

**Social Network Analytics Lab**

**Digital Assignment -3**

**Report of the Assignment**

Register Number: 22MCB0033

Name: Sathwik Shetty B

Course Name: Social Network Analytics Lab

Course code: MCSE618P

**Topic: Sentiment Analysis of Twitter Dataset**

**Index of Contents:**

1. Abstract
2. Introduction
3. Description of Dataset
4. Methodology
5. Results
6. Conclusion
7. Future Cope

**Abstract:**

In order to comprehend the underlying sentiment conveyed by users and to extract insights from textual data, sentiment analysis is a critical component of social network analytics. In this context, we investigated the use of Naive Bayes classifier, Support Vector Machine (SVM), Logistic Regression, Random Forest classifier, Faatext word embedding and Transformer based model with BERT-based word embedding. To train and assess these models, we used a dataset of Twitter data that included text and accompanying sentiment classifications (positive, negative, or neutral).

The complexity of sentiment expressions and the amount of processing power available determine which model is used for sentiment analysis. We can choose the best model for a specific social network analytics task by comparing and assessing these models' performance.

**Introduction:**

Sentiment analysis, commonly referred to as opinion mining, is essential for social network analytics because it draws important conclusions from the massive amounts of textual data produced on sites like Twitter. Machine learning models are used to identify the sentiment polarity associated with a given text passage by examining patterns, grammatical hints, and contextual data contained in the text. We use a dataset of Twitter data that includes numerous tweets and their related sentiment classifications to address this issue. We use the TfidfVectorizer to convert the raw text data into a numerical representation appropriate for machine learning methods. The four models being examined display various skills and traits for sentiment analysis.

The Naive Bayes classifier offers simplicity and effectiveness in training and prediction, while SVM can capture complex correlations between features and uses a maximum margin technique. Logistic regression provides interpretability by estimating the likelihood of a binary result. The Random Forest classifier is best at handling high-dimensional feature fields and can detect both linear and nonlinear correlations between words and attitudes. Faatext word embedding and Transformer based model with BERT-based word embedding can capture nuances in the relationships between different variables in the dataset. To improve sentiment analysis' accuracy and efficacy, businesses and organisations should examine and contrast various machine learning models. This will enable businesses and organisations to make data-driven decisions and obtain insightful knowledge from the large amount of textual data available on social networks.

**Description of Dataset:**

The collection of tweets and the related sentiment classifications make up the Twitter dataset used in this analysis. The dataset is a useful tool for sentiment analysis since it makes it possible to investigate the sentiments expressed by Twitter users.

The collection includes a number of features that offer useful details about each tweet. The characteristics are described as follows:

textID: A distinct identity for each tweet that makes it simple to refer to and identify a particular tweet.

text: The tweet's exact text. The textual material that users have submitted on Twitter is contained in this attribute.

selected\_text: A piece of the tweet's text that has been chosen to best express the main idea or message. As the target variable for training and evaluation in sentiment analysis tasks, this property is frequently utilised.

Sentiment: The label for the tweet's feelings. It represents the polarity of the sentiment that was stated in the tweet and can be either positive, negative, or neutral.

The dataset acts as a sample of tweets from various people that represent a variety of subjects and moods. It offers a chance to investigate how well sentiment analysis models function across various sentiments and textual phrases.

The Twitter dataset is an invaluable resource for sentiment analysis research, social network analytics, and related applications since it provides insightful information on public sentiment and opinion on a variety of topics. Researchers and practitioners can better comprehend sentiment patterns, identify trends, and make data-driven decisions based on the feelings expressed by Twitter users by analysing this dataset and using machine learning models.

**Methodology:**

Data preparation: Dealing with missing values Look through the dataset for any missing values, and then use the right techniques to address them, like imputation or elimination.

Choose pertinent columns: Choose the columns that will be used for sentiment analysis, including the ones that will include the tweets' text and their related sentiment labels.

Data Division: Create training and test sets from the dataset: Separate the data into two subsets, one for modelling training and the other for performance assessment.

Vectorization of text: Textual data to numerical features conversion: To vectorize the raw text data into a numerical representation, use the TfidfVectorizer. In this stage, words are given numerical weights depending on their relevance and frequency within each tweet and across the entire dataset.

Model Education and Assessment: Fit the Naive Bayes classifier using the training set and the accompanying sentiment labels to train it. The probabilities of each sentiment class are determined by this model, which implies independence between features. Use the training data and sentiment labels to train the Support Vector Machine model. SVM looks for the best hyperplane in a high-dimensional feature space to divide various sentiment classes. Fit the Logistic Regression model using the training set and sentiment labels to train the model. The odds of a binary outcome are estimated using logistic regression, which also prioritises the features.

Instructions for training the Random Forest classifier: The training set and sentiment labels should be used. In order to capture intricate interactions between features, this ensemble model mixes various decision trees.

Faatext word embedding and Transformer based model with BERT-based word embedding are built on the dataset which prioritises the important parameters in the dataset.

Model contrast: Compare the models' performance: Examine and evaluate the findings of the Naive Bayes classifier, SVM, Logistic Regression, Random Forest classifier, Faatext word embedding and Transformer based model with BERT-based word embedding. To assess each model's efficacy in sentiment analysis, take into account the prime parameter i.e, accuracy.

Identify your talents and shortcomings. Depending on the results, determine each model's advantages and disadvantages. Take into account elements like interpretability, handling complex relationships, processing efficiency, and the capacity to convey subtle emotions.

Model Choice and Implementation: Choose the most appropriate model: Select the model that best satisfies the requirements and goals of the sentiment analysis task in the context of social network analytics based on the comparative analysis and performance evaluation.

Release the chosen model: Apply the selected model to new, unexplored data to perform sentiment analysis. This can entail incorporating the model into a bigger application or applying it to gauge sentiment in social media streams as they are being generated in real-time.

**Github link:**

**Results:**

|  |  |
| --- | --- |
| **MODELS** | **ACCURACY** |
| Naïve Bayes classifier | 65.01% |
| Support Vector Machine | 71.2% |
| Logistic Regression | 69.81% |
| Random Forest | 67.65% |
| Faatext word embedding | 68.32% |
| Transformer based model with BERT-based word embedding | 68.50% |

**Conclusion:**

In conclusion, we used these distinct machine learning models—Naive Bayes classifier, Support Vector Machine (SVM), Logistic Regression, Random Forest classifier, Transformer based model with BERT-based word embedding —to perform sentiment analysis on a Twitter dataset. The objective was to evaluate the performance of these models and compare how well they classified the emotions represented in tweets.

The following findings were reached after rigorous testing and experimentation:

The Naive Bayes Classifier had a sentiment classification accuracy of 65.01%. Although it performed reasonably well, it had the lowest accuracy of the four evaluated models. Naive Bayes makes the assumption that each feature is independent and may have trouble capturing complicated relationships in the data.

SVM surpassed the competition, delivering the greatest sentiment classification accuracy of 71.2%. SVM is useful for handling complex relationships and identifying patterns in the data since it seeks to identify the best hyperplane to divide several sentiment classes.

The sentiment categorization accuracy for logistic regression was 69.81%. The odds of a binary outcome are estimated using logistic regression, which also prioritises the features. Compared to SVM, it showed competitive performance but significantly worse accuracy.

The sentiment classification accuracy of the Random Forest classifier was 67.65%. In order to capture intricate interactions between features, this ensemble model mixes various decision trees. Although it performed similarly to SVM, it was a little less accurate.

Faatext word embedding and Transformer based model with BERT-based word embedding provided accuracies of 68.32% and 68.50% respectively.

These results lead us to the conclusion that, of the investigated models, the Support Vector Machine (SVM) model performs the best in terms of sentiment analysis on the Twitter dataset. Its success in achieving the highest accuracy shows that it can correctly identify the feelings represented in tweets. SVM is a good option for sentiment analysis jobs because of its capacity to handle complex relationships and detect patterns in high-dimensional data.

**Future Scope:**

There are several potential directions for future research and development in the fields of sentiment analysis and social network analytics. Here are some possible research areas:

Sentiment analysis at a finer scale: At the moment, sentiment analysis algorithms categorise tweets into general sentiment categories as positive, negative, or neutral. The development of models that can capture more nuanced emotions, such as happiness, sadness, rage, or surprise, could be the main goal of future study. This would offer more subtle insights into the feelings and opinions of users.

Aspect-based sentiment analysis: A lot of tweets discuss particular facets or qualities of a good, service, or occasion. The goal of aspect-based sentiment analysis is to locate and examine emotions connected to various features mentioned in the text. Future research can concentrate on creating models that can extract and analyse sentiments at the aspect level, allowing for more focused analysis and decision-making.

Sarcastic and sardonic statements are frequently made on social media sites, which might make it difficult for sentiment analysis models to process them. Future study can investigate methods to correctly recognise and understand irony and sarcasm in tweets, increasing the precision of sentiment analysis in such circumstances.

Sentiment analysis models based on generic Twitter data may not function well in circumstances that are specific to a certain domain. Sentiment analysis algorithms can be tailored for certain industries, such as banking, politics, or healthcare, to produce more precise and context-aware results. This may entail domain adaption strategies, the development of a vocabulary, or domain-specific feature engineering.

Multimodal sentiment analysis: In addition to textual information, tweets frequently contain images, emoticons, and other visual components. A more thorough understanding of user sentiments on social media platforms can be achieved by integrating multimodal data sources and creating models that can successfully analyse sentiment across different modalities.

Real-time sentiment analysis: Since social media platforms produce a significant amount of data in real-time, it can be beneficial to construct effective and scalable sentiment analysis models that can evaluate sentiments in real-time. Real-time sentiment analysis may help businesses respond quickly to user comments, keep an eye on new trends, and identify changes in target audience mood.

Pre-trained models and transfer learning: For sentiment analysis on huge datasets, transfer learning techniques might be investigated. The performance of sentiment analysis models may be improved by fine-tuning pre-trained models with domain-specific Twitter data or by using methods like contextual embeddings (e.g., BERT, GPT).

Sentiment analysis on social media can be done in multiple languages, as opposed to only one. Sentiment analysis can be used in a variety of linguistic contexts by extending models for the technique to support many languages. The difficulties of multilingual sentiment analysis can be addressed by investigating methods like cross-lingual transfer learning and multilingual embeddings.

By providing a more precise, context-aware, and nuanced analysis of user feelings on websites like Twitter, these future possibilities can progress the field of sentiment analysis in social network analytics. Researchers and practitioners can increase the potential of sentiment analysis and its applications across a variety of fields by solving these issues and looking into new possibilities.