lexically clean unit, it further classifies the tweets as neutral and non-neutral tweets on the basis of the subjective content of the tweet by using natural language processing approaches. Lexical classifier unit retrieves tweets from the lexically offensive unit and classifies the tweets containing less offensive words as neutral and non-neutral using Rules and LDA based approach. In order to show a comparison of the best feature generated Naive Bayes is used for Rules and a Hybrid of LDA and Naive Bayes.

In Last unit, analytics unit, the various machine learning and natural language processing techniques used to build classifiers are compared on the basis of precision, recall and accuracy.

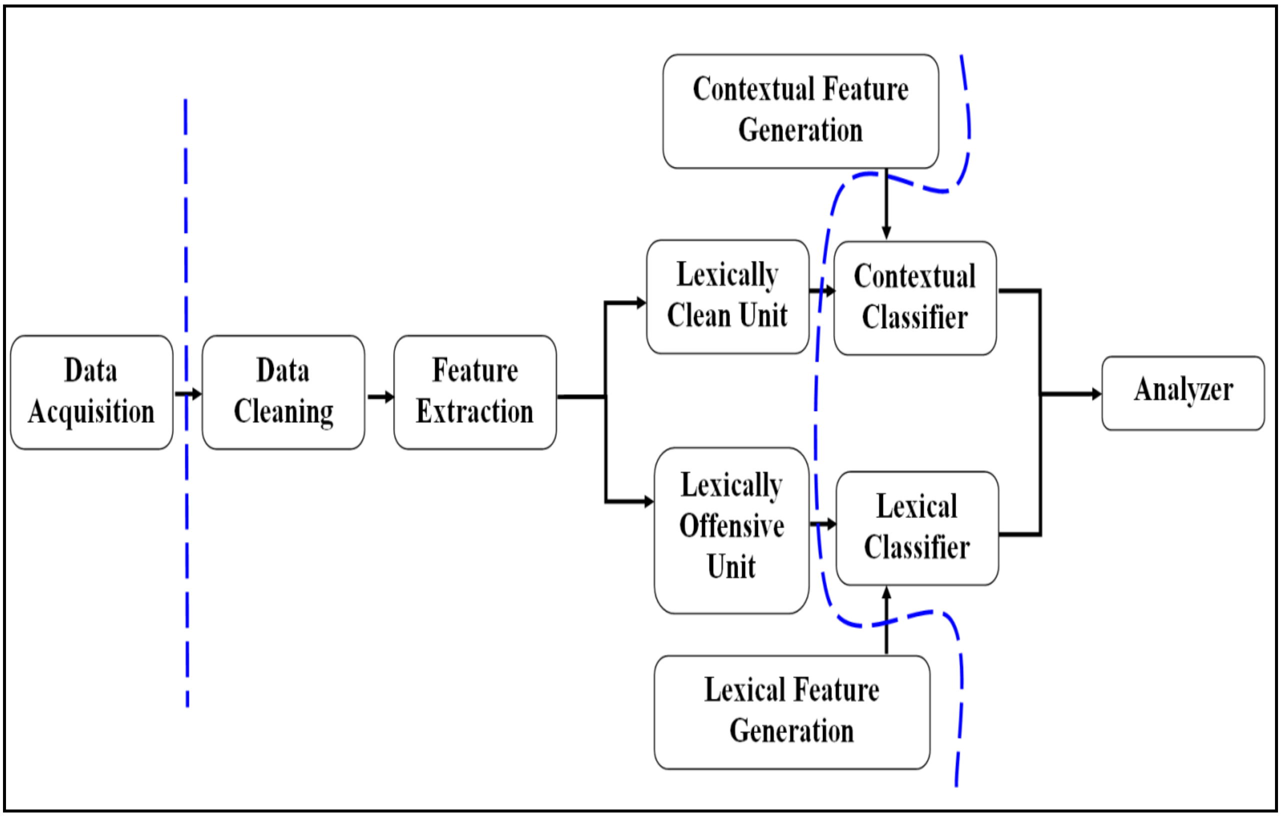


Figure 1 Offensive Tweet Detection Multi-Component Model

III. DATASET

Dataset is collected over a period of time as per the requirement. These tweets are collected for specific terms i.e., both offensive and non-offensive. This task was done using Tweepy API for streaming data. The key personality on whom dataset was acquired was Mr. Donald Trump. During the time of US Presidential Election there exist several kinds of reactions for Donald Trump. Therefore, tweets regarding those reactions, feeling of people etc was taken as the main filter for extracting tweets. The dataset consisted of around 1, 00,000 distinct tweets

A. Data Set- Data Acquisition

For the dataset we first select a high profile event victory of Mr. Donald Trump. It was the most trending topic on social media and news channels. Thus, large amount of tweets related to this topic were posted and retweeted. Among these tweets many were containing offensive statements and phrases. For the proposed system we collected 40,000 offensive tweets related to Donald Trump which were used as a training dataset. To ensure the tweets in training set is offensive, we have collected tweets containing hashtags like "#antitrump", "#notMyPresident" and "#dumbtrump". For test data we collected 1, 00,000 random tweets containing keywords like "trump" and "Donald trump". Data which was extracted with the help of a specific query is shown in figure 3 and overall offensive and non-offensive tweet data distribution is shown using pie chart in figure 2.

B. Cleaning of Raw Tweets:

Some data cleaning techniques have been applied to make is useable for offensive tweet detection. Techniques applied are- conversion to lowercase, standardizing slang words, removal of punctuation marks and stop words. Further, edit distance algorithm was used to correct the spellings in tweets. Also, as tweets are highly informal, repeated characters was removed from words. For example, awesomeeeeeeeee is converted to awesome.

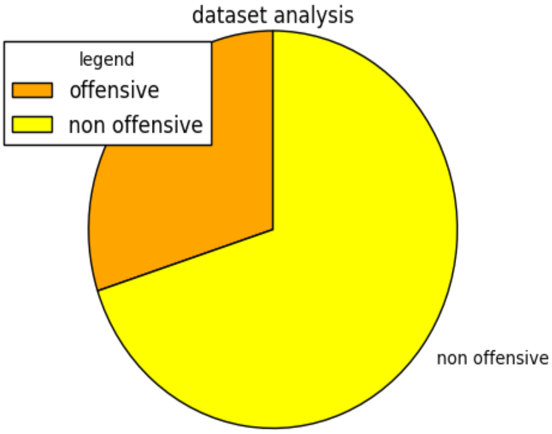
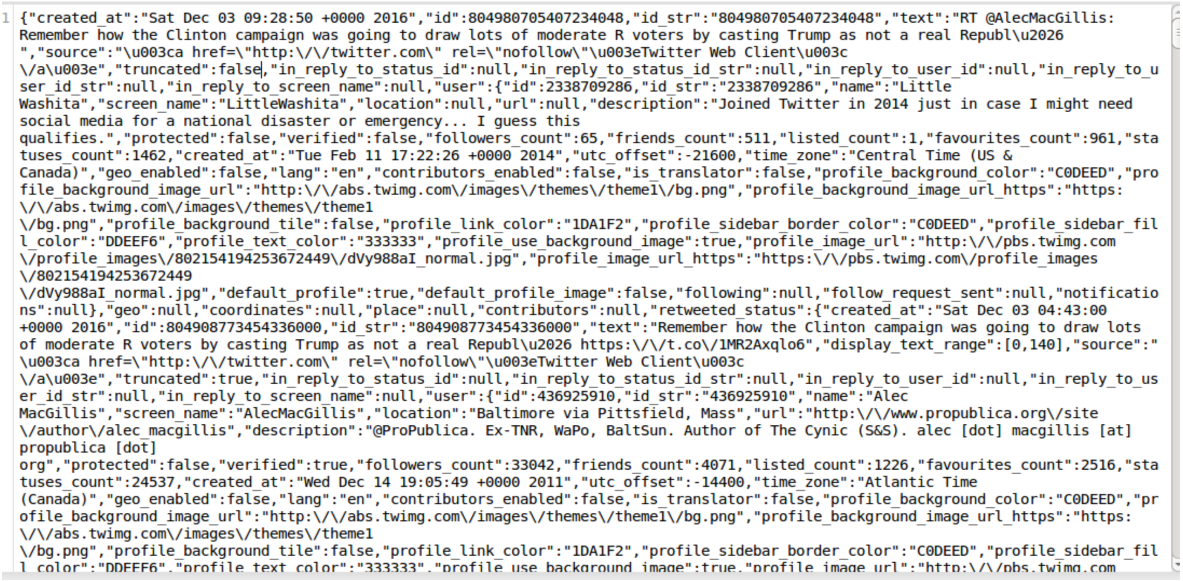


Figure 2 Data Set



```
{"created_at":"Sat Dec 03 09:28:50 +0000 2016","id":804980705407234048,"id_str":"804980705407234048","text":"RT @AlecMacGillis:
Remember how the Clinton campaign was going to draw lots of moderate R voters by casting Trump as not a real Republ\u2026
","source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c
\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_u
ser_id_str":null,"in_reply_to_screen_name":null,"user":{"id":2338709286,"id_str":"2338709286","name":"Little
Washita","screen_name":"LittleWashita","location":null,"url":null,"description":"Joined Twitter in 2014 just in case I might need
social media for a national disaster or emergency... I guess this
qualifies.","protected":false,"verified":false,"followers_count":65,"friends_count":511,"listed_count":1,"favourites_count":961,"sta
tuses_count":1462,"created_at":"Tue Feb 11 17:22:26 +0000 2014","utc_offset":-21600,"time_zone":"Central Time (US &
Canada)","geo_enabled":false,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","pro
file_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:
\/\/abs.twimg.com\/images\/themes\/theme1
\/bg.png","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fil
l_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com
\/profile_images\/802154194253672449\/dVy988aI_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images
\/802154194253672449
\/dVy988aI_normal.jpg","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notificatio
ns":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"Sat Dec 03 04:43:00
+0000 2016","id":804908773454336000,"id_str":"804908773454336000","text":"Remember how the Clinton campaign was going to draw lots
of moderate R voters by casting Trump as not a real Republ\u2026 https:\/\/t.co\/1MR2Axqlo6","display_text_range":[0,140],"source":"
\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c
\/a\u003e","truncated":true,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_us
er_id_str":null,"in_reply_to_screen_name":null,"user":{"id":436925910,"id_str":"436925910","name":"Alec
MacGillis","screen_name":"AlecMacGillis","location":"Baltimore via Pittsfield, Mass","url":"http:\/\/www.propublica.org\/site
\/author\/alec_macgillis","description":"@ProPublica. Ex-TNR, WaPo, BaltSun. Author of The Cynic (S&S). alec [dot] macgillis [at]
propublica [dot]
org","protected":false,"verified":true,"followers_count":33042,"friends_count":4071,"listed_count":1226,"favourites_count":2516,"sta
tuses_count":24537,"created_at":"Wed Dec 14 19:05:49 +0000 2011","utc_offset":-14400,"time_zone":"Atlantic Time
(Canada)","geo_enabled":false,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","pr
ofile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:
\/\/abs.twimg.com\/images\/themes\/theme1
\/bg.png","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fil
l_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com
```

Figure 3: sample tweet extracted

IV. LEXICAL FEATURE BASED OFFENSIVE TWEET DETECTION

A. FEATURE GENERATION

For lexical analysis two types of bag of words (BOW) are created. BOW-1 contains the highly offensive words and the other contains the less offensive words (BOW-2) which when used with other offensive words or nouns or proverbs then can be termed as highly offensive as shown in Table 1. For example, phrase such as “this is a stupid thing” cannot be termed as offensive but a phrase like “you are so damn stupid” where stupid is being used with a second person can be termed as offensive.

TABLE 1:TYPES OF BAG OF WORDS USED FOR LEXICAL ANALYSIS

Examples	**Types of Bow**

Fuck, fucking etc	Type – 1
Stupid, silly etc	Type – 2

B. FEATURE EXTRACTION

In the feature extraction phase, the preprocessed tweets are then classified as highly offensive. For this the BOW-1 created in feature generation phase is used. If the tweets contain any of the words that are present in BOW-1, then they are classified as offensive.

In the feature extraction phase, the remaining preprocessed tweets are then classified as lexically clean and lexically offensive. For this the BOW-2 created in feature generation phase is used. If the tweets contain any of the words that are present in BOW-2, then they are classified as lexically offensive.

C. APPROACH1: RULE BASED NAÏVE BAYES (RNB)

For classifying the tweets, we have 2 approaches for generating the features and later those features are used as input to Naive Bayes` Machine Learning Algorithm.

Rule based Naïve Bayes machine learning algorithm is based on Bayes theorem of probabilistic classifiers and work on the strong assumption between the independent features of the data set. Rule Based Approach: Certain rules were created from various references and understanding of content that may be offensive like 'You are a stupid boy', 'Stupid Modi, Stupid Demonetization' etc. Later, these rules are used to classify data into 2 classes- Neutral and Non-Neutral which are taken as input to Naive Bayes. The rules are presented in Table II.

TABLE II:TYPES OF BAG OF WORDS USED

Examples	Language Features
<You, gay>	Second person pronoun + perojective
<stupid, boy>	Offensive adjective + people referring terms
<stupid, Modi>	Adjective + proper noun

D. APPROACH2: LDA DEPENDENT NAÏVE BAYES (LDANB)

LDA Based Approach: Latent Dirichlet Allocation was used to divide the dataset into 2 clusters like above then the same type of input is given to Naive Bayes`. LDA does not need the features to be independent and use for the dimensionality reduction. It is a preprocessing step for machine learning application and pattern classifications.

Later, Naive Bayes` tell which technique for feature generation for machine learning can be used in order to classify whether tweets are offensive or not.

V. CONTEXTUAL FEATURE BASED OFFENSIVE TWEET DETECTION

A. FEATURE GENERATION:

For contextual analysis again two types of lists are generated. These lists are used as training data for contextual analysis. One is a list of tweets containing the hashtags which are anti to the topic on which our testing dataset is based. The second one is of list of offensive adjectives from the most popular tweets which are anti to the topic of testing dataset. The tweets antis to the topic were extracted though the most used anti hashtags for Donald Trump such as “#notMyPresident”, “#Antitrump”.

Popularity of the tweets was determined by the followers of the tweet. Tweets contain more than 2000 followers are identified as popular and stored in the dataset.

The contextual classifier used one Similarity indexing approach- Cosine Similarity Measure and another one is natural language processing approach-Adjective Based Offensive tweet detection. The llater approach used the adjective based contextual features.

B. APPROACH1: COSINE SIMILARITY MEASURE:

In Cosine similarity [23] documents are represented as term vectors and similarity of two documents correspond to the correlation between the vectors. This is quantified as the cosine of the angle between vectors.

Each term in a document vector represents a dimension which is according to the weight of the term in document. Cosine similarity between 2 vectors v1 and v2 are computed by the following equation.

$$cos\theta = (v1.v2)/(|v1|.|v2|)$$

It uses the tweet list feature generated in feature generation phase as training data. Each tweet in the training data as well as the tweet under test is treated as a document. Each of them is converted into a vector form based on the count frequency of each term in the tweet. Then, we compared the cosine similarity of the tweet being processed with the training data. If the cosine similarity between the tweets is being tested and any of the tweet in the training dataset is greater than 60% then that tweet is classified as offensive.

C. APPROACH2: ADJECTIVE BASED APPROACH

A new natural language processing-based approach was developed for contextual analysis. In English language adjectives are basically used to identify or quantify individual people and unique things. Hence, the offensiveness of a tweet highly depends on the adjectives present in the tweet. For example - “how can trump be so ignorant”. Therefore, instead of giving importance to each word of the text of tweet, focus had been put only on the adjective present in the text. For the proposed methodology, the adjective feature generated in the feature generation phase was used to classify the tweets. This adjective feature contains all the offensive adjectives related to Donald Trump. These adjectives were extracted from the tweets which had content that was against Donald trump. The tweets containing any of these adjectives were classified as offensive. The rest remaining tweets were classified as neutral tweets.

VI. EXPERIMENTAL RESULTS- COMPARATIVE ANALYSIS AND FINDINGS

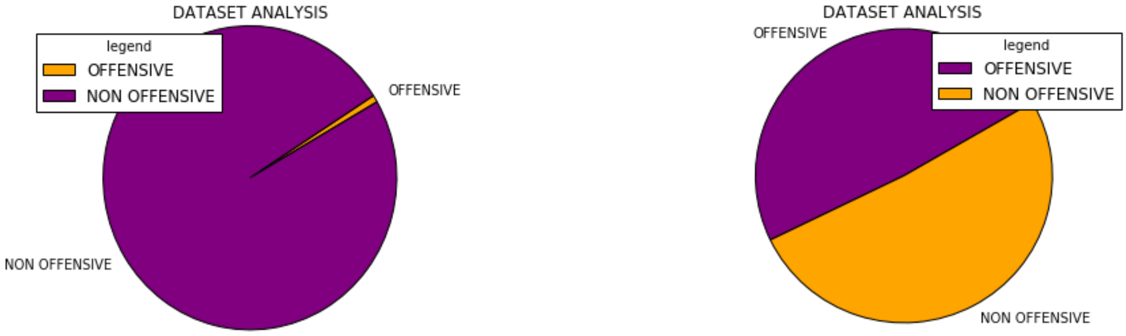
In our tweets dataset which is used to validate the experiments contain more non-offensive tweets as compared to offensive tweets (See figure 2). The three standard performance measures have been used to confirm accuracy of the Multi-Component offensive tweet detection model. These measures are Precision, Recall and Accuracy [24] where,

- Precision denotes the fraction of classified (Offensive/ Non-Offensive) tweets that are relevant to the requirement.
- Recall denotes the fraction of relevant classified tweets those are retrieved
- Accuracy defines as correct classified tweet percentage.

This section is further divided into two subsections where, first section details the outcome of offensive tweet detection using lexical analysis and second section details Contextual Analysis based offensive tweet detection outcome.

A. LEXICAL ANALYSIS RESULTS:

It is observed that Rule based Naïve Bayes detects extremely less number of offensive tweets and classifies all as non-offensive tweets. Reason of such degraded in performance is due to the self generated rules for naïve bayes. Whereas, on the other end, LDANB gives refined and improved performance because rules are generated according to their actual usage in the tweets content (See Figure 4).

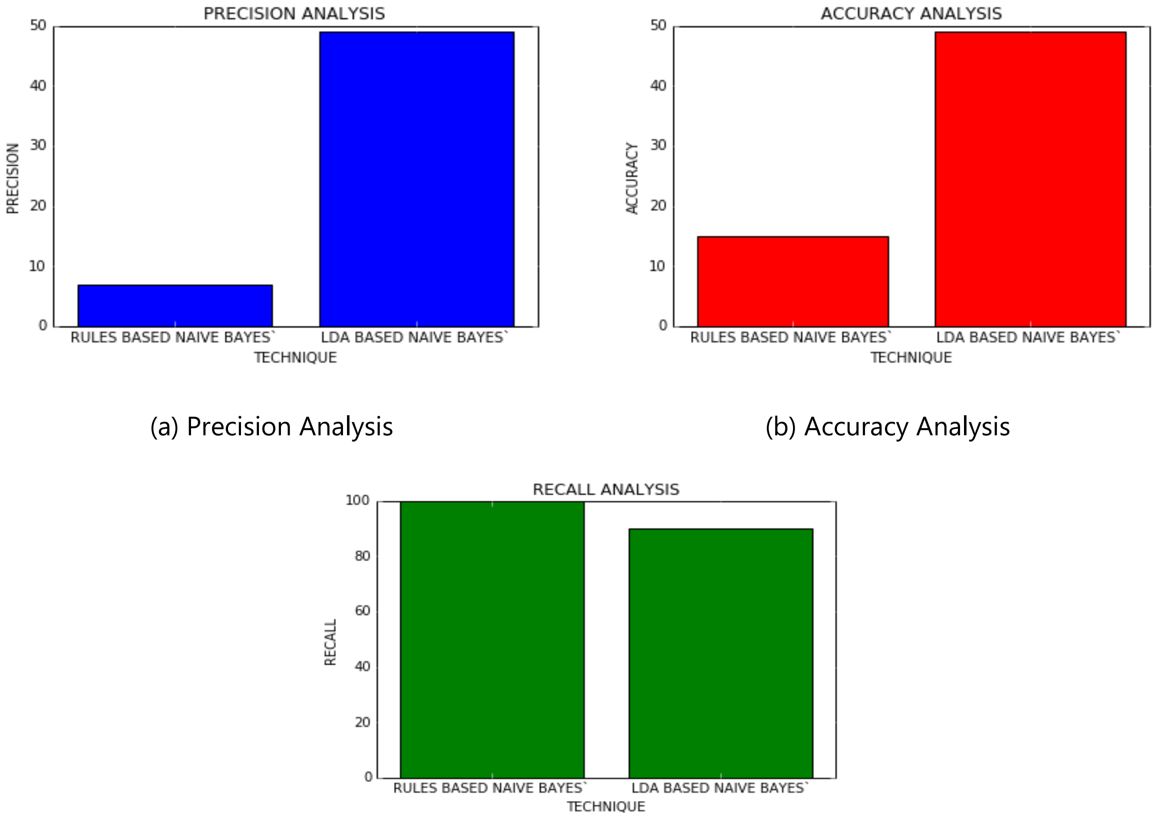


(a) Using Rule Based Naïve Bayes (RNA)

(b) LDA Based Naïve Bayes (LDANB)

Figure 4: Comparative Analysis in between RNB and LDANB

Table 3 details the comparative confusion matrix of RNB and LDANB results which shows count of true positive, true negative, false positive and false negative. Figure 5 (a) shows that RNB gives a precision of 7% while LDANB produce a precision of 49%. Thus, LDANB gives high precision as compared to Rule Based Naive Bayes and this is due to the fact that Rule based approach only considers the words occurrence once in tweet and LDA considers inverse document frequency probability as well. Though, both the techniques RNB and LDANB provide low precision. Figure 5(b) shows accuracy comparison in RNB and LDANB, out of which RNB provides 15% accuracy which increased to 49% using LDANB. Further, smaller dataset produces high recall and RNB makes clusters on a specific word. Henceforth, Figure 5(c) shows Recall Comparison where RNB has nearly 100% recall and LDANB has 90% recall.



(a) Precision Analysis

(b) Accuracy Analysis

(c) Recall Analysis

Figure 5: Lexical Feature based Offensive Tweet Detection Results

TABLE 3: LEXICAL ANALYSIS CONFUSION MATRIX

	TP	TN	FP	FN
RNB	3	3	389	0
LDANB	174	23	179	19

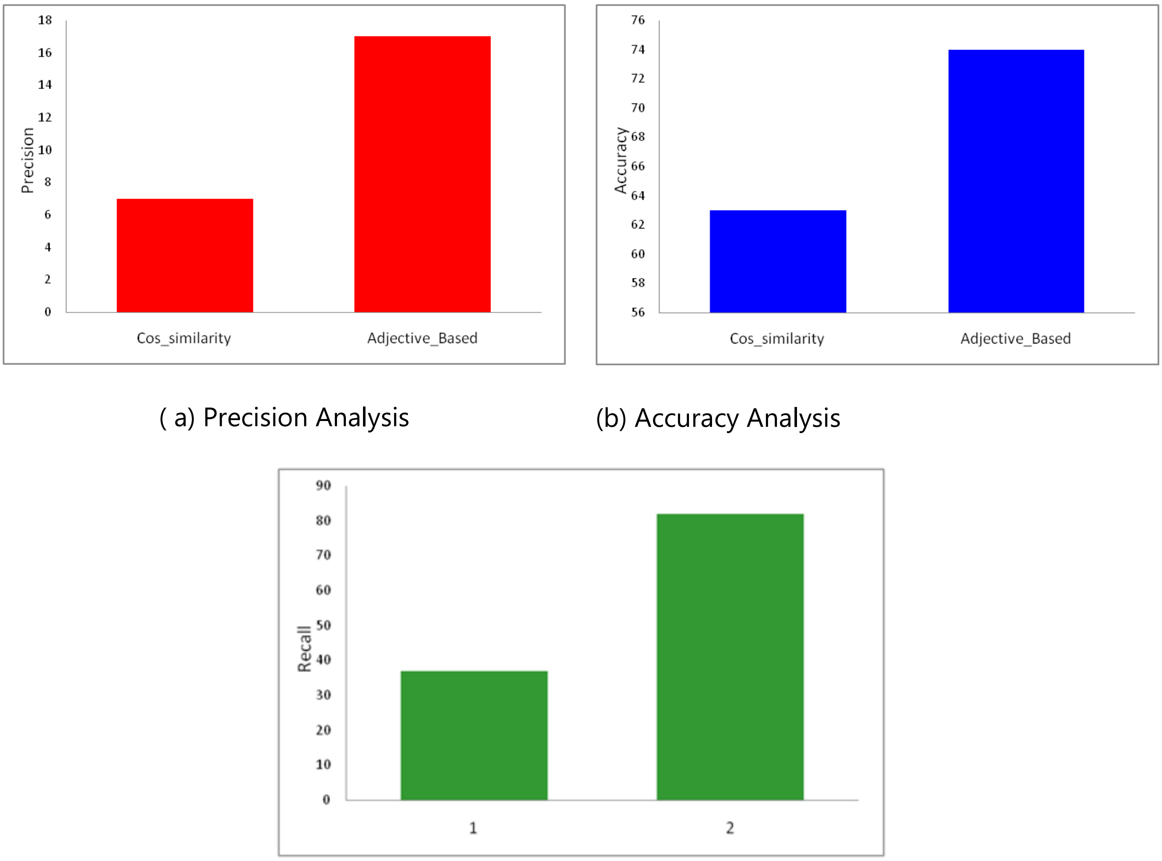
B. CONTEXTUAL ANALYSIS RESULTS

As it can be seen in the figure, the test dataset for contextual analysis contains more of non offensive tweets than neutral tweets. The ratios of offensive and non offensive tweets remain same irrespective of the topic taken. Hence the classification of tweets becomes difficult.

Adjective based approach has a precision of 17% while cosine similarity has a precision of 7%. Hence, Adjective based approach has a twice as more precision than cosine similarity. Therefore, it outperforms the cosine similarity approach in terms of precision. The main reason for this is that cosine similarity gives equal importance to all words in tweet while adjective based approach emphasizes only on adjectives. Still both the approaches have a considerably low precision. Adjective based approach has an accuracy of 74% while cosine similarity has an accuracy of 94%. Therefore, Cosine similarity approach outperforms the adjective based approach in terms of accuracy. The main reason for this is that cosine similarity concentrates on all terms in tweets and hence catches more positive results. Both the approaches have relatively high accuracy. Adjective based approach has a recall of 82% while cosine similarity has a precision of 16%. Adjective based approach has 5 times as more recall than cosine similarity. hence, outperforms the cosine similarity approach in terms of recall. The main reason for this is that cosine similarity gives equal importance to all words in tweet. Hence the number of false positives is more while adjective based approach emphasizes only on adjectives which are the main parameters on which the offensiveness of a tweet depends.

TABLE 3: CONTEXTUAL ANALYSIS CONFUSION MATRIX

	TP	TN	FP	FN
Cosine Similarity	45	1167	600	104
Adjective based Approach	98	1303	453	21



(a) Precision Analysis

(b) Accuracy Analysis

(c) Recall Analysis

Figure 7: Contextual Feature Based Offensive Tweet Detection Results

VI. CONCLUSION

Lexical Classification was done on 2 techniques out of which the newly devised Collaborative Model of LDA and Naive Bayes is better than the Rule Based Naive Bayes approach. Although, the rules were created on thorough research [1] but even though the results generated are not up to the mark as compared to the other technique in terms of accuracy and precision. The main reason for the failure of Rule based Naive Bayes approach is that it only takes into consideration adjective and proper nouns. It did not take into consideration the frequency of word in the tweet which was the game changer in case of Collaborative Model of LDA and Naive Bayes approach. The Collaborative was the showstopper of the research, as it has considerably good precision and accuracy. Though the precision and accuracy of Collaborative Model were better, but we cannot neglect the recall of Rule Based Approach. This was due to the fact that we had less number of documents. The Collaborative Model has its own advantages as the two techniques just fits each other quite well. It was because the clusters formed by LDA were more detailed and systematic. Also, the shortcoming of Rule Based Naive Bayes` approach was also resolved by taking into consideration the frequency of words in the document. Hence, it is concluded from above results that Collaborative Model of LDA and Naive Bayes` produced better result and can be used in order to detect offensive and non-offensive tweets.

Contextual classifier used two approaches for detection of offensive content which presented different aspects of natural language processing. While the cosine similarity approach was based on document similarity between two documents, the newly devised adjective based approach used parts of speech methodology to extract offensive adjectives from the training set and used them to analyze the testing set.The performance of adjective based approach is independent of the volume of the offensive content present. It performs equally well whether the volume of offensive content is low or high. This is because it has no relation to the volume of the data being tested. It is only concerned with the individual tweet being tested. From the comparative study of context based classifiers it was found that the newly devised adjective based approach outperforms the existing cosine similarity based approach in terms of recall and precision when the offensive content volume is low in the test dataset. The main reason for this is as the adjective based approach only takes in consideration the most important part of speech of text for classification i.e. adjectives, it will always perform optimally irrespective of the volume of the offensive content in the test data. The cosine similarity based approach has a upper hand in terms of accuracy of classification. But this approach offers very low precision and recall with nearly unacceptable values. Hence it can be concluded that cosine similarity based approach is not optimal for context based classification when the volume of offensive content is low.

REFERENCES

[1] Chen, Ying, et al. "Detecting offensive language in social media to protect adolescent online safety." Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom). IEEE, 2012.

[2] Cambria, Erik, et al. "The CLSA model: a novel framework for concept-level sentiment analysis." International Conference on Intelligent Text Processing and Computational Linguistics. Springer International Publishing, 2015.

[3] ODonovan, John, et al. "Credibility in context: An analysis of feature distributions in twitter." Privacy, Security, Risk and Trust (PASSAT), 2012 international conference on and 2012 international conference on social computing (SocialCom). IEEE, 2012.

[4] Dhanake, M. S. A., & Nandedkar, V. S. (2014). An Automated System to Filter Unwanted Message from OSN User Wall. Information Technology, 5, 1.

[5] Jinju Joby, P. and Korra, J., Message Filtering on Social Media Content. structure, 1, p.3.

[6] Aghila, G. "A Survey of Naive Bayes Machine Learning approach in Text Document Classification." arXiv preprint arXiv:1003.1795 (2010).

[7] Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A. and Menczer, F., 2011. Detecting and Tracking Political Abuse in Social Media. ICWSM, 11, pp.297-304.

[8] S. Venkata Lakshmi, K. Hema Filtering Information for Short Text Using OSN International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)317 Vol. 2, Issue 2, Ver. 2

[9] Frakes, W., Baeza-Yates, R. (eds.): Information Retrieval: Data Structures & Algorithms.Prentice-Hall (1992)

[10] Manning, C., Raghavan, P., Schutze, H: Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK (2008)

[11] Boykin, P.O., Roychowdhury, V.P.: Leveraging social networks to fight spam. IEEE Computer Magazine 38, 61–67 (2005)

[12] Gavrilis D, Tsoulos I G and Dermatas E (2006), Neural Recognition and Genetic Features Selection for Robust Detection of E-mail Spam, Lecture Notes in Computer Science, pp. 39-55, 498-501

[13] Chelmis C, Prasanna VK. Social networking analysis: A state of the art and the effect of semantics. Privacy, security, risk and trust (Passat), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (socialcom). IEEE, 2011.

[14] Bonchi F, Castillo C, Gionis A, Jaimes A. JaimesA. Social Network Analysis and Mining for Business Applications. ACM Transactions on Intelligent Systems and Technology (TIST). 2011

[15] Salton G., Buckley C., Term-weighting approaches in automatic text retrieval. Information Processing and Management 24(5), 513–523 (1988)

[16] Kim, Y.H., Hahn, S.Y., Zhang, B.T.: Text filtering by boosting naive bayes classifiers. In: SIGIR’00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 168–175. ACM, New York, NY, USA (2000)

[17] Yerazunis W (2004), The Spam FIltering Accuracy Plateau at 99.9 Percent Accuracy and How to Get Past it, Proceedings of the MIT Spam Conference.

[18] Zhao, Zhe, Paul Resnick, and Qiaozhu Mei. "Enquiring minds: Early detection of rumors in social media from enquiry posts." In Proceedings of the 24th International Conference on World Wide Web, pp. 1395-1405. ACM, 2015.

[19] Huang, Dongxu, and Dejun Mu. "Topic detection in twitter based on label propagation model." In Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2014 13th International Symposium on, pp. 97-101. IEEE, 2014.

[20] Kumar, R. Kishore, G. Poonkuzhali, and P. Sudhakar. "Comparative study on email spam classifier using data mining techniques." In Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, pp. 14-16. 2012.

[21] Zielinski, Andrea, Ulrich Bügel, L. Middleton, S. E. Middleton, L. Tokarchuk, K. Watson, and F. Chaves. "Multilingual analysis of twitter news in support of mass emergency events." In EGU General Assembly Conference Abstracts, vol. 14, p. 8085. 2012

[22] Jay, Timothy, and Kristin Janschewitz. "The pragmatics of swearing." Journal of Politeness Research. Language, Behaviour, Culture 4.2 (2008): 267-288.

[23] Huang, Anna. "Similarity measures for text document clustering." Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008.

[24] Fabrice Guillet, Howard J. Hamilton. “Quality Measures in Data Mining.” Springer, 17-Jan-2007

[25] Aggarwal, Niyati, et al. "Analysis the effect of data mining techniques on database." Advances in Engineering Software 47.1 (2012): 164-169.

[26] Aggrawal, Niyati, and Anuja Arora. "Vulnerabilities Issues and Melioration Plans for Online Social Network Over Web 2.0., CDQM, An Int. J., Volume 19, Number 1, 2016, pp. 66-73"

[27] Aggrawal, Niyati, and Anuja Arora. "Visualization, analysis and structural pattern infusion of DBLP co-authorship network using Gephi." Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on. IEEE, 2016.

[28] Aggrawal, Niyati, et al. "Brand analysis framework for online marketing: ranking web pages and analyzing popularity of brands on social media." Social Network Analysis and Mining 1.7 (2017): 1-10.