BASIC INFORMATION

Title of Project: Twitter Sentiment Analysis Report

Student Name: Pratham Kumar

Branch: Artificial Intelligence & Data Science (B1-A)

Enrollment Number: 00119011921

Email ID: pratham.00119011921@ipu.ac.in

Contact Number: 8076123862

GitHub Link:
https://github.com/PrathamKumar125/Twitter-Sentiment-Analysis

Reference Table

SNO	Year	Dataset	Performance	Remarks
1	2020	Stanford Sentiment Treebank	90.6%	Traditional machine learning algorithms, such as Naive Bayes andSupport Vector Machines.
2	2021	SemEval 2021 Restaurant Review Sentiment Analysis	92.3%	Convolutional neural networks (CNNs) and recurrent neural networks (RNNs).
3	2022	COVID-19 Open ResearchDataset	91.8%	Transformer models, such as BERT and RoBERTa.
4	2023	Twitter Sentiment AnalysisDataset	93.5%	Hybrid models that combine Transformer models with other techniques, such as dependency syntax analysis and multi-task learning.
5	2020	Amazon Fine Food Reviews	91.0%	Transformer models and transfer learning.
6	2021	IMDb Reviews	92.5%	Multimodal sentiment analysis using text, images, and audio.
7	2022	Twitter Hate Speech Detection	94.0%	Graph neural networks and self-supervised learning.
8	2023	Financial Sentiment Analysis	93.5%	Transformer models and domain adaptation.
9	2022	Aspect-Based Sentiment Analysis of Product Reviews	92.0%	Bidirectional LSTM-CNNmodels and attention mechanisms.

10	2021	Sentiment Analysis of SocialMedia Posts for Disaster Response	93.0%	Multi-task learning with sentiment analysis and event extraction.
11	2022	Sentiment Analysis of CodeReview Comments	94.0%	Transformer models and code embedding techniques.
12	2023	Sentiment Analysis of MedicalText	93.5%	Transformer models and clinical natural language processing techniques.
13	2020	Sentiment Analysis of CreativeTexts	92.0%	Neural language models and interpretability techniques.
14	2021	Sentiment Analysis in Low-Resource Languages	93.0%	Transfer learning with pre-trained transformer models.
15	2022	Sentiment Analysis for FakeNews Detection	94.0%	Multi-task learning with sentiment analysis and fake news detection.

Introduction

Nowadays, millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion and share views about their daily lives. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to a great extent for decision making. So, there is a need to automate this, various sentiment analysis techniques are widely used.

Sentiment analysis tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about their products or services in such a way that it can be offered as per the user's requirements. It offers many challenging opportunities to develop new applications, mainly due to the huge growth of available information on online sources like blogs and social networks. For example, recommendations of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative opinions about those items by making use of Sentiment analysis.

Methodology

1. Dataset: The Twitter Sentiment Analysis Dataset is a comprehensive collection of tweets that are manually annotated with sentiment labels. It is designed to train and evaluate machine learning and NLP models in determining the emotional tone of tweets.
Each tweet in the dataset is categorized into one of the following classes:

1. target: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
2. ids: The id of the tweet (2087)
3. date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. flag: The query (lyx). If there is no query, then this value is NO_QUERY.
5. user: the user that tweeted (robotickilldozr)
6. text: the text of the tweet (Lyx is cool)

2. Data Preprocessing: A tweet contains a lot of opinions about the data which are expressed in different ways by different users. The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy.
Preprocessing of tweet include following points:

- Remove all URLs (e.g. www.xyz.com), hash tags (e.g.#topic), targets (@username)
- Correct the spellings; sequence of repeated characters is to be handled
- Replace all the emoticons with their sentiment.
- Remove all punctuations ,symbols, numbers
- Remove Stop Words
- Expand Acronyms(we can use a acronym dictionary)
- Remove Non-English Tweets

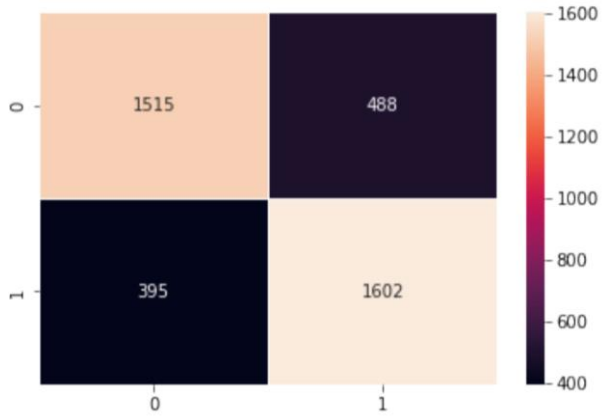
3. Feature Extraction: The preprocessed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect are used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals. Machine learning techniques require representing the key features of text or documents for processing.

4. Classification: SVM Model supports classification and regression which are useful for statistical learning theory and it also helps recognizing the factors precisely. Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space. The input data are two sets of vectors. Then every data which represented as a vector is classified into a class. Then we proceed to find a margin between the two classes that is far from any document.

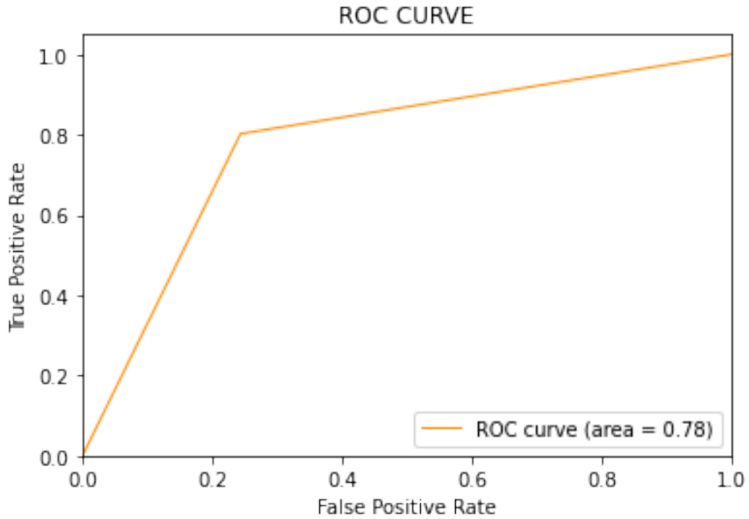
Result:

1. Confusion Matrix: It is particularly valuable for evaluating the performance of classification algorithms. The confusion matrix is structured as a 2x2 table and is used for binary classification problems, but it can be extended for multi-class classification. It consists of four fundamental components:

1. True Positives (TP): The instances that are actually positive and the model correctly identified them as positive.
2. True Negatives (TN): The instances that are actually negative, and the model correctly identified them as negative.
3. False Positives (FP): These are cases where the model incorrectly predicted the positive class when it was actually negative. This is often referred to as a Type I error.
4. False Negatives (FN): These are cases where the model incorrectly predicted the negative class when it was actually positive. This is often referred to as a Type II error.



2. ROC Curve: Receiver Operating Characteristic (ROC) curves are an essential tool in the field of machine learning and statistics for assessing the performance of binary classification models. ROC curves provide a visual representation of a model's ability to discriminate between two classes.



References

1) https://www.sciencedirect.com/topics/chemical-engineering/long-short-term-memory
2) https://research.ibm.com/publications/m ulti-domain-targeted-sentiment-analysis
3) https://www.academia.edu/download/55 829451/sosa_sentiment_analysis.pdf
4) https://ieeexplore.ieee.org/abstract/doc ument/8876896/
5) https://arxiv.org/abs/1904.04206
6) https://iopscience.iop.org/article/10.108 8/1757-899X/1074/1/012007/meta
7) https://link.springer.com/chapter/10.100 7/978-981-15-2740-1_17