# MUET LOGO

# 

**Data Science and Analytics**

**Sentiment Analysis using Machine learning**

**Submitted by:**

Bilawal Mehmood (18CS26)

Shifa Ishfaque (18CS34)

Mahnoor Bhurio (18CS16)

**Submitted to:**

Dr. Sanam Narejo

Department of Computer Systems Engineering

Mehran University of Engineering and Technology, Jamshoro.

**Table of Contents**

1. Abstract ……………………………………………………………………
2. Introduction ……………………………………………………………….
3. Literature Review………………………………………………………
4. Objective of the Project …………………………………………………...
5. Methodology for implementation (Formulation/Algorithm)…………....
6. Implementation Details …………………………………………………….
7. Results/Sample output ……………………………………………………..
8. Conclusion……………………………………………………………………

**Abstract**

This project addresses the challenge of sentiment analysis in Twitter, which involves categorizing tweets as favorable, negative, or neutral. Twitter is an online microblogging and social networking website that allows users to post 140-character status updates. It is a rapidly growing site with over 200 million registered users, of which 100 million are active users, and half of them log in daily, resulting in almost 250 million tweets per day. We expect that through analyzing the sentiments conveyed in tweets, we would be able to obtain a reflection of popular sentiment. Analyzing public mood is useful for a variety of applications, including corporations attempting to determine the market response to their products, predicting political elections, and anticipating macroeconomic events such as stock exchange. The goal of this project is to create a functional classifier that can accurately and automatically classify the sentiment of an unknown tweet stream.

**Introduction**

The use of microblogging networks such as Twitter has increased dramatically in recent years. Companies and media organizations are increasingly looking for ways to mine Twitter for information about what people think and feel about their products and services as a result of this development. Twitter (twitrratr.com), Tweet feel (www.tweetfeel.com), and Social Mention (www.socialmention.com) are just a handful of the companies that market Twitter sentiment research as one of their services.

The online medium has become an important way for people to express their thoughts, and social media provides a wealth of opinion information. The polarity of opinions, such as positive, negative, or neutral, can be determined via sentiment analysis by evaluating the text of the opinion. Sentiment analysis has proved valuable for businesses in obtaining customer feedback on their products, predicting election outcomes, and obtaining feedback from movie reviews. Sentiment research provides useful information for businesses making future decisions.

Sentiment analysis is a vast topic of natural language processing that deals with the computational examination of text-based views, attitudes, and emotions. Sentiment Analysis (SA) or Opinion Mining (OM) seeks to learn about people's sentiments and emotions regarding a particular entity. Individuals, events, or concepts can all be represented by the entity. An enormous amount of study has been conducted in the field of sentiment analysis. However, the majority of them concentrated on categorizing formal and longer pieces of text data, such as reviews. Because of the widespread popularity of social networking

Twitter is an innovative microblogging service that debuted in 2006 and now has over 550 million users 1. This service refers to user-created status messages as tweets. Twitter's public timeline displays tweets from all users globally and is a comprehensive source of real-time information. The first purpose of microblogging was to deliver personal status updates. However, the current situation is unexpectedly witnesstweets spanning everything under the sun, from current political problems to personal experiences. Add to the list with movie reviews, vacation experiences, current happenings, and so on. Tweets (and microblogs in general) differ from reviews in their fundamental nature. While reviews are distinguished by formal writing patterns and are summaries of authors' thoughts, tweets are more casual and limited to 140 characters of text. Tweets provide businesses with an additional channel for gathering feedback. Sentiment analysis to investigate items, movie reviews, and so on can help buyers make decisions before making a purchase or going to the movies.

Businesses utilize this area to explore public opinion about their company and products, as well as to analyses consumer satisfaction. Organizations use this data to get input on freshly released goods, which helps to improve future design. Various approaches, including as machine learning (ML) algorithms, sentiment lexicons, hybrid approaches, and so on, have been shown to be useful for sentiment analysis on formal texts. However, their effectiveness in extracting sentiment from microblogging data will need to be investigated. A careful examination of tweets demonstrates that the 140 character limit restricts the vocabulary used to convey the sentiment. The hyperlinks that are frequently present in these tweets, in turn, limit the vocabulary size. The various domains listed would undoubtedly present training challenges. On the other hand, the enormous volume of data available from microblogging websites on many domains is unmatched to other data repositories. Microblogging language is distinguished by expressive punctuation that conveys a wide range of emotions. Bold lettered sentences, exclamations, question marks, quoted material, and so forth provide opportunities for sentiment extraction. The proposed work takes a novel approach to twitter data by combining a modified polarity lexicon learned from product reviews in the areas under consideration, tweet specific attributes, and unigrams to develop a classifier model using machine learning techniques.

## Literature Review

Because of the increasing rise of e-commerce shopping, online customer reviews play an important part in product sales. A significant study topic has evolved around the subject of how to extract the best and most accurate approach while also categorizing customers' written evaluations as negative or positive.

Main focus on data analysis of mobile reviews data set Single and Randhawa [1]. They concentrate on data analysis of a data set of mobile reviews. They discover the connection between various features. Can perform positive and negative sentiment classification, which is beneficial to both consumers and manufacturers. They use the unstructured data to perform pre-processing on mobile phone evaluations for sentiment analysis. After cross-validation, they obtained 84.87% accuracy using the Support Vector Machine (SVM) Machine Learning Model.

In Research [2], they used the Support Vector Machine (SVM) Classification Technique to identify text from smartphone product reviews as favorable or negative. Precision, recall, and F-measure were used to assess the model's performance. The projected Model was shown to be highly accurate**.**

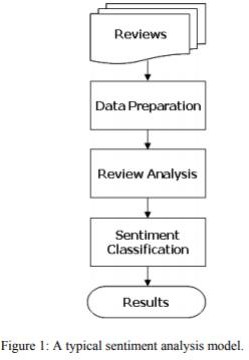
In Research [3], Deep learning approach proposed for sentiment analysis of smartphone reviews using the Twitter Corpus. Convolutional Neural Networks were used for the Deep Learning technique. (CNN). They also employed Machine Learning Models such as Nave Byes and Support Vector Machine (SVM). They discovered that deep learning approaches are more efficient than machine learning techniques.

Main proposed of Souza and Oliviera[4] a sentiment analysis technique based on Deep Learning. They classified the movie reviews using several supervised machine learning models and compared the findings using Convolutional Neural Networks (CNN), a Deep Learning Technique. They found that deep learning model CNN gives reliable performance.

In Research [5], learning technique proposed for sentiment analysis of Amazon Product Reviews on Mobile Phones. They employed the word2vec word-embedding technique in conjunction with the CNN Model to categories Mobile Phone Reviews as Positive or Negative. They used Google's pre-trained word2vec model to convert text to 300-dimensional word vectors. If words were not accessible in the pre-trained model, the word vectors were generated by randomly selecting values between -0.25 and +0.25. The CNN Model has been trained to predict the sentiment of new phone reviews. A comparison is performed, and it is discovered that the proposed approach has a higher accuracy of 0.9123 than the Machine learning model. They used the deep learning libraries Keras and Tensorflow to create the CNN Model. A comparison is performed, and it is discovered that the proposed approach has a higher accuracy of 0.9123 than the Machine learning model. They used the deep learning libraries Keras and Tensorflow to create the CNN Model.

In Research [6], they employed a unigram and bigram method, as well as machine learning classifiers like Naive Bayes, maximum entropy classification, and SVM. They evaluated classifier performance using precision, recall, and the F-measure. It was discovered that bi-grams produced good results.

**Objective of the Project**

* Analyze sentiment on dataset from document level (review level). Categorization or classification of opinion sentiment into-
* Positive
* Negative

**Methodology**

**DATA COLLECTION**: Data which means product reviews collected from Github from May 1996 to July 2014. Each review includes the following information: 1) reviewer ID 3) rating 4) time of the review 5) helpfulness 6) review text.

Every rating is based on a 5-star scale, resulting all the ratings to be ranged from 1-star to 5-star with no existence of a half-star or a quarter-star. SENTIMENT SENTENCE **EXTRACTION & POS TAGGING**: Tokenization of reviews after removal of STOP words which mean nothing related to sentiment is the basic requirement for POS tagging. After proper removal of STOP words like “am, is, are, the, but” and so on the remaining sentences are converted in tokens. These tokens take part in POS tagging in natural language processing; part-of-speech (POS) taggers have been developed to classify words based on their parts of speech. For sentiment analysis, a POS tagger is very useful because of the following two reasons:

1) Words like nouns and pronouns usually do not contain any sentiment. It is able to filter out such words with the help of a POS tagger;

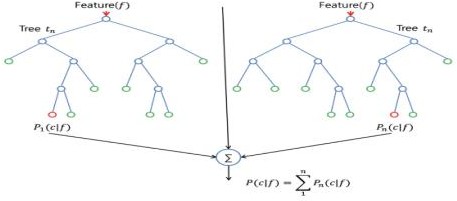
2) A POS tagger can also be used to distinguish words that can be used in different parts

of speech.

**NEGATIVE PHRASE IDENTIFICATION**: Words such as adjectives and verbs are able to convey opposite sentiment with the help of negative prefixes. For instance, consider the following sentence that was found in an electronic device’s review: “The built in speaker also has its uses but so far nothing revolutionary." The word, “revolutionary" is a

positive word according to the list in. However, the phrase “nothing revolutionary" gives more or less negative feelings. Therefore, it is crucial to identify such phrases. In this work, there are two types of phrases have been identified, namely negation-of-adjective (NOA) and negation-of-verb (NOV).

**Random forest:** The random forest classifier was chosen due to its superior performance over a single decision tree with respect to accuracy. It is essentially an ensemble method based on bagging. The classifier works as follows: Given D, the classifier firstly creates k bootstrap samples of D, with each of the samples denoting as Di. A Di has the same number of tuples as D that is sampled with replacement from D. By sampling with replacement, it means that some of the original tuples of D may not be included in Di, whereas others may occur more than once. The classifier then constructs a decision tree based on each Di. As a result.



**Logistic Regression**

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

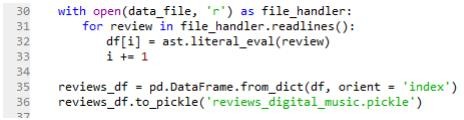
A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)

* Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line. On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group. Logistic regression uses maximum likelihood estimation (MLE) to obtain the model coefficients that relate predictors to the target. After this initial function is estimated, the process is repeated until LL (Log Likelihood) does not change significantly

1. **Implementation Details**

The training of dataset consists of the following steps: Unpacking of data: The huge dataset of reviews obtained from amazon.com comes in a .json file format. A small python code has been implemented in order to read the dataset from those files and dump them in to a pickle file for easier and fast access and object serialization. Hence initial fetching of data is done in this section using Python File Handlers. Preparing Data

for Sentiment Analysis: i) The pickle file is hence loaded in this step and the data besides the one used for sentiment analysis is removed. As shown in our sample dataset in Page 11, there are a lot of columns in the data out of which only rating and text review is what we require. So, the column, “review Summary” is dropped from the data file



Preparing Data for Sentiment Analysis: i) The pickle file is hence loaded in this step and the data besides the one used for sentiment analysis is removed. As shown in our sample dataset in Page 11, there are a lot of columns in the data out of which only rating and text review is what we require. So, the column, “review Summary” is dropped from the data file.

ii) After that, the review ratings which are 3 out of 5 are removed as they signify neutral

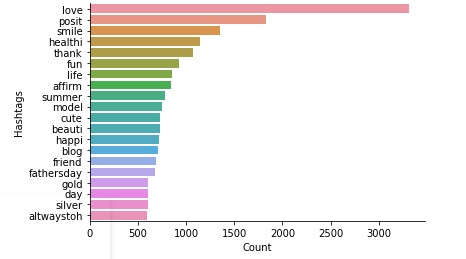
review, and all we are concerned of is positive and negative reviews. iii) The entire task of preprocessing the review data is handled by this utility class- “NltkPreprocessor

**Code for Project:**

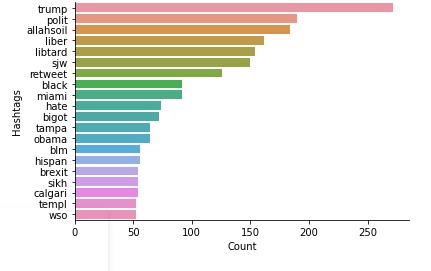
**Results and Sample Output**

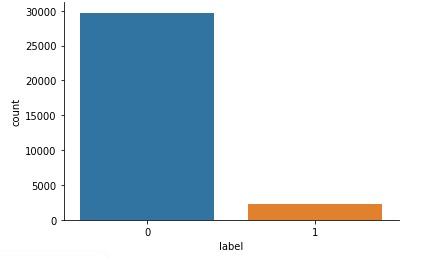
The ultimate outcome of this Training of Public reviews dataset is that, the machine is capable of judging whether an entered sentence bears positive response or negative response. Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance.

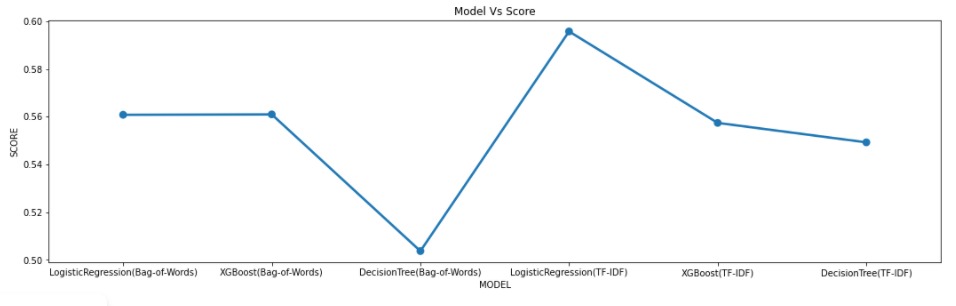
Positive Sentiments:



Negative Sentiments:







**Conclusion**

Sentiment analysis deals with the classification of texts based on the sentiments they contain. This article focuses on a typical sentiment analysis model consisting of three core steps, namely data preparation, review analysis and sentiment classification, and describes representative techniques involved in those steps. Sentiment analysis is an emerging research area in text mining and computational linguistics, and has attracted considerable research attention in the past few years. Future research shall explore sophisticated methods for opinion and product feature extraction, as well as new classification models that can address the ordered labels property in rating inference.

Applications that utilize results from sentiment analysis is also expected to emerge in the near future.

**References:**

[1] Zeenia Singla, Sukhchandan Randhawa, and Sushma Jain, “STATISTICAL AND SENTIMENT ANALYSIS OF CONSUMER PRODUCT REVIEWS” IEEE-2017

[2] Upma Kumari, Dr. Arvind K sharma,Dinesh Soni, “Sentiment analysis of Smart Phone Product Review Using SVM Classification Technique” 2017 IEEE

[3] R. Pavithra and A. R. Mohamed Shanavas ,“Sentiment Analysis about Smart Phones Using Twitter Corpus by Deep Learning Approach “ Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.8 No.S2, 2019

[4] Joana Gabriela Ribeiro de Souza, Alcione de Paiva Oliveira,”Deep Learning Approach for Sentiment Analysis Applied to Hotel’s Reviews “Springer International Publishing 2018

[5] Jagadeesh Panthati, Jasmine Bhaskar , Tarun Kumar Ranga , Manish Reddy Challa, “Sentiment Analysis of Product Reviews using Deep Learning” 2018 IEEE

[6] Satuluri Vanaja,Meena Belwal,”Aspect-Level Sentiment Analysis on E-Commerce Data ”IEEE 2018

* [**https://www.kaggle.com/kazanova/sentiment140**](https://www.kaggle.com/kazanova/sentiment140)
* [**https://github.com/KingstonKrunal/Sentiment-Analysis-of-Tweeter-Data**](https://github.com/KingstonKrunal/Sentiment-Analysis-of-Tweeter-Data)