Sentimental Analysis Using Machine Learning Models

**INTRODUCTION:**

Sentiment is a machine learning (AI) model that measures consumer satisfaction based on a huge number of customer interactions. It is also known as opinion mining that employs natural language processing to assess whether data is positive, negative, or neutral. It is a proven predictive indicator of customer satisfaction and can be used to analyze processes, behaviors, products, and more for optimal customer experiences and operational efficiency. Businesses frequently utilize it to detect sentiment in social data, assess brand reputation, and gain a better understanding of their customers.

Sentiment analysis is becoming a crucial tool for monitoring and understanding client sentiment as they share their opinions and feelings more openly than ever before. Brands can learn what makes customers happy or frustrated by automatically evaluating customer feedback, such as comments in survey replies and social media dialogues. This allows them to customize products and services to match their customers' demands.

With this project I can learn the machine learning models which are to be performed on the dataset and positive and negative outlook of people regarding a topic.

**PROBLEM STATEMENT:**

With this data set which has a collection of tweets, can be used to detect the sentiment associated with a particular tweet and define it as negative or positive accordingly using Machine Learning models.

**DATA SET DESCRIPTION:**

I used twitter sentiment analysis dataset. My dataset contains 31962 labeled tweets and 17191 unlabeled tweets. My dataset contains three attributes id, tweets, and label

1. Id: The number associated with each tweet
2. Tweets: Tweets made by the various people and having either the negative or positive sentiment
3. Label: A positive sentiment is given as label ‘0’ and negative sentiment is given as label ‘1’.

I took the data set from <https://raw.githubusercontent.com/dD2405/Twitter_Sentiment_Analysis/master/train.csv>.

**DATA PRE-PROCESSING:**

Data cleaning:

1. I removed twitter handles(@User) clearly see that the Twitter handles do not contribute anything significant to solve our problem.
2. Punctuation, numbers, and special characters do not help much. It is better to remove them from the text just as we removed the twitter handles.
3. I have decided to remove all the words having length 3 or less. These words are also known asStop Words.

I performed other Data Pre-processing steps like tokenization and stemming.

**RELATED WORK**:

Sentiment Analysis is the study of how one's viewpoints and perspectives are linked to one's mood and attitude as expressed in natural language in response to an event.

Recent events demonstrate that sentiment analysis has advanced to a point where it can go beyond positive vs. negative and address the entire spectrum of behavior and feelings for various communities and themes. In the field of sentiment analysis, a large number of research have been done utilizing various methodologies to anticipate social attitudes. *Ankita Gupta et al*. To improve classification accuracy, a hybrid model based on SVM and KNN is described in this paper. While much of the work in this sector is related with 2-way categorization, the proposed method categorized the tweets in positive, negative, and neutral moods. The suggested model's work has gone through three stages: preprocessing, feature creation, and classifier learning. *Christianini and Taylor.* SVM is a machine learning algorithm that has been published and shared. The authors are able to provide a thorough grasp of the technique and how to approach the SVM algorithm in order to apply it to real-world challenges. *Malhar and Ram* SVM surpasses all other classifiers in a case study of Presidential and Assembly elections, using supervised machine learning techniques and artificial neural networks to classify twitter data. Using the user influence factor, the authors suggested a mechanism for predicting election results. The authors used a combination of Principle Component Analysis and SVM to achieve dimension reduction. *Neri et al.* performed sentiment analysis on more than 1000 Facebook postings and then compared the sentiment for La7, a dynamic firm, and Rai, an emerging Italian social broadcasting provider. The observations of the authors were compared to a study undertaken by the Osservatorio di Pavia, an Italian research institute specializing in the empirical and theoretical study of media and obsessed with the study of political communication in the mass media. *Po-Wei Liang et.al* In the 2012 presidential elections in the United States, a system based on real-time sentiment analysis on the online microblogging site Twitter was proposed. Traditional election analysis takes a long time to obtain poll data, but this technique uses twitter, a microblogging service, to collect data from a larger number of people. It enables social figures like as academics, the media, and politicians to broadcast their future views on public opinion and the political process. Finally, the authors stated that the system and technique are generic, and that they should be easily adopted and expanded over a variety of other fields.

**OBJECTIVES OF THE STUDY**:

The main goal of this project is to identify tweets having negative sentiments. If a tweet contains racist or sexist tweets, we identify it as hate speech using this dataset.

From this project I will have an analysis to determine the sentiment used in a tweet using machine learning models to have feedback as positive or negative or neutral based on the nature of the word.

**RESEARCH DESIGN AND METHODOLOGY**:

To execute sentiment analysis, the data needs to be obtained from the appropriate source. This data goes through a series of pre-processing stages to make it more machine-friendly than it was before.

Diagram, schematic

Description automatically generated

General Methodology of sentimental Analysis

One of the best things about machine learning is that the algorithms can memorize the data, but when we try to apply it for new data, it performs poorly. This is known as overfitting. To circumvent this issue, I employ a test-driven approach. Each dataset is divided into three parts at random, with each portion further divided into three parts:

• Train (60%): Used to feed machine learning into the learning process' algorithm.

• Test (20%): Used to determine whether our algorithm is overfitting or not.

• Validation (20%): Used to evaluate the algorithm's execution.

The data analytic methods I used in this work are:

1. Logistic Regression
2. XGBOOST
3. Decision Tree

**DATA VISUALIZATION:**

Data visualization is the process of transforming huge data sets into Bars, graphs, and other visuals. The final visual representation of data makes it easier to identify, outliers, and new insights about the information represented in the data.

I have done the data visualization in terms of WordCloud

WordCloud: A WordCloud is a visual representation in which the most often used words are displayed in large font and the less frequently used words are displayed in smaller font.

I have generated WordCloud for Positive sentiment Tweets

Result:

Text

Description automatically generated

Most of the words in WordCloud are either positive or neutral Sentiment.

I also generated WordCloud for Negative Sentiment Tweets

Text

Description automatically generated

Most of the words in WordCloud are Mostly has Negative Sentiment.

**FEATURE EXTRACTION TECHNIQUES:**

I used two Feature extraction techniques

1. Bag-of-Words
2. TF-IDF
3. Bag-of-words Feature: A approach for extracting features from text documents is known as Bag of Words. Machine learning algorithms can benefit from these qualities. It generates a vocabulary of all the unique words that appear in the training set's papers.
4. TF-IDF Feature: TF-IDF stands for Term Frequency-Inverse Document Frequency

Term Frequency: The number of times a word appears in a document, divided by the total number of words in that document.

IDF: The logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

**MACHINE LEARNING MODELS**:

I applied three machine learning models they are:

1. Logistic Regression: Logistic Regression is a Machine Learning algorithm that is used to solve classification problems. It is a predictive analytic approach that is based on the probability notion.
2. XGBoost: XGBoost is a machine learning method that has recently dominated Kaggle tournaments for structured or tabular data. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees.
3. Decision Tree: by learning simple decision rules inferred from past data, the purpose of employing a decision tree is to develop a training model that can be used to predict the class or value of the target variable.

F1-score: The weighted average of Precision and Recall is used to get the F1 Score. As a result, both false positives and false negatives are considered while calculating this score. F1 isn't as intuitive as accuracy, but it's usually more useful, especially if your class distribution isn't even.

**RESULTS:**

I applied all three models and compared between the models by using F1-Score.

I compared F1-score between models using Bag-of-words feature:

|  | **1** | **2** | **3** |
| --- | --- | --- | --- |
| **Model** | LogisticRegression(Bag-of-Words) | XGBoost(Bag-of-Words) | DecisionTree(Bag-of-Words) |
| **F1\_Score** | 0.572135 | 0.581102 | 0.514178 |

I compared F1-score between models using TF-IDF feature:

| **1** | **2** | **3** |
| --- | --- | --- |
| **Model** | LogisticRegression(TF-IDF) | XGBoost(TF-IDF) | DecisionTree(TF-IDF) |
| **F1\_Score** | 0.586207 | 0.5792 | 0 |

I can clearly say that best possible model for Bag-of-words feature is XGBoost and the best possible model for TF-IDF feature is Logistic Regression

**CONCLUSION:**

Sentiment analysis is a technique for determining people's beliefs, attitudes, and emotional states. Human beings' (man/woman) perspectives might be positive or negative. Sentiment analysis can be used in a variety of commercial situations, including brand monitoring and product analytics, as well as customer service and market research. Leading brands can work faster, more accurately, and toward more meaningful purposes by embedding it into their existing systems and analytics. This analysis has strong commercial interest because companies want to know how their products are being perceived and also consumers want to know what the existing customers think about the products.

**BIBLIOGRAPHY:**

1. A Gupta, J Pruthi, N Sahu “Sentiment Analysis of Tweets using Machine Learning Approach” International Journal of Computer Science and Mobile Computing, Vol.6 Issue.4, April- 2017, pg. 444-458, from <https://www.ijcsmc.com/docs/papers/April2017/V6I4201799.pdf>.
2. Andrew, “An Introduction to Support Vector Machines and Other Kernel‐based Learning Methods” 20016Nello Christianini and John Shawe‐Taylor. An Introduction to Support Vector Machines and Other Kernel‐based Learning Methods. Cambridge University Press, 2000. xiii + 189 pp., ISBN: ISBN 0‐521‐78019‐5 Hardback: £27.50, Kybernetes, vol. 30, no. 1, pp. 103- 115, 2001.
3. M. Anjaria and R. Guddeti, “Influence factor-based opinion mining of Twitter data using supervised learning”, Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on, pp. 1--8, 2014. From <https://ieeexplore.ieee.org/document/6734907>.

4. F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By,‖ “Sentiment Analysis on Social Media ‖, Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*”*, pp.919—926, 2012. From <https://ieeexplore.ieee.org/document/6425642>.

5. P Liang, B 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, <https://ieeexplore.ieee.org/document/6569069>.