**Twitter Sentiment Analysis Using machine learning**

# Harsh Dubey

@harsh.dubey1228@gmail.com

*UG Student at Amity University Madhya Pradesh*

1. **ABSTRACT:**

In this project I am doing some basic NLP means processing the text data to machine learning

To predict the output. So, we just taken two date sets one for the training purpose and other one for the testing purpose. This project is a very vast topic and because I am a beginner in machine learning so I it was a very difficult for me to complete this. In this project we focused on the tweets of people to analyze that the tweet is positive negative or neutral we performed some basic algorithm and some advanced algorithms to train the machine for predication of the output. Using machine learning algorithms like logistics regression, SVM, decision tree, confusion matrix etc.

**2. Introduction:**

Nowadays, internet is taking the world and machine learning is the future of our world so to performing well in future we are developing algorithms for machine learning to learn by its self we don’t need to be programmed every time. In this project we are doing text analysis on data set and making model for predication. We know to express our feelings and thoughts we developed some applications like Facebook, WhatsApp, Twitter, Snapchat, and Telegram. All these are the most common ways to communicate each other, and all the feeling and their thoughts are In the form of text or images but our topic focus on text analysis so we are going to analyze the text data. We are just going to analyze the tweets and predict that whether the tweet is negative, positive, or neutral.

form of text or images but our topic focus on text analysis so we are going to analyze the text data. We are just going to analyze the tweets and predict that whether the tweet is negative, positive, or neutral. There are many ways to do this project, but we have to apply many algorithms so we will do in a way that we can convert them to algorithms.

So, what we have to do in this project is to first download the data set and performing some basic data cleaning and the data preprocessing to achieve the neat and clean data. After applying data preprocessing, we will head on to NLP.

Now what is NLP?

NLP is the Natural language pre-processing in this we build the algo which understand by machine automatically as humans do it is a branch of Computer Science and AI (artificial Intelligence) giving the ability to computers Understand the text or spoken language by own not to be Programmed every time for this. Some of the basic task of NLP is Speech recognition **Word sense disambiguation, Named entity recognition,** **Sentiment analysis etc.**

### LITERATURE REVIEW

### In this literature review is made to analyse the Domain. In some past year, the interest of sentiment analysis is increasing very fast, and this sentiment analysis can be done by 2 different approaches first is Supervised learning and second one is unsupervised learning. In Supervised approach it required a training labelled data and in Unsupervised we are using only input data to train the model for prediction. Using predefined rules, the citation sentences are annotated as positive, negative and neutral. For the purpose of annotations, human annotators are required. While in the case of an unsupervised approach, there is no need for labelled training data. Instead, there is a need for sentiment lexicon to assign polarities to citation sentences. This approach is very difficult because it requires different varieties of a lexicon for different genres. From the literature review, we found that many researchers used supervised while others relied on unsupervised approach. In this paper I applied various things just to make predictions like reshaping dataset, labelling the data set to positive comments and negative comment and after that plotting the graph between positive and negative sentiment to build relationship between data.

### after that extracting the hashtags from the data set. From this we can set the parameters for what we must predict. Means after extracting the hashtags we have a particular direction to work with it.After that we created a word cloud which can help us to learn which word is high in frequency and what is in low frequency. So in this paper we are also applied linear regression, logistic regression, SVM, lasso and ridge.We simply fit and train the data set and predicted output for the same and also drawn confusion matrix and roc curve for each algorithm.

**4. Data set:**

The data set is available on Kaggle and it available in two parts the first part is the train data set and other is the test data set.

The train data set contain 3 row and 31962 columns and test data set contain 2 row and 17197 columns.

Data set contain **id** (which is unique of every tweet) **label** (which defining the positive and negative tweet) and then **Tweets** (these are in the form of text).

Text

Description automatically generated with low confidence

**5. Sentiment analysis:**

It is the process of analyzing sentiment of text whether the text contain a positive sentiment or a negative sentiment. By this we can train our model in a way that the model can predict the sentiment of text. We will be doing all this by the help of NLP (natural language processing)

In this project we are doing some steps to prediction.

1. Collecting Data:
2. Preparing the data
3. Converting data set
4. Training the model
5. Making prediction

These steps we are performed in our project.

And some are the algorithms performed by me is:

1. Logistic regression
2. Word cloud
3. Random Forest
4. Confusion matrix
5. Svc
6. Linear Regression
7. Lasso
8. Ridge
9. Roc curve

**5.1 collecting the data:**

In this step we are collecting the data from the different websites, and I collected the data set form the Kaggle.com. Selecting a perfect data set is play a very important role because what we do in machine learning all the algorithm to build the model is perform on the data set. So selecting the data set is an difficult and important task.

**5.2 Preparing the data:**

In this step we are doing some pre processing steps like removing special characters and dropping the column and row if they are showing the NAN value and also Removing Stop Words. All this we are doing to increase the accuracy of the model. Because all these are special character will create noise in data which can be harmful for the model.

Step 1.

We can do this by code:

Text

Description automatically generated with low confidence

Text

Description automatically generated

Step 2.

After removing the noisy data we will check positive and negative comments in the data set to analyse the data and train the model . after checking positive and negative tweet will print a graph for the positive and negative comments for better understanding.

Chart, bar chart

Description automatically generated

Fig 2. Graph for positive and negative comments (0= positive and 1= negative)

**Step 3.**

After checking the positive and negative comments, now we are checking frequency of training and testing data set.

Chart, histogram

Description automatically generated **Step 4.**

Now we are adding a column called Len which is representing the length of each row containing the tweets or comments.

Fig 1. Graph representing the testing and training model frequency.

(Pink = train and orange =test)

**Step 5.**

Now we are Groupby the len and label (positive and negative comments )and plotting graph between them.

Chart, histogram

Description automatically generated fig 3. For variation of length of text data

A picture containing chart

Description automatically generated

Fig 4. Graph representing the most frequent word used in data set (top 30) here the blue part is the frequency and x label is word.

Step 6.

now we are going to create a word cloud with the help of WordCloud library this will give form of two types negative and positive

this word cloud builds in the form of reviews

Text

Description automatically generated

Fig 5. Word cloud showing all positive comments as a frequency will raise the font will also bigger.

Step 7.

Now in this step we will extract the hashtags and label it and then draw a graph for all hashtags present int the set this will help us to know more about the data set and to give a little idea that how we will train our model.

Text, chat or text message

Description automatically generated

Fig 6. Word cloud showing all neutral comments as a frequency will raise the font will also bigger.

Text

Description automatically generated

Fig 7.Word cloud showing all neutral comments as a frequency will raise the font will also bigger.

Chart, bar chart

Description automatically generated

Fig7. For positive Hashtags we printed the Graph

Chart, bar chart

Description automatically generated

Fig8. For negative Hashtags we printed the Graph

**5.3 converting the model:**

From the sklearn library we are importing the feature extraction of text. means it is a conversion of that format which machine learning algorithm will understand.

Count vectorizer: it is a tool by scikit learn library in python the use of CountVetorizer is to convert the text into token or to count each word and convert to words.

By using this library, we plotted a graph which shows the most uses of word in whole data set.

After this we will tokenize the data set Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens **help in understanding the developing of model.**

Then in next step we will remove unwanted patterns from the data by the “stopwords”.

Now we will start stemming. Stemming is **the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language**."

And In this step we will do BOW method. Bag of Words (BOW) is **a method to extract features from text documents**. These features can be used for training machine learning algorithms. It creates a vocabulary of all the unique words occurring in all the documents in the training set

**5.4 training the model:**

In this step we will train the model from the sklearn.model\_selection library we will import the train\_test\_split this will help in to split the data set for training and testing purpose. By this we can easily train our model.

After training the model we can apply different type of algorithm to get the accuracy.

5.5 Applying algorithms to Model:

1. **RANDOM FOREST CLASSIFICATION**

Text

Description automatically generated

**The accuracy we get after training and testing**

**The accuracy we get after training and testing.**

**Confusion matrix:**

A picture containing chart

Description automatically generated

Chart, line chart

Description automatically generated with medium confidence

1. **Logistic Regression.**

Text

Description automatically generated

**Confusion matrix:**

A picture containing graphical user interface

Description automatically generated

Chart, line chart

Description automatically generated

# **Decision Tree**

**Confusion matrix:**

Chart

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

# **4. SUPPORT VECTOR MACHINE**

**Confusion matrix:**

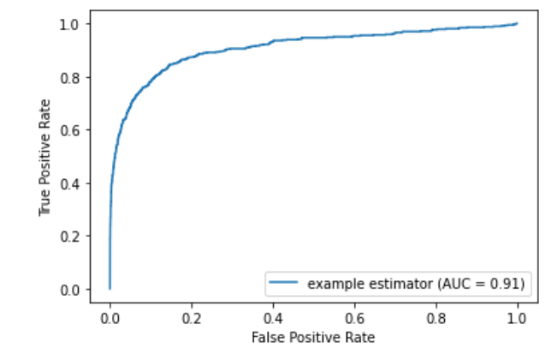
Chart

Description automatically generated

Chart, line chart

Description automatically generated

# **5. Linear Regression**

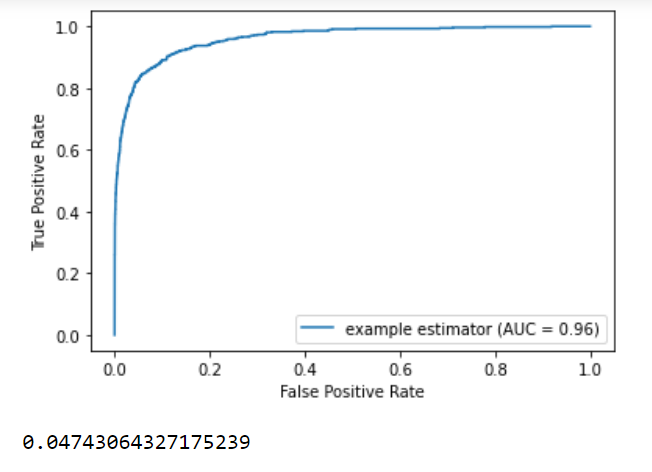


**6. Lasso:**

A picture containing chart

Description automatically generated

**7. Ridge**



Diagram

Description automatically generated

Fig 9. Sentiment Analysis Architecture

LEVELS OF SENTIMENT ANALYSIS

Diagram

Description automatically generated

1. **EVALUATION OF SENTIMENT CLASSIFICATION**

The performance of sentiment classification can be evaluated by using four indexes calculated as the following equations:

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

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

F1 = (2×Precision×Recall)/(Precision+Recall)

In which TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negativeinstances, the number of false positive instances and the number of true negative instances, as defined in the table 1

Table

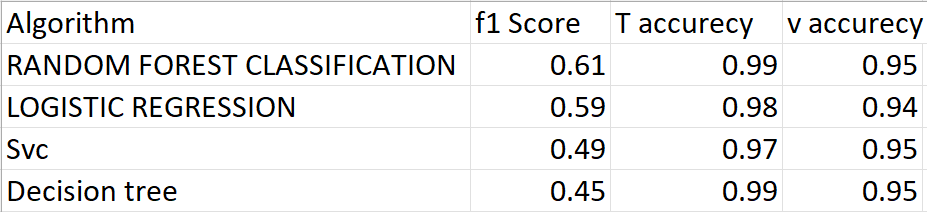
Description automatically generated

1. **APPLICATIONS OF SENTIMENT ANALYSIS**
2. As a website reviewer
3. Spam detector
4. In business intelligence
5. Smart home
6. Customer support
7. Product analysis
8. **CHALLENGES IN SENTIMENT ANALYSIS:**

Sentiment Analysis is a very challenging task. Some challenges I faced that are.

1. I**dentifying subjective parts of text:** The same word can be treated as subjective in one case, or an objective in some other. This makes it difficult to identify the subjective portions of text.
2. **Domain dependence:** The same sentence or phrase can have different meanings in different domains.
3. **Noisy data:** so many comments having a same meaning leads to a challenge
4. **Converting the data set:** many errors comes during this converting period.
5. **Displaying roc curve:** it a very challenging for me.
6. **RESULTS AND DISCUSSION**

We used the data set available on the Kaggle.com. we used NLP and many algorithms with the ROC curve and scores.



Precision score: 0.9416843949443123

Recall score: 0.96676534

1. **CONCLUSION:**

In this term paper we survey my Twitter sentiment analysis project on machine learning using different algorithms like

1. Random Forest

2. linear regression

3. logistic regression

4. SVC

5. lasso

6. ridge

Our model is done all the algorithms.

Table

Description automatically generated

Precision score: 0.9416843949443123

Recall score: 0.96676534

Our conclusion is that our model can predict the sentiment, and this will help in

Various situation like in hotels, hospitals, and many more places.

11. 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". University, Technical Paper,2009 Stanford

[4] L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38

[7] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010

[8] Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, <http://doi.ieeecomputersociety.org/10.1109/MDM.2013>.

[9] Pablo Gamallo, Marcos Garcia, “Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[10] Neethu M,S and Rajashree R,” Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661

[11] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

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

[13] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[14] ZhunchenLuo, Miles Osborne, TingWang, An effective approachto tweets opinion retrieval", Springer Journal onWorldWideWeb,Dec 2013, DOI: 10.1007/s11280-0130268-7.

[15] Liu, S., Li, F., Li, F., Cheng, X., &Shen, H.. Adaptive cotraining SVM for sentiment classification on tweets. In Proceedings of the 22nd ACMinternational conference on Conference on information & knowledgemanagement (pp. 2079-2088). ACM,2013.

[16] Pan S J, Ni X, Sun J T, et al. “Cross-domain sentiment classification viaspectral feature alignment”. Proceedings of the 19th internationalconference on World wide web. ACM, 2010: 751-760.

[17] Wan, X..“A Comparative Study of Cross-Lingual SentimentClassification”. In Proceedings of the The 2012 IEEE/WIC/ACMInternational Joint Conferences on Web Intelligence and IntelligentAgent Technology-Volume 01 (pp. 24-31).IEEE Computer Society.2012

[18] Socher, Richard, et al. "Recursive deep models for semanticcompositionality over a sentiment Treebank." Proceedings of theConference on Empirical Methods in Natural Language Processing(EMNLP). 2013.

[19] Meng, Xinfan, et al. "Cross-lingual mixture model for sentimentclassification." Proceedings of the 50th Annual Meeting of theAssociation for Computational Linguistics Volume 1,2012

[20] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M..“Lexicon basedmethods for sentiment analysis”. Computational linguistics, 2011:37(2), 267-307.

[21] Li, S., Xue, Y., Wang, Z., & Zhou, G..“Active learning for cross-domainsentiment classification”. In Proceedings of the Twenty-Thirdinternational joint conference on Artificial Intelligence (pp. 2127-2133).AAAI Press,2013

[22] Bollegala, D., Weir, D., & Carroll, J.. Cross-Domain SentimentClassification using a Sentiment Sensitive Thesaurus. Knowledge andData Engineering, IEEE Transactions on,

[23] Pang, B.and Lee, L. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”. 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278. [24] V. M. K. Peddinti and P. Chintalapoodi, “Domain adaptation in sentiment analysis of twitter,” in Analyzing Microtext Workshop, AAAI, 2011.