Manobhaav : Twitter Sentimental Analysis

Aditi Chaudhary [1], Harshita Srivastava[2], Amritanshu Dwivedi[3], Upendra Mishra[4]

[1,2,3,4]*Department of Information Technology , IMS Engineering College ,Ghaziabad , India*

Abstract –*Social media websites have emerged as one of the platforms to elevate users' perceptions and influence the way any business is sold. People's perspective is very important to analyze how information dissemination impacts health on a large network such as Twitter. The conceptual analysis of tweets determines the unity and inclination of a large number of people to a particular topic, thing or thing. It is one of the fastest growing research languages in the natural language (NLP), making it challenging to keep track of all activities. The first purpose is to provide a way of analyzing sensitive points in the noisy twitter streams. The paper reports on the structure of sentiment analysis, producing a large number of tweets. The results distinguish the user's perception of the tweets positively and negatively.*

Key Words *– Sentiment Analysis, Twitter, Opinion mining Naïve Bayes.*

1. INTRODUCTION

In the last few years, there has been a significant increase in the use of microblogging platforms such as Twitter. Spurred on by this growth, companies and media organizations are increasingly looking for ways to introduce Twitter with information about what people think and feel about their products and services for Sentiment analysis. Whether it's a product or a movie, people's opinions are important, and it also affects people's decision-making process. The first thing a person does when they want to buy a product online, is to see the type of reviews and comments written by people. Social media such as Facebook, blogs, twitter have become places where people post their ideas on specific topics. The feeling of tweets on a particular topic has made many uses, including the company's stock market, movie reviews, psychology to analyze the human condition with various uses, and more. Ideally, it is a paradox to split conversations into positive, negative or neutral labels.Many people use social media sites to keep in touch with people and to keep up-to-date with current news and events. These sites (Twitter, Facebook, Instagram, google +) provide a platform for people to voice their opinions. For example, people are quick to post their reviews online as soon as they watch a movie and start a series of comments to discuss the acting skills displayed in the movie. This kind of information forms the basis for people to measure the effectiveness of not only the movie but also other products and to know whether they will be successful. This kind of big data on these sites can be used for marketing and social studies [1]. So, sentiment analysis has many applications and includes mood mining, solidarity, categorization and impact analysis Twitter is an online communication platform driven by 140 character limited tweets of character.Twitter's sentiment analysis includes the use of natural language processing to extract, to identify the content of the content. Sentiment analysis is usually done on two levels 1) the level of encounter and 2) the quality level. At a quantitative level, all textual analysis is done while at the right level, the character analysis is done [3]. However, making the analysis of the tweets presented is not an easy task. Many challenges are involved in relation to the unity, dictionary and language of tweets. They tend to be less structured and non-linguistic. It becomes difficult to interpret their meanings. In addition, the widespread use of nicknames, annotations, and vocabulary words is very common while online. The isolation of such words in each polarity becomes difficult for the affected natural processors.

2. LITERATURE REVIEW

This section summarizes some of the scholarly and research works in the field of Machine Learning and data mining to analyze sentiments on the Twitter and preparing prediction model for various applications. As existing social forums work up and down, information gets bigger and can be released for conversion for business purposes, social campaigns, marketing and other promotional strategies as outlined in [4]. The benefit of social media to know public perceptions and extract their emotions are considered by authors in [2] and explained how twitter gives advantage politicallyduring elections. They opted two stage approach for their framework preparing training data from twitter using mining conveying relevant features and further they applied

Naive Bayes classification algorithm for the accumulation of tweets. Naive Bayes methods are a set of supervised learning algorithms that are based on applying Bayes' theory to the "irrational" concept of conditional independence between both elements given the amount of phase difference.

The common approach found in almost all relevant research works constitutes data collection using Twitter API, preprocessing of data, filtering of data then approaches in feature extraction, classification and pattern analysis makes the distinction. Following on Twitter, the authors also predicted prejudice - positive, negative or neutral tweets by creating a classifier. In addition, they have used many algorithms and methods to determine the influence of the active business on the twitter patterns of users that display particular emotions,judgment on tweets sent by users.

3. SENTIMENTAL ANALYSIS

Our project involves the use of machine learning algorithms and natural language processing we can extract the data used for a document and try to classify it according to its smallness as positive, neutral or negative. A really useful analysis because we can get a general ideaof what a stock is selling, or predict stock markets for a given company such as, if most people think good of it, maybe its stock markets will go up, and so on.

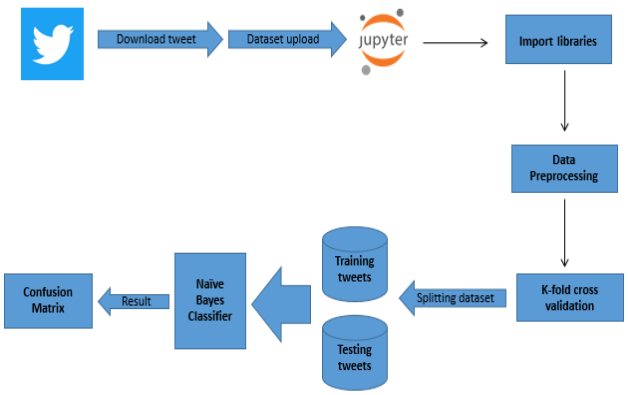


Fig. 1 : Modular Structure

A. Data collection:

Data in the form of raw tweets is retrieved by using the time twitter streaming API. The API requires us to register a developer account with Twitter and fill in parameters. Here we used a dataset of 1578612 tweets in english coming from two sources: Kaggle and Sentiment140. It is composed of four columns that are ItemID, Sentiment, Sentiment Source and SentimentText . We are only interested by the Sentiment column corresponding to our label class taking a binary value, 0 for negative tweet, 1 for positive tweet and the SentimentText columns containing the tweets in a raw format.

B. Data Processing:

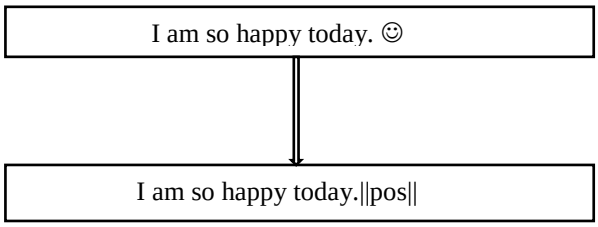
Now that we have a bunch of tweets and all the resources that can be useful, we can resize the tweets.It is very important because all the changes we will make during this process will directly impact the performance of the classifier.Thepreprocessing includes cleaning, normalization ,transformation, feature extraction and selection, etc. The result of preprocessing will be consonant and uniform data that are viable to maximize the classifier's performance.

Following are the steps –

EMOTICONS: Replacing emoticons by their sentiment polarity i.e.||pos|| and ||neg|| using the emoticon dictionary. To do the substitution, we pass through each tweet and by using a regex we find out if it contains emoticons, if yes they are substituted by their corresponding polarity.

	Smiley	Sentiment
0	:-)	1
1	:)	1
2	:D	1
3	:o)	1
4	:]	1

Fig. 2 : emoticons with their sentiments



URL: We substitute all URLs with the tag||url|| There is about 73825urls in the data set and we proceed as the same way we did for the emoticons.

56	57	0	Sentiment140	friends are leaving me 'cause of this stupid love http://bit.ly/ZoxZC
57	58	1	Sentiment140	go give ur mom a hug right now. http://bit.ly/azFwv
58	59	1	Sentiment140	Going To See Harry Sunday Happiness
59	60	0	Sentiment140	Hand quilting it is then...

56	57	0	Sentiment140	friends are leaving me 'cause of this stupid love ||url||
57	58	1	Sentiment140	go give ur mom a hug right now. ||url||
58	59	1	Sentiment140	Going To See Harry Sunday Happiness
59	60	0	Sentiment140	Hand quilting it is then...

Fig. 3 : Replacing URLs

UNICODE: For simplicity and because the ASCII table should be sufficient, we choose to remove any.unicode character that could deceiving for the classifier.

	ItemID	Sentiment	SentimentSource	SentimentText
1578592	1578608	1	Sentiment140	'Zu SpÃ¤t' by Die Ã„rzte. One of the best bands ever
1578593	1578609	1	Sentiment140	Zuma bitch tomorrow. Have a wonderful night everyone goodnight.
1578594	1578610	0	Sentiment140	zummie's couch tour was amazing....to bad i had to leave early

	ItemID	Sentiment	SentimentSource	SentimentText
1578592	1578608	1	Sentiment140	'Zu Spt' by Die rzte. One of the best bands ever
1578593	1578609	1	Sentiment140	Zuma bitch tomorrow. Have a wonderful night everyone goodnight.
1578594	1578610	0	Sentiment140	zummie's couch tour was amazing....to bad i had to leave early

Fig. 4 : Replacing Unicodes

HTML ENTITIES: HTML entities are characters reserved in HTML. We need to decode them in order to have characters entities to acknowledge them

CASE: The case is something that can appears useless but in fact it is really important for differentiating between proper noun and other kind of words. As we know,"Induction Motor" is the same thing that "induction motor", or "BSc" and "bsc". So reduction of all letters to lowercase should be normally done properly. In this project, for simplicity we will ignore it since we assume that it should not impact too much the classifier's performance.

TARGETS: The target related to usernames is grabbing the attention. We substitute all usernames/targets by the tag
||target||.

	ItemID	Sentiment	SentimentSource	SentimentText
45	46	1	Sentiment140	@ginaaa <3 go to the show tonight
46	47	0	Sentiment140	@spiral_galaxy @ymptweet it really makes me sad when i look at muslims reality now
47	48	0	Sentiment140	- all time low shall be my motivation for the rest of the week.

	ItemID	Sentiment	SentimentSource	SentimentText
45	46	1	Sentiment140	||target|| <3 go to the show tonight
46	47	0	Sentiment140	||target|| ||target|| it really makes me sad when i look at muslims reality now
47	48	0	Sentiment140	- all time low shall be my motivation for the rest of the week.

Fig. 5 : Replacing Targets

ACRONYM: We substitute all acronyms with their full forms. An acronym is a word formed from the initial letters of each compound term.At this point, tweets are going to be tokenized by getting rid of the punctuation and using split in order to do the process really fast.

```
1) lol => laughing out loud : 59000
2) u => you : 54557
3) im => instant message : 51099
4) 2 => too : 42645
5) gonna => going to : 23716
6) 4 => for : 18610
7) dont => don't : 18363
8) wanna => want to : 16357
9) ok => okay : 16104
10) ur => your : 12960
11) omg => oh my god : 12178
12) n => and : 10415
13) ya => yeah : 9948
14) gotta => got to : 9243
15) r => are : 8132
16) tho => though : 7696
17) tv => television : 6246
18) o => oh : 6002
19) kinda => kind of : 5953
20) pic => picture : 5945
```

Fig 6: Top 20 of acronyms in the data set of tweets with their

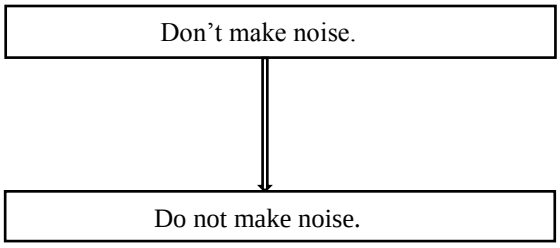


Fig 7 :Replacing Acronyms with their translations

NEGATION: We substitute all negations such as"not", "no", "never" by the tag ||not|| using the negation

dictionary in order to take more or less of sentences like "I don't like it". Here we will substitute "don't" by ||not|| and the word like will not be counted as positive.

	Negation	Tag
0	not	||not||
1	don't	||not||
2	doesn't	||not||
3	aren't	||not||
4	isn't	||not||

Fig. 6 : Replacing negation by tag

SEQUENCE OF REPEATED CHARACTERS: Now, we substitute repeated characters by two characters like (e.g.: "hiiiiiiii"= "hii") to keep the emphasized usage of the word.

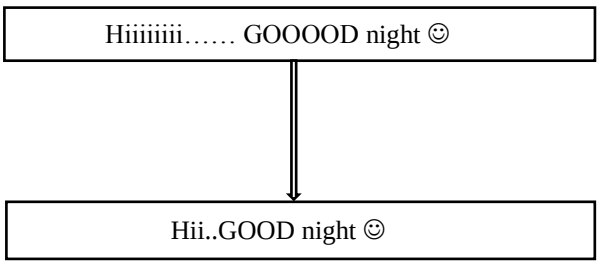


Fig. 7 : Replacing repeated characters .

4. RESULT

A confusion matrix helps to visualize and analyse how the model did during the classification and evaluate its accuracy. In our case we get about 156716 false positive tweets and 139131 false negative tweets. The confusionmatrix of the naive bayes classifier can be expressed using a color map where dark colors depicts high values and light colors depicts lower values as shown in the corresponding color map of the naive bayes classifier below

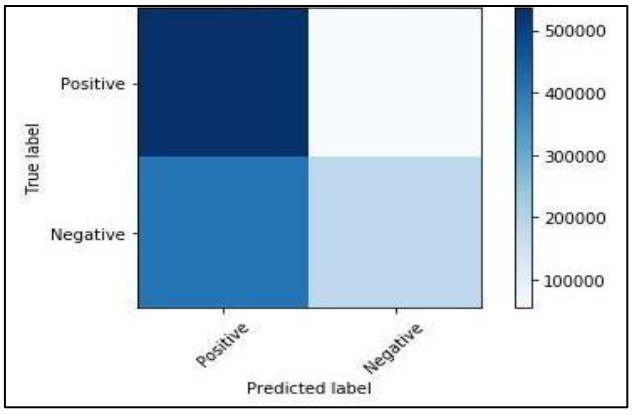


Fig. 7 : Confusion Matrix(Output)

5. FUTUREWORK

From future perspective, we would like to widen this project by implementing some machine learning algorithms for applications like election results, product ratings, movies' outcomes and running the project on clusters to expand its functionalities. Moreover, we would like to make a web application for users to input keywords and get analyzed results.

6. CONCLUSION

Twitter is enormous and a great source of unstructured and inconsistent data that can be processed to analyze interesting patterns and trends. In this project we tried to show the basic and simple way of classifying tweets into positive or negative category using Naive Bayes as baseline and how language models are related to the Naive Bayes and can produce better results. We could further improvise our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the naïve Bayes classifier, or trying another classifier all together.

ACKNOWLEDGEMENT

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend my sincere thanks to all of them.

We are highly indebted to Mr. Upendra Mishra for their guidance and constant supervision as well as for providing necessary information regarding the

project & also for their support in completing the project.

We would like to express our gratitude towards our parents & member of IMS Engineering College for their kind cooperation and encouragement which help us in completion of this project.

Many people, especially our classmates and team members itself, have made valuable comment suggestions on this proposal which gave us an inspiration to improve our project. We thank all the people for their help directly and indirectly to complete our project.

7. REFERENCES

1. Dr. Khalid N. Alhayyan& Dr. Imran Ahmad "Discovering and Analyzing Important RealTime Trends in Noisy.
2. J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, "ElectionInventive ComputationTechnologies(ICICT),Internationa lConference on, 2016, vol. 1, pp. 1–5.
3. M. Desai and M. Mehta, "Techniques for sentiment analysis of Twitter data: A comprehensive survey", 2016 International Conference on Computing, Communication and Automation (ICCCA), 2016.
4. Alexander Pak and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining". In Proceedings of the Seventh International Conference on Language.
5. Alec Go, RichaBhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision.
6. Jin Bai, JianYunNie. Using Language Models for Text Classification.
7. Apoorv Agarwal, BoyiXie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau.Sentiment Analysis of Twitter Data.
8. Fuchun Peng. 2003, Augmenting Naive Bayes Classifiers with Statistical Language Models.