International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206456

Twitter Sentiments Analysis Using Machine Learning

Saurabh Singh
Department of Computer Science and Engineering, IMS Engineering College, Ghaziabad ,Uttar Pradesh, India

Article Info
Volume 6, Issue 4
Page Number: 312-320
Publication Issue :
July-August-2020

Article History
Accepted : 20 July 2020
Published : 27 July 2020

ABSTRACT

Twitter sentiment analysis is the method of Natural Language Processing (NLP). In this project named Twitter sentiment Analysis we analyze the sentiments behind the twitter's tweet. We have three type of sentiment: Positive, Neutral and Negative. Analyzing the sentiments behind every tweet is the biggest problem in the early days but now it can be solved with the help of Machine Learning. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters and through the Twitter Sentimental Analysis we can analysis the mood of the person who tweet which can helps in the industries to analyze the market and their product reviews or we can know the sentiments behind the opinion on any topic on which the group of people tweet and through this we can find the final result that the people point on view on the particular topic, product and any other tweets suggestions.

Keywords : Machine learning, Twitter Sentiment Analysis, Natural Language Process, Data Minning, Bag of Words(BoG), Embedded layer, Naïve Bayes Classifier, Keras, Natural language toolkit(nltk).

I. INTRODUCTION

Sentiment Analysis is the process of determining the sentiment behind the tweet. whether a piece of written text(tweet) is positive, neutral or negative.

It is also referring as opinion analysis, it is a sub machine learning task where we want to determine which is the general sentiment of a given document. Using machine learning techniques and natural language processing we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative. It is a really useful analysis since we could possibly determine the overall opinion about a selling object, or predict stock markets for a given company.

Let's suppose There is an election in our country so government starts their campaigning and they want to analyses the people's reaction on their campaigning's advertisements and the tweets of their leaders or party members so they have to know the mood of the public on their actions for this they can analyze the sentiments of replies by the public on social media platforms and calculate the average that their action has been liked by public or not. Other example is like a company launch any product so they have to know the opinions of their buyers who used it. that, are they liked that product or not but they can read each and every reviews so here sentiment analysis plays a very important role in calculating the sentiments through the reviews given by the users in few seconds.

Sentiment analysis of public is highly critical in macro-scale socioeconomic phenomena like predicting the stock market rate of a particular firm. This could be done by analyzing overall public sentiment towards that firm with respect to time and using economics tools for finding the correlation between public sentiment and the firm's stock market value. Firms can also estimate how well their product is responding in the market, which areas of the market is it having a favorable response and in which a negative response (since twitter allows us to download stream of geo-tagged tweets for particular locations.

These are the steps used in Machine Learning for analyzing the sentiment.

- Stemming
- Tokenization
- Part of speech tagging
- Parsing
- Lexicon analysis (depending on the relevant context)

Prior Knowledge:

In every field of education, we need prior knowledge to understand an analyze that field very well, prior knowledge become base for successful understanding and analyses of any study. So, before we start to study the actual content of any paper. We have to understand the basic concepts related to the paper that will help us to understand and comprehend the paper very well.

Natural language processing (NLP)

Natural language processing is a subfield linguistic; computer science, information engineering, and artificial intelligence describes the interaction between human language and computers, in particular to program computers to process and analyze large amount of natural language data.

Examples of NLP that we use in our everyday life:

- Spell check
- Autocomplete
- Spam filter
- Voice text messaging
- Siri, Alexa or Google search engines

Data Mining:
Data mining is the method to find the useful patterns in large datasets involving methods at the interaction of machine learning, statistics, and database system in order to generate the new useful information from that datasets.

Bag of Words(Bog)
Bag of words model is the NLP technique of text modeling. Whenever we apply any algorithm in NLP, it works on numbers.

We cannot directly feed out text into that algorithm. Hence Bag or Words model is Used to preprocessing the text by converting it into a bag of words, which keeps a count of the total occurrences of most frequent used words.

Embedded layer:
The **Embedding layer** is used to create word vectors for incoming words.
The Embedding layers is defined as the hidden layers of a network. It must specify 3 arguments:

- **Input_dim:** this is the size of the vocabulary in the text data. As we know the maximum length twitter allows is 140 words. So, if your data is integer encoded to values between 0-139, then the size of the vocabulary would be 140 words.
- **output_dim:** It is the size of the vector space in which words will be embedded. It defines the size of the output vectors from the layers for each word.
- **Input_length:** it is the length of input sequence, as you would define for any input layers of a keras mode.

In this project we use **Word2Vec** methods of learning word embedding from text.

Stopswords:

Stopwords(these are the word we don't want to use in tweets after cleaning the text which are not relevant to predict the sentiments are Positive, Negative and Neutral.

These words may be like 'The', and etc. These are the word which did not give hint about the sentiments is Positive, Negative and Neutral.

Natural language Toolkit(NLTK):

It is a classic library in NLP which allow us to download the ensemble of Stopwords.

The Natural Language Toolkit (**NLTK**) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning.

Stemming:

It is the method of NLP use in cleaning the text. It transforms all the conjugates of the words for exam it treated loved & love same just simplify the tweets because both means the positive sentiments

luv	0.5732780694961548
loves	0.5732780694961548
loved	0.5373271703720093
amazing	0.5026600360870361
adore	0.4942743480205536
awesome	0.4598265290260315
loveee	0.4531649351119995
loooove	0.44260522723197937

Machine Learning

It is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it learn for themselves.

Machine learning are of three types:

1. Supervised leaning
2. Unsupervised learning
3. Reinforcement learning

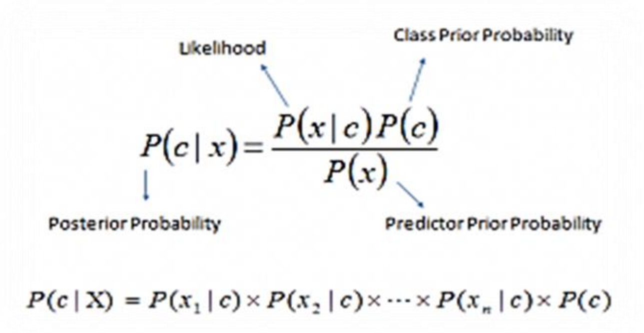
In this project we use **Naïve Bayes classifier** methods which are the part of Supervised learning .

Naïve Bayes Classifier:

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive)independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum likelihood training can be done by evaluating a closed form expression (mathematical expression that can be evaluated in a finite number of operations), which takes linear time.

It is based on the application of the Baye's rule given by the following formula:

Bayes theorem provides a way of calculating posterior P(c|x) from P(c) and P(x|c). Look at the equation below:



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Above,

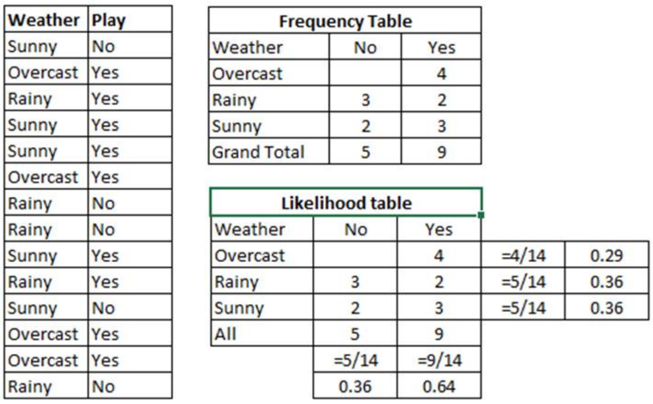
- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

How Naive Bayes algorithm works?

Let's understand it using it an example. Below I have a training data set of weather and corresponding target variable 'Play' (suggestion possibilities of playing). Now, we need to classify whether player will play or not based on weather condition Let's follow the steps to perform it.

Step 1: Convert the data set into a frequency table
Step 2: Create Likelihood table by finding the probabilities like Overcast

Probability = 0.29 and probability of playing is 0.64.



Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Step 3: Now, use Naive Bayes equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Problem: player will play if weather is sunny. Is this statement is correct?

We can solve it using discussed method of posterior probability.

P(Yes|Sunny) = P(Sunny|Yes) * P(Yes)/P(Sunny)
Here we have P(Sunny|Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(yes) =9/14 =0.64

Now, P (Yes|Sunny) = 0.33 * 0.64/0.36 = 0.60, which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

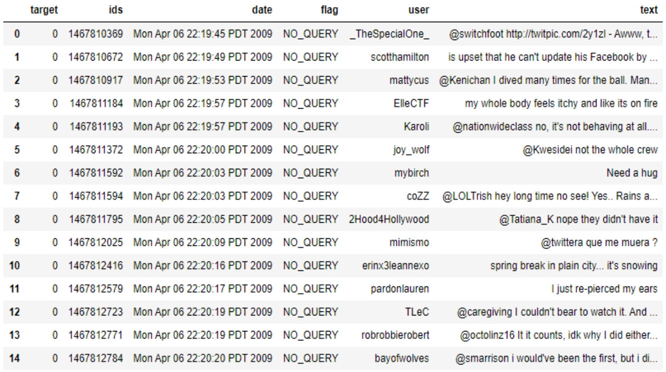
Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is we used in text classification through which we can analysis the sentiment behind the text(tweet).

Cross Validation:

Cross-validation is a technique to evaluate predictive models by dividing the original dataset into a training set and test set. Training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly divided into k equal size subsets. Of the k subsets, a single subset is taken as the validation data for testing the model, and the remaining k-1 subsets are used for training the model. The cross-validation process is then repeated k times (the folds), with each of the k subsets used exactly once as the validation data and average accuracy of k-folds is taken as final accuracy. In most experiments 10-fold cross validation technique issued. In10-fold cross validation all the instances of the dataset are used and are divided into 10 disjoint groups, where nine groups are used for training and the remaining one is used for testing. The algorithm runs for10 times and average accuracy of all folds is calculated.

II. METHODOLOGY

This Twitter sentiment Analysis Dataset is taken from Kaggle Website that consist 12894 of rows and 6 columns (target, ids, date, flag, user, text). Sample of dataset give in this figure:



	target	ids	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
5	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
6	0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
7	0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a...
8	0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
9	0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?
10	0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
11	0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren	I just re-pierced my ears
12	0	1467812723	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	TLeC	@caregiving I couldn't bear to watch it. And ...
13	0	1467812771	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	robrobbierobert	@octolinz16 It it counts, idk why I did either...
14	0	1467812784	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	bayofwolves	@smarrison i would've been the first, but i di...

Attributes of Dataset:

1- Target

0 – Negative Tweets

1 – Positive Tweets

2 – neutral Tweets

2-Ids

3-Date

4-Flag

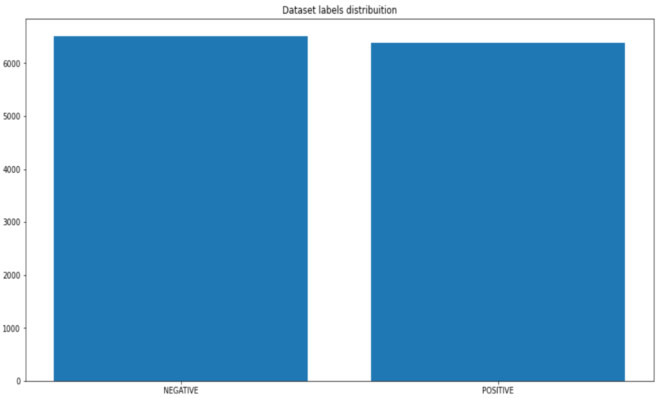
5-User

It consists the name of the users who tweet.

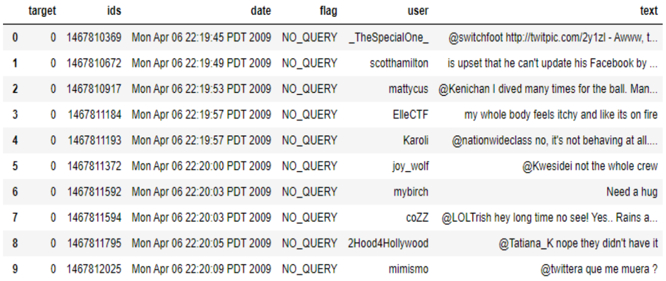
6-text

It consists all the tweet (positive, negative, neutral)

Graphical representation of Dataset:



Dataset without removing punctuation:



	target	ids	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
5	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
6	0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
7	0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a...
8	0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
9	0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?

Dataset after cleaning:

For cleaning the dataset, we apply following methods:

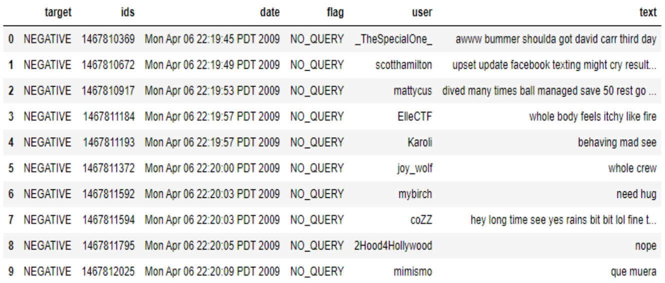
1-removing all the punctuations (letters expect [a-zA-Z])

2-Removing all the Stopwords(these are the word we don't want to use in tweets after cleaning the text which are not relevant to predict the sentiments are Positive, Negative and Neutral.

These words may be like 'The', 'and', 'is', 'are' etc. These are the word which did not give any hint about the sentiments is Positive, Negative and Neutral.

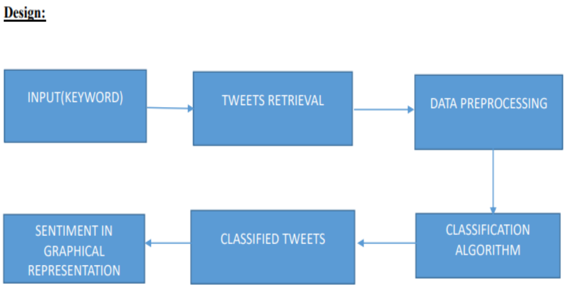
3-Make all the words of text in lower case.

4- Apply Stemming



	target	ids	date	flag	user	text
0	NEGATIVE	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	awww bummer shoulda got david carr third day
1	NEGATIVE	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	upset update facebook texting might cry result...
2	NEGATIVE	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	dived many times ball managed save 50 rest go ...
3	NEGATIVE	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	whole body feels itchy like fire
4	NEGATIVE	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	behaving mad see
5	NEGATIVE	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	whole crew
6	NEGATIVE	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	need hug
7	NEGATIVE	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	hey long time see yes rains bit bit lol fine t...
8	NEGATIVE	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	nope
9	NEGATIVE	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	que muera

Proposed Model:



Design:

INPUT(KEYWORD)

TWEETS RETRIEVAL

DATA PREPROCESSING

CLASSIFICATION ALGORITHM

CLASSIFIED TWEETS

SENTIMENT IN GRAPHICAL REPRESENTATION

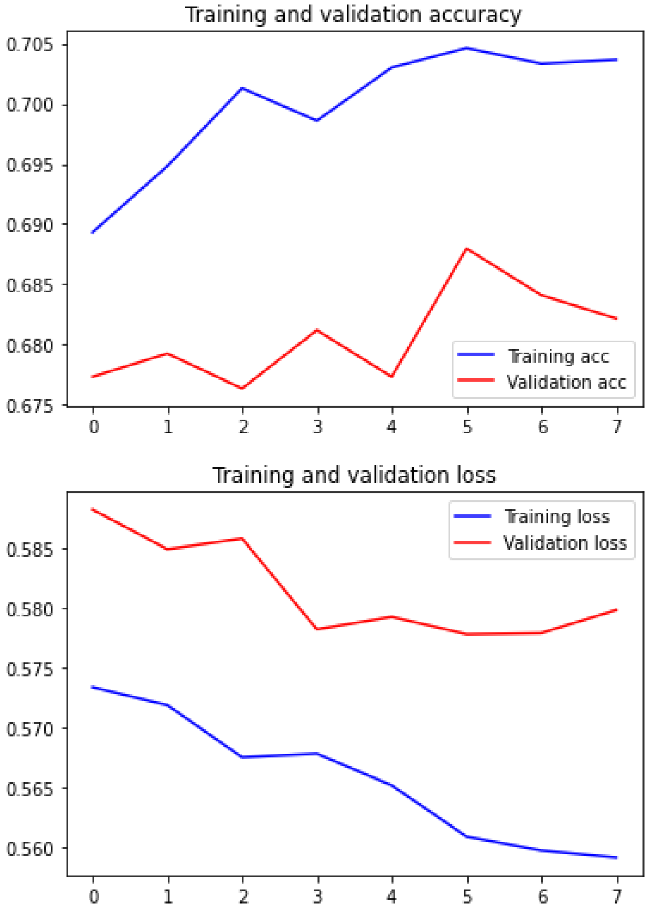
Observation:

To make the validation set, there are two main options:

- Split the training set into two parts (60%, 20%) with a ratio 2:8 where each part contains an equal distribution of example types. We train the classifier with the largest part, and make prediction with the smaller one to validate the model. This technique works well but has the disadvantage of our classifier not getting trained and validated on all examples in the data set (without counting the test set).
- The K-fold cross validation. We split the data set into k parts, hold out one, combine the others and train on them, then validate against the held-out portion. We repeat that process k times (each fold), holding out a different portion each time. Then we average the score measured for each fold to get a more accurate estimation of our model's performance.

Epoch	Accuracy	Training_Loss	Validation_Loss	Validation_accuracy
Epoch- 1	0.6893	0.5734	0.5882	0.6773
Epoch- 2	0.6948	0.5719	0.5849	0.6793
Epoch -3	0.7013	0.5675	0.5858	0.6764
Epoch-4	0.6986	0.5678	0.5782	0.6812
Epoch-5	0.7030	0.5652	0.5793	0.6773
Epoch-6	0.7046	0.5689	0.5778	0.6880
Epoch-7	0.7033	0.5597	0.5779	0.6841
Epoch-8	0.7037	0.5591	0.5798	0.6822

Graphical representation of accuracy, validation_loss and validation_accruracy:

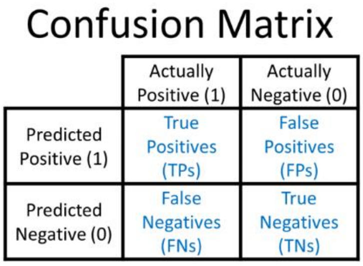


We split the training data into 8 folds and cross validate on them using scikit learn as shown in the figures above. The number of K-folds is arbitrary and usually set to 8 it is not a rule. In fact, determine the best K is still an unsolved problem but with lower K: computationally cheaper, less variance, more bias. With large K: computationally expensive, higher variance, lower bias.

Observation:

Confusion Matrix:

Confusion matrix it the table that is used to describe the performance of the model on the set of the test data. It is also known as error matrix. In this the row of the matrix represented the instances in a predicted class and column of the matrix represent the instances in an actual class (or vice versa).

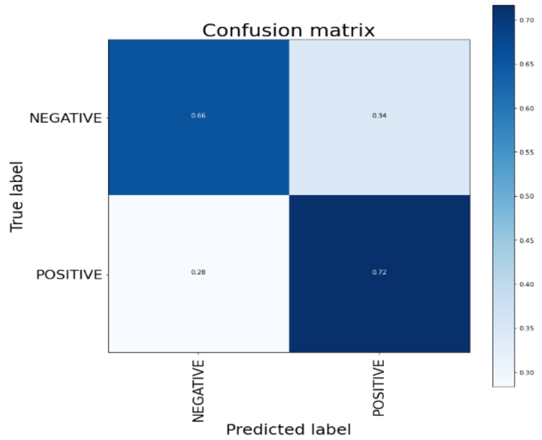


Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

- **True Positive (TP):** Observation is Positive, and is predicted to be positive.
- **False Negative (FN):** Observation is positive, but predicted negative
- **True Negative (TN):** Observation is negative, and is predicted to be negative.
- **False Positive (FP):** Observation is negative, but is predicted positive.

Confusion Matrix For Twitter Sentiment Analysis:



Accuracy:

Accuracy is the ratio of total number of Correct predictions to the total number of input samples. It is also known as Classification Rate. This is the formula of accuracy/Classification Rate.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Accuracy for Twitter sentiment analysis data is = 0.6991081833839417

Recall:

Recall can be defined as the ratio of the total number of correctly classified positive example divide to the number of positive examples.

High recall indicates the class is correctly recognized.

Recall = TP/(TP+FN)

Recall for Twitter sentiment analysis data is:

For Negative	0.66
For positive	0.72

Precision:

Precision is the Total positive divided Total number of positive prediction (TP+FP)

Precision = TP/(TP+FP)

Precision score for Twitter sentiment Analysis dataset is:

For Negative	0.72
For positive	0.66

F1-Score:

F1- score is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. It is calculated as

F-1 Score=2*(Recall*Precision)/(Recall+Precision)

F1 score for Twitter sentiment Analysis dataset is:

For Negative	0.69
For positive	0.68

Macro_average:

In this method we Just take the average of the precision and recall of the system on different sets.

1-**Macro_average for Precision:**

Just like we have 8 epochs so for the macro_average precision we take average of all the precision, for example

(P1+P2+P3+P4+P5+P6+P7+P8)/8

Macro_average for precision of our dataset is : 0.69

2-**Macro_average for recall:**

For this we take average of all the recall for each epoch (8 epochs).

(R1+R2+R3+R4+R5+R6+R7+R8)/8

Macro_average for recall of our dataset is: 0.69

3-**Macro_average for f-1 Score:**

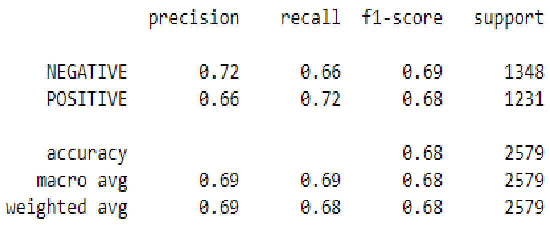
For this we take average of all the f-1 Score for each epoch (8 epochs).

(FS1+FS2+FS3+FS4+FS5+FS6+FS7+FS8)/8

Macro_average for F-1 score of our dataset is: 0.68

III. RESULT

After Implementing the Machine learning Algorithm or method, the table in below figure represent the Accuracy, Precision, F-1 Score, macro_average, Weighted_average.



```
              precision    recall  f1-score   support

    NEGATIVE       0.72      0.66      0.69      1348
    POSITIVE       0.66      0.72      0.68      1231

    accuracy                           0.68      2579
   macro avg       0.69      0.69      0.68      2579
weighted avg       0.69      0.68      0.68      2579
```

Testing for Tweets:

IV. CONCLUSION

Machine Learning is a hot topic for the industries. In this project we are tried to analyze the sentiments of the tweets. But we are still far to detect the sentiments of corpus of texts very accurately of the complexity in the English language and even more if we consider the other countries languages like chines.

In this project we tried to show the basic way of classifying tweets into positive, neutral and negative category using Naïve Bayes as baseline and how language models are related of the Naïve Bayes and can produce better results.

We could further improve our class classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the Naive Bayes Classifier.

V. ACKNOWLEDGEMENT

I have completed this work under the guidance of Dr. Pankaj Agarwal (Professor and head) & Ms. Sapna Yadav (Assistant Professor), Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad. I have been doing Summer Internship of Machine Learning and Deep Learning under their mentorship and worked with various machine learning and Deep learning Projects. This work has been assigned by my instructors to showcase my efforts and learning that I have obtained. They believed on me and showed me this opportunity to write a paper on my project on Twitter Sentiment Analysis. I express my heartfelt thanks to my mentors who have encouraged me, guided me throughout the project work. I really appreciate their cooperative support and moral boosting that they continuously provided me. I even thank my parents for their support and encouragement. My mentors as well as my parents both believed on my potential to successfully complete this paper. I even show my gratitude towards 'IMS Engineering College' for this wonderful opportunity. Any omission in this brief acknowledgement doesn't show my lack of gratitude.

VI. REFERENCES

[1]. Alexander Pak, Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

[2]. Twitter sentiments analyze dataset from Kaggle

[3]. Jin Bai, Jian-Yun Nie. Using Language Models for Text Classification.

[4]. AnalyticsVidya: For Naïve Bayes, (https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/)

[5]. Twitter sentiment analyses report on www.cse.ust.hk

[6]. Natural language processing from Wikipedia

Author

Saurabh Singh is B.Tech 3rd year student in the department of Computer Science & Engineering at IMS Engineering College, Ghaziabad , UP , India . His areas of interest is programming in python, Machine Learning, Deep Learning. Her research interests are Machine Learning and Deep Learning.

Cite this article as :

Saurabh Singh, "Twitter Sentiments Analysis Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 312-320, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206456
Journal URL : http://ijsrcseit.com/CSEIT206456