NATURAL LANGUAGE PROCESSING (CS F429)
FIRST SEMESTER: 2021-22

RESEARCH PAPER

SENTIMENT ANALYSIS-TWITTER COMMENTS

Group Number: 27

Group Members:	ID Numbers:
ROHITH KUMAR GATTU	2019A7PS0049H
T SAI PRASOONA	2019A8PS0603H
SIREESHA RALLABHANDY	2019AAPS0198H
KASINA SATWIK	2019A7PS0011H

Department of Computer Science and Information Systems
BITS Pilani Hyderabad Campus

ABSTRACT

With the introduction of digital technology and its expansion, the web now contains a massive amount of data for internet users, as well as a large amount of data being created. The internet has evolved into a platform for online learning, idea exchange, and opinion sharing. Social networking sites like Facebook, Twitter, Instagram allow everyone to express themselves and have discussions on various topics with other people on that platform through comments and posts. With the increasing connection between people across the world, toxicity, negativity is also spreading, impacting people's mental health and sometimes physical health too.Our objective through this project is to build a bot that automatically hides/ deletes the negative Twitter comments specific to a tweet of a user. We perform sentiment analysis on Twitter comments, classifying them as positive, neutral, and negative and hiding the negative tweet comments using Twitter API.
On Twitter data, we investigate sentiment analysis. This study makes the following contributions:

1) We introduce POS-specific prior polarity features.
2) We look into using a tree kernel to avoid the need for time-consuming feature engineering. Both the additional features (when used in conjunction with previously proposed features) and the tree kernel outperform the current state-of-the-art baseline.

Keywords -- NLP, Twitter Comments, Sentiment Analysis, Model Training, Machine learning Algorithms.

INTRODUCTION

The Internet has changed the way people express their thoughts and opinions today. It is currently mostly accomplished through blog entries, internet forums, product review websites, social media, and other similar mediums. Millions of individuals use social media sites such as Facebook, Twitter, Google Plus, and others to express their feelings, exchange opinions, and share perspectives about their everyday lives. Furthermore, social media gives a platform for businesses to communicate with their customers for advertising purposes. People rely heavily on user-generated content on the internet when making decisions. Marketers and firms use this sentiment analysis data to understand about their products or services in such a way that it can be offered as per the user“s requirements. Textual information retrieval strategies are primarily concerned with processing, finding, and interpreting the factual information available.

Although facts are objective, there are certain textual components that represent subjective traits. Opinions, feelings, assessments, attitudes, and emotions are the most common contents in Sentiment Analysis . Due to the massive expansion of available information on internet sources such as blogs and social networks, it presents numerous difficult chances for developing new applications.For example, using Sentiment analysis, it is possible to forecast recommendations of goods provided by a recommendation system by taking into account factors such as favourable or negative comments about such products.Not only for recommendation, marketing purposes but sentiment analysis can be used for many other purposes too. One of them is in microblogging websites.

Microblogging websites have evolved as sources of varied information. People in real-time post their microblogs, write messages about their opinions on various topics, discuss contemporary issues, and sometimes criticize them.

This free opinion flow has sometimes become toxic, which is affecting the mental health of the users due to negative or inappropriate comments. The purpose of the project is to reduce this toxic behavior towards a single user and give the user the ability to hide or delete the comments that show hatefulness or negative sentiments.This helps to improve the mental health of the user by not exposing them to such hateful speech. It also helps in making microblogging websites a positive and healthy social networking platform. In this paper, we are looking at one of the most popular microblogging sites called Twitter and building models for classifying the comments of the user-specific tweet into positive, negative, and neutral sentiment.

RELATED WORK

Sentiment analysis is a procedure that uses Natural Language Processing to determine attitudes, opinions, perspectives, and emotions from text, audio, tweets, and database sources. In the past few years, much work has been done in this area and initially, the classification was binarized into only positive and negative.

This binarized model was proposed by Pak and Paroubek[1]. They created a Twitter corpus by collecting tweets using Twitter API and annotating those tweets using emoticons. They developed a classifier using that corpus-based on that multinomial Naive Bayes theorem using N-gram and POS tags as features. Their model was less efficient as they used emoticons to classify.

Parikh and Movosatte[2] tried a different approach i.e Maximum Entropy Model to classify the tweets. They found that this model worked better than the Naive Bayes Classifier model.

Go and L.Huang[3] performed sentiment analysis by using distant supervision, in that the training data they used was made up of tweets with emoticons which served as noise. They build models using Naive Bayes Classifier and SVM. The features they used consisted of unigrams, bigrams, and parts of speech (POS) tagging. Their conclusion was that SVM surpassed other models and unigram was more effective.

Kamps et al[4] used the lexical database WordNet to determine the sentiment of the word along different dimensions. They used WordNet to build a distance metric and estimate the semantic polarity of adjectives.

Luoet. al.[5] discussed the difficulties and effective ways for extracting opinions from Twitter tweets. Opinion retrieval on Twitter is difficult due to spam and widely differing wording.

APPROACH/ METHODOLOGY

To begin, we used the Twitter API to gather tweets and their comments, which we then pre-processed to do data cleansing. Secondly, using any of the feature selection methods, the key characteristics are retrieved from the clean text. Next, to construct a training set, a subset of the data is manually labeled as good or negative Twitter comments. Finally, the extracted features and labeled training set are used to categorize the remaining Twitter comments using the created classifier.

TWITTER COMMENTS DATA COLLECTION METHODS

The three possible ways to collect Tweets for research are given below [6]:

- Data repositories such as UCI, Friendster, Kdnuggets, and SNAP.
- APIs: Twitter provides two types of APIs as search API and stream API. Search API is used to collect Twitter data on the basis of hashtags and stream API is used to stream real-time data from Twitter.
- Automated tools are further classifIed into premium tools such as Radian6 [7]. Sysmos, Sumplify360, Lithium, and non-premium tools such as Keyhole, Topsy, Tagboard, and SocialMention.

DATA SOURCES

While performing the sentiment analysis on Twitter comments, selecting the correct data sources plays a very vital role. A microblogging site like Twitter has gained higher popularity because of its limited strength of content and public availability of the data. Twitter was chosen as a data source for sentiment analysis because of its statistics.

TWITTER STUDIES:
Twitter works as an arena for topics like literally everything whether it's healthcare-related, politics, advertising, marketing, sports, etc. Twitter has been picked as the most promising source for research into community or impact identification, subject discovery, recommendation systems, tweet categorization, and other related topics.

TWITTER GROWTH RATE:
On average, every second around 6000 tweets are being tweeted all around the world. That means every minute 3,50,000 tweets are tweeted and 500 million tweets per day are tweeted. In Twitter's history, the number of tweets increased from 5000 tweets per day in 2007 to 50,00,00,000 tweets per day in 2013. It had a rise of almost 6 times its previous magnitude of

tweeting amount. Taking into account all these statistics, we figured to take Twitter for our project.

OTHER SOCIAL NETWORKING SITES:

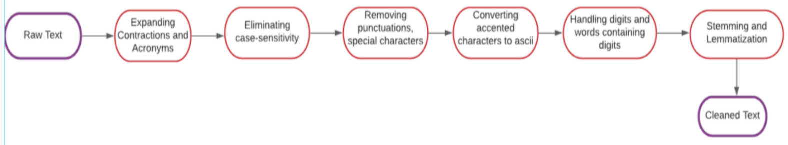
We considered using other famous networking sites like Facebook, Instagram, LinkedIn, TikTok for our project too. Facebook has on average, 2.68 billion active users out of which 1.82 billion people log in to it daily. According to research done in October 2020, Facebook users make an average of 5 comments per day. Instagram has 1.38 billion users. About 6 in every 10 people log in to Instagram daily. Even though these sites have more users logging in every day than compared to Twitter, the reason we chose Twitter over them was that Twitter revolves around every issue and not only entertainment. Twitter can be assumed to be a formal version of other social media sites.

TWEET COMMENTS:

A tweet is a message posted on Twitter and the reply to it is called a comment of that tweet which is generally limited to 280 characters. A tweeted comment generally consists of text, emoticons, and rare images and links. Mining is used to categorize text, links, photos, emoji or emoticons, and even films based on these components. There are two notations in the Tweet comments: hashtags (#), and account ID (@).

DATA PREPROCESSING

Twitter data mining is a challenging task. The data collected is raw and needs cleaning. Text preprocessing is a technique for cleaning text data and preparing it to be used in a model. Text data is one of the most amorphous forms of data available and when it comes to dealing with Human Languages involving slang and colloquialisms it's even more complicated so, we have to preprocess and feed it to the model so the machine can understand it. The pre-processing techniques are as follows:


Raw Text
Expanding Contractions and Acronyms
Eliminating case-sensitivity
Removing punctuations, special characters
Converting accented characters to ascii
Handling digits and words containing digits
Stemming and Lemmatization
Cleaned Text

EXPANDING CONTRACTIONS AND ACRONYMS:

Contractions are the shortened form of a word whereas **acronyms** are the abbreviations formed from the initial letters of each word and pronounced as a word. This expansion will give us better analysis and the original form will also help us in text standardization.

ELIMINATING CASE-SENSITIVITY:

If every word in the text data is in the same case then it will be easy for a machine to interpret the words. As in this model of ours, we only need information, not case distinctions.

REMOVING PUNCTUATIONS AND SPECIAL CHARACTERS

Special characters and punctuation marks don’t add any value to the text understanding and only induce noise into the algorithms, so it’s better to get rid of them. The 32 punctuation marks Python’s **string** module consists of are:

!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~

CONVERTING ACCENTED CHARACTERS TO ASCII

Accented characters are important elements that signify emphasis on a certain word during pronunciation or understanding but we need to remove them as most of the corpus we might come across will be accented characters free.

HANDLING DIGITS AND WORDS CONTAINING DIGITS:

There’s a high probability that the comments will include words formed from combining alphabets and numbers. We need to either delete such words or find the word that has many similarities to replace. And remove words formed with only numbers as they don’t convey any emotion.

REMOVE STOPWORDS:

Stopwords are the most commonly occurring words in the text which carry minimal to no importance. Stopwords are generally added to the sentences to make them grammatically correct.

STEMMING AND LEMMATIZATION:

Stemming is the process of reducing derived words to their stem, base, or root word. It removes the prefix or suffixes from the word like **ing, s, es**. But sometimes the conversion is not desirable. Whereas **lemmatization** uses vocabulary and converts properly by aiming to remove inflectional endings only. While choosing between these we need to keep in mind the performance.

FEATURE EXTRACTION

The data that has been pre-processed has a number of features. We extract different factors like adjectives, verbs, and nouns using feature extraction methods, and then classify them as "positive" or "negative" to determine the polarity of the whole comment.

The feature extraction methods used are

TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY:

These characteristics define separate and distinct words, as well as their frequency of recurrence.

NEGATIVE PHRASES:

The presence of negative words has the potential to alter the meaning or direction of a viewpoint. As a result, it is obvious that negative word orientation should be considered.

PARTS OF SPEECH (POS):

Finding nouns, verbs, adjectives, and so on as they are important gauges of opinion.

SENTIMENT CLASSIFICATION TECHNIQUES

There are generally two techniques to identify the sentiment of the text [8] [9]: knowledge-based technique and machine learning techniques.

The knowledge-based technique is also called Lexicon based technique. The lexicon-based technique focuses on deriving the opinion based lexicons from the text and then identifying the polarity of those lexicons. Lexicons are the collection of known and precompiled sentiment terms. This approach is further classified into the Dictionary-based approach and the Corpus-based approach. In the Dictionary-based approach, we find the opinion-oriented words, and then examine the dictionary to collect their synonyms and antonyms. Whereas in the Corpus-based approach, we create a list of opinion words and then based on their context-specific orientations, we find additional related opinion words in a vast corpus.

The main objective of machine learning techniques is to develop an algorithm that optimizes the performance of the system using training data such as examples and/or past knowledge and experiences. Machine learning provides a solution to the sentiment classification problem in two sequential steps:

1. Develop and train the model using the training set data i.e., already labeled data.
2. Classifying the unlabeled or unclassified data based on the trained or skilled model.

Machine learning techniques are further classified into supervised and unsupervised techniques. To carry out sentiment analysis, typically the supervised machine learning techniques are used as we are dealing with subjective data. Supervised machine learning techniques highly depend on training data which are already labeled data unlike in the case of unsupervised machine learning techniques. Based on the provided training data, the classifier will classify the rest data i.e. test data. A large number of supervised machine learning algorithms such as Logistic Regression, Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, Maximum Entropy, and Bayesian Network are used for sentiment analysis [8]. Choice of an appropriate algorithm for selected data and domain is a crucial step.

EXPERIMENTS

DATASET

We performed the sentiment analysis on Twitter comments, extracting all the comments up to 20,000 of a tweet and classifying them based on their sentiment. Data collection is done using the 3 aforementioned steps.

The comments of a tweet are numbered 0 indexed. The sentiment score is normalized to -1 to 1, depicting the minimum and maximum sentiment. The value -1 represents that the given text or comment is extremely negative and +1 represents the sentiment is extremely positive.

EVALUATION METHODS

Sentiment analysis is the evaluation step. It is used to analyze the sentiment/emotion of the text and detect its polarity. But, there is a challenge here, that is, the user might write their opinions in different ways, some people express it straightforwardly, but others might add sarcasm to it which will be difficult for our model to detect the emotion. The solution we came up with is using VADER. Vader is Valence Aware Dictionary and sEntiment Reasoner. It is a rule based and lexicon sentiment analysis tool that is used especially for detecting sentiment/emotion in social media expressions. It is used to analyse the emotion of text that has both positive and negative polarity. VADER is a tool for calculating the amount of good and negative emotion in a text, as well as the intensity of that emotion. The main reasons behind using VADER are:

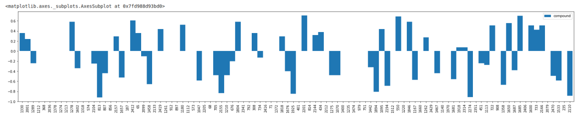
1) No training data is required for it.
2) It can understand the slangs, emoticons, punctuations, capital words, conjunctions and many more which will help a lot in sentiment analysis.

EXPERIMENTAL SETUP

We import a lot of different packages in the setup of the experiment. Notable ones are pandas, numpy, VaderSentiment, nltk, etc. We also imported a GitHub project in our experiment. It is called emoji, which was used to replace emojis with its English counterpart during text preprocessing. As we ran our experiment on google colab, we mounted the google drive on our code. We use twitter API to access the tweet replies. So we need to have a developer account and the relevant information while trying to request the API.

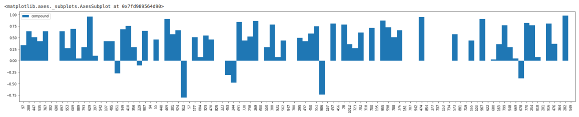
RESULTS

We have a variable size of dataset to perform sentiment analysis. For the showcase of results, we take a random of 100 replies and plot the graph of sentiment value. Due to the limitations of the API, we can only take the replies until 1 week before.



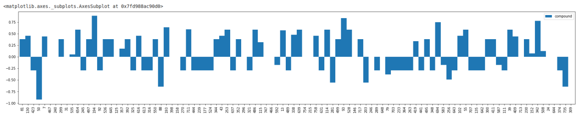
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd988d93bd0>
```

https://twitter.com/guardian/status/1467073775602745347



```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd989564d90>
```

https://twitter.com/kunalb11/status/1465516995235893250

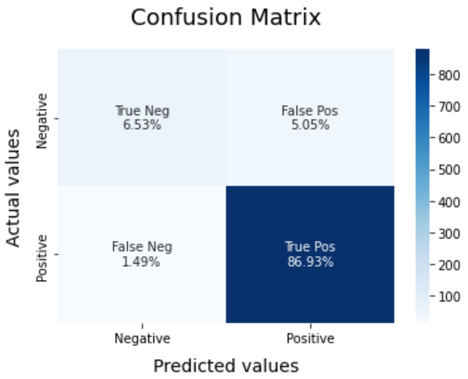


```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd988ac90d0>
```

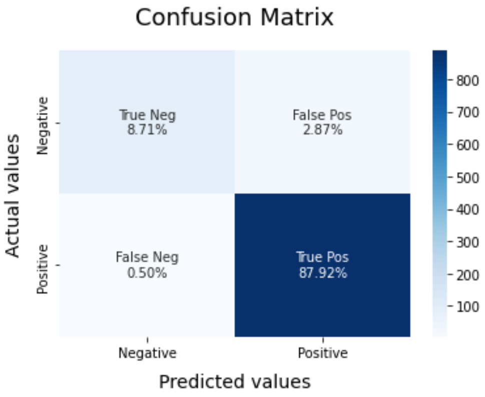
https://twitter.com/ColorsTV/status/1467537620217446404

We also did an analysis of 3 different models to perform sentiment analysis using the dataset we got during the initial sentiment analysis using the VaderSentiment. We took a tweet which has a huge reply count. The result of the sentiment analysis is stored in a csv file, which is in turn fed to the train model code. 5% of the dataset is taken as test data and the remaining is taken to train the model. For a dataset of 20,000, we got the following confusion matrix for each model(We take tweets with sentiment value as 0 as positive tweets)

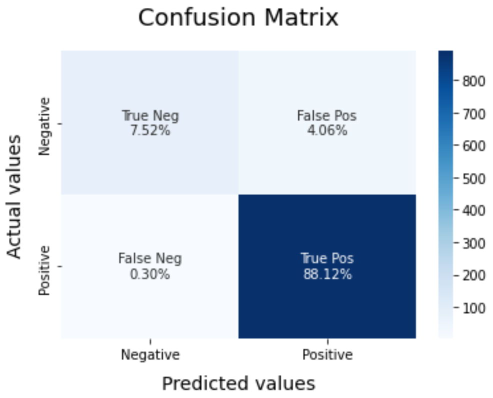
Model-1 Bernoulli Naive Bayes



Model-2 SVM (Support Vector Machine)



Model-3 Logistic Regression



DISCUSSION

We observe that after doing the sentiment analysis, the tweets with polarity value as 0 is relatively very high.

We observe in our model training that SVM model has the least false positives and false negatives. In the next place there is the Logistic Regression model. In the last place is the Bernoulli Naive Bayes model.

The extracted data set many 0 as their polarity values. They are neither positive nor negative but we considered them as positive and performed the training of the model.

This may have given rise to some imperfections in training of the model and it is to be rectified by removing the 0 polarity values from the data set that we have collected and training the model with the remaining data set.

It has been also discussed that “Does increasing the training data set increases the efficiency of the model?”

We have observed that the increase in the data set increases the efficiency of the model. We have trained the model using various scales of sizes of data set and results are inferred.

CONCLUSION

In this paper, we have firstly presented the detailed procedure to carry out sentiment analysis processes to classify highly unstructured data of Twitter into positive or negative categories. Secondly, we have discussed various techniques to carry out sentiment analysis on Twitter data including knowledge based techniques and machine learning techniques.

We have trained three models using the twitter data that we have extracted from twitterAPI and found their accuracy and the confusion matrix.

It is observed that SVM is the most accurate model followed by logistic regression. Naive Bayes had been the least accurate among the three.

REFERENCES

[1] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2] R. Parikh and M. Movassate, "Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques", CS224N Final Report, 2009

[3] Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[4] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.

[5] ZhunchenLuo, Miles Osborne, TingWang, ``An effective approach to tweets opinion retrieval", Springer Journal onWorldWideWeb, Dec 2013, DOI: 10.1007/s11280-013- 0268-7.

[6] "Three Cool and Inexpensive Tools to Track Twitter Hashtags", June 11, 2013. [Online]. Available http://dannybrown.me/2013/06/11/three-cool-tooltwitterhashtags/

[7] B. Gokulakrishna, P. Plavnathan, R. Thiruchittampalam, A. Perera and N. Prasath "Opinion Mining and Sentiment Analysis on a Twitter Data Stream", in *Int. Conf. on Advances in ICT for Engineering Regions,* 2012, p

[8] X. Chen, M. Vorvoreanu and K. Madhavan, "Mining Social Media Data to Understand Students' Learning Experiences", *IEEE Transaction*, 2014, vol. 7, no. 3, pp. 246-259.

[9] N. Kasture and P. Bhilare, "An Approach for sentiment analysis on social networking sites", *Computing Communication Control and Automation (ICCUBEA),* 2015, pp. 390-395.