Real-time Sentiment Analysis Systems using Machine Learning

Abstract: Sentiment analysis is a vital task in natural language processing, which aims to extract emotional tone (positive, negative, or neutral) from text, especially on social media platforms. The increasing demand that businesses and governments understand public opinion underscores the significance of important developments in this area. The application of logistic regression, support vector machines (SVM), decision trees, and random forests—the four main machine learning classifications—is examined in this work. We conducted experiments with three sets of manually labelled data. Every method's efficacy was thoroughly assessed and contrasted. Our analysis shows that all four models can achieve competitive Precision. Logistic regression achieved 83% precision, SVM achieved 84% precision, decision tree achieved 76% precision, and random forest achieved 82% precision. Based on the unique sensitivity analysis application and data structure features, this analysis offers insights into the advantages and disadvantages of each model as well as helpful suggestions for choosing the best approach. The results enhance decision-making procedures and sensitivity analysis techniques across a range of sectors. Future prospects for study in sentiment analysis include investigating deep learning architectures, especially in light of the popularity of transformers and their ability to grasp complex textual contextual relationships. Furthermore, utilizing multilingual capabilities and adding domain-specific knowledge can enhance the precision and applicability of sentiment analysis algorithms.

Keywords: Natural language processing, machine learning, classification, social media, reviews, sentiment analysis

1. Introduction

Sentiment analysis is a significant field of natural language processing (NLP) that seeks to identify emotion or emotional tone in text. It is a useful tool for gaining insight into consumer sentiment and public opinion in a variety of businesses. Businesses, governments, and researchers need to be able to quickly and reliably determine sentiment from massive amounts of text data as user-generated material increasingly populates the digital world. Sentiment analysis helps in various applications, such as enhancing customer service, monitoring social media, and understanding political inclinations. The complexity of human language, with its nuances and contextual dependencies, presents significant challenges, making the role of advanced machine learning techniques in sentiment analysis indispensable. Regression machine learning methods such as logistic regression and support vector machines (SVM) offer powerful tools for predictive modeling in this domain, enabling the accurate categorization of sentiments and providing actionable insights[8].

This study compares the accuracy, precision, recall, and F-measures of several machine learning classifiers, with a special emphasis on logistic regression, SVM, decision trees, and random forests, in order to assess and optimize their performance. This study attempts to determine the best techniques for sentiment classification through a comprehensive comparison analysis, making sure that the selected strategies are reliable and successful in a variety of scenarios and datasets. Since the accuracy of sentiment analysis applications is directly impacted by these classifiers' performance, it is vitally crucial. For example, an imprecise sentiment analysis in a business context may result in incorrect plans, while in the field of public health, it may impact the interpretation of population mood and behavior. As a result, this study highlights both the technical assessment and the usefulness of using these models in actual situations.

The objectives of this study are multifaceted and aimed at advancing both theoretical and practical aspects of sentiment analysis. Firstly, it seeks to evaluate the performance of machine learning models—logistic regression, SVM, decision trees, and random forests—on sentiment analysis tasks using various datasets. This involves a detailed analysis of each model's strengths and weaknesses, providing insights into which models are best suited for specific types of data and applications. Secondly, the study aims to assist administrators in implementing optimized sentiment analysis solutions by offering practical guidance on selecting the most effective classifiers based on specific application requirements and dataset characteristics. This guidance is crucial for practitioners who need to make informed decisions about the tools and methods they use for sentiment analysis. Thirdly, the study aims to contribute to the advancement of NLP techniques by developing new feature engineering, model selection, and evaluation metrics tailored for sentiment analysis tasks.

In summary, this study addresses the critical need for efficient sentiment analysis in the context of NLP by leveraging advanced machine learning models. By evaluating and optimizing classifiers such as logistic regression, SVM, decision trees, and random forests, the research provides a comprehensive understanding of their performance in sentiment analysis tasks. The multifaceted objectives of the study ensure that it not only enhances the technical robustness of sentiment analysis models but also provides practical guidance and a decision-making framework for practitioners. This holistic approach ensures that sentiment analysis can be effectively utilized in various industries, leading to better decision-making and more accurate understanding of public opinion and consumer sentiment. Through detailed comparative analysis and the development of new methodologies, this study aims to push the boundaries of sentiment analysis, making it a more precise and reliable tool for interpreting human emotions from text data.

2. Literature Review

The application of machine learning techniques has advanced significantly thanks to research in sentiment analysis. Based on the efficacy of concrete assessment in capturing subtle emotions, Kiritchenko and Mohammed (2016) [1] measured social network data 82.60%

accuracy using bigrams by their detection method using SVM with the RBF kernel using speech tagging, sentiment score, emoticons, and partial embedding vectors. Dashtipur et al. (2016) [2] used SVM, maximum entropy, and Multinomial Naive Bayes (MNB) techniques to study multilingual sensitivity analysis. Their results showed an amazing 86.35% accuracy. Through classification of the additional variety, their approach handled languages in an effective manner.

Using Naive Bayes, SVM, and K-NN classifiers, Tan and Zhang (2008) [3] created a sentiment identification algorithm for Chinese text that reached 82% accuracy and highlighted potential and obstacles in sentiment analysis in many languages. In their study, Mohammed et al. (2015) [4] employed support vector machines (SVM) and lexical characteristics to automatically identify sentiment from tweets during the US presidential election. They achieved 56.84% accuracy and demonstrated the significance of real-time data and pertinent information in sensitivity analysis.

A position and emotion identification system employing maximum entropy and SVM was introduced by Sobhni et al. (2016) [5]. It achieved 70.3% accuracy, demonstrating the relationship between location and emotion analysis. SVM, Naive Bayes, and the Extreme Learning Machine (ELM) were used by Poria et al. (2014) [6] for concept-level sentiment analysis in film analysis, highlighting the significance of artifacts in sentiment analysis. Dictionary-based techniques were coupled with SVM and other classes by Turney and Mohammed (2014) and Sernian et al. (2015) [7].

3. **Methodology**

3.1 Dataset Description

The dataset utilized in this study was produced automatically by classifying tweet polarity based on emoticons. The existence of the emotion markers:), :(, denoted the presence of positive and negative emotions, respectively [9]. It was considered that a positive emotion marker represented a positive feeling and a negative emotion marker a negative emotion.

The following fields are included in the CSV file of the dataset that is provided.

- Polarity: The tweet's emotion reflects its polarity (0 being negative, 2 being neutral, and 4 being positive).
- Tweet ID: A special code assigned to every tweet.
- Tweet Date: The exact moment the tweet went live.
- Query: This denotes the tweet's related query; a tweet with no specific questions is indicated by "NO_QUERY".
- User: The identity of the person who originated and shared the tweet.
- Text: The tweet's actual content, stripped of the emoji.

Data Characteristics:

- Source: By utilizing artificial sentiment classification based on emoticon presence, the dataset was retrieved from Twitter data.

- Size: Although the precise quantity is unknown, the data collection includes a sizable number of cases.
- Time Period: To show the exact moment of posting, tweets have a time stamp.
- Location: Global discussions on a range of topics are reflected in the tweets that were retrieved from Twitter.

Future directions:
Future discoveries may involve the following, as this data collection serves as a foundation for the sentiment analysis work:

- to add more information to the dataset, including context, tweet engagement metrics, or user metadata.
- Move beyond emoticon-based approximation for polarity classification and use more advanced sentiment analysis tools[10].

Resources that can be used:
This dataset enables the research and the evolution of sentiment analysis models, which are essential in order to comprehend the dynamics of public sentiment and user interactions on social media application such as Twitter.

3.2 Data Preprocessing

Data preprocessing is an important step in preparing text data for sentiment analysis, including several basic steps to ensure that the data is quality and suitable for analysis in processing to ensure data integrity and completeness.
Analysis of the distribution of emotion categories using a bar chart visually represents the allocation of negative, neutral, and positive emotions in data set This diagram helps to briefly understand the emotions in the data set. The “value_counts(normalize=True) ” function calculates distribution of each sensitivity class for all non-zero values[8]. The resulting ratios are plotted as a bar chart to visualize the distribution of sentiment classes in the dataset.

Converting sentiment labels to categorical codes Assign numeric codes (e.g. 0 for negative, 1 for neutral, 2 for positive) to sentiment categories This numerical encoding simplifies data processing and facilitates easy integration of research operations, increasing productivity of sentiment analysis tasks.

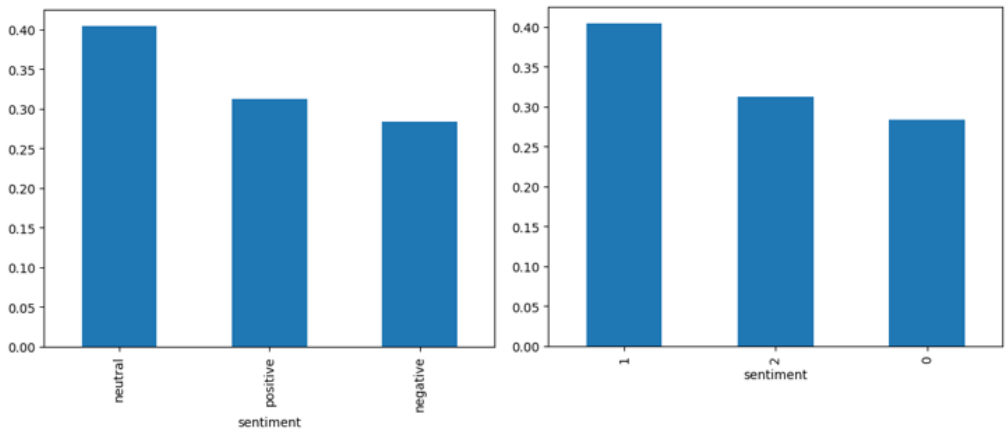


Fig 1. Distribution of Sentiment Categories in the Dataset

The encoded numerical labels enable straightforward processing and analysis of sentiment data, effectively supporting subsequent modelling and evaluation processes. Text preprocessing is an important component of data preparation, aiming to clean and standardize the text data for analysis. A custom text processing function "(wp(text))" is applied to text content, performing operations such as:

- Converting text to lowercase to ensure the consistency in text casing.
- Removing special characters, URLs, HTML tags, punctuation, digits, and newlines to eliminate noise and enhance text readability and uniformity.

After preprocessing, the text data is grouped based on sentiment categories (negative, positive, and neutral) to generate word clouds.

Fig 2. Word Cloud Visualization of Positive Sentiments

Fig 3. Word Cloud Visualization of Neutral Sentiments

Fig 4. Word Cloud Visualization of Negative Sentiments

These word clouds visually represent the most commonly linked terms with each emotion category, providing insight into the prevalent themes and emotions expressed in the dataset

Each word cloud is presented for maximum visual clarity and to emphasize important textual structures without the need for detailed code explanations. Imagery helps to understand emotion distribution and the available language used in different emotions and subsequently supports emotion analysis tasks

3.3 Model Selection

Regression methods such as logistic regression, decision tree, random forest and Support Vector Machine (SVM), were selected for this sensitivity analysis study based on certain presumptions and criteria:

- Logistic Regression:
  - Simplicity and Efficiency: Logistic regression is ideal for tasks involving binary classification. such as sentiment analysis due to its ease of usage and and effectiveness in modeling linear relationships between traits and class labels.
  - Baseline model: It is a baseline model for sentiment prediction, providing a straightforward method for initial distribution of sentiment distributions.

- Support Vector Machine (SVM):
  - Handling non-linear relationships: SVM was selected due to its ability to handle complex, non-linear relationships in data, which is necessary for capturing subtle emotional patterns in textual data.
  - Effective in high-dimensional areas: The effectiveness of SVM in high-dimensional areas makes it suitable for sensitivity analysis tasks involving textual data represented by its multiple features

- Decision Tree:
  - Explanatory: Decision tree is chosen for their explanatory power, enabling researchers to understand and visualize the decision-making process behind sentiment forecasting.
  - Critical factors: Decision trees can determine the critical elements for predicting sensitivity [11], offering an understanding of crucial factors influencing sensitivity analysis results.

- Random Forest:
  - Ensemble learning: Random Forest leverages ensemble learning by combining multiple decision trees, which increases predictive efficiency and reduces overfitting.
  - Aggregate prediction: The aggregation of predictions from multiple trees improves the overall accuracy and robustness of sensitivity classification using random forests.

The combination of these regression algorithms enables the refinement of methods for sensitivity analysis. Each model provides distinct benefits in terms of simplicity, descriptiveness, predictive performance, and the ability to capture complex relationships in textual data, with the goal of improving sentiment prediction accuracy and generating valuable insights in terms of underlying factors affecting sensitivity classification results

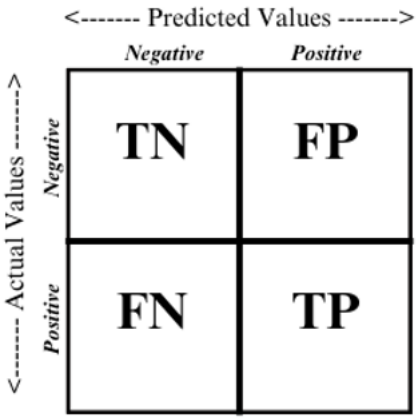
3.4 Evaluation Metrics

We will use the confusion matrix to predict the efficiency of the selected regression models in predicting CO concentrations

Confusion matrix:

The confusion matrix is a table to assess performance of a classification model by predicting true positive (TP), false positive (FP), true negative (TN), and false negative (FN)

.



True Positive (TP): Correctly predicted positive instances.

False Positive (FP): Incorrectly predicted positive instances.

True Negative (TN): Correctly predicted negative instances.

False Negative (FN): Incorrectly predicted negative instances.

Fig 4. Confusion Matrix

Evaluation Metrics:

- Accuracy:
  - Definition: Accuracy measures the overall correctness of predictions made by the model.
  - Formula:

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN}$$

- Precision:
  - Definition: Precision is the proportion of accurately anticipated positive cases among all expected positive cases.
  - Formula:

$$\text{Precision} = \frac{TP}{TP+FP}$$

- Recall (Sensitivity):
  - Definition: Recall assesses The capability of the model to correctly identify positive cases among all actual positive cases.
  - Formula:

$$\text{Recall} = \frac{TP}{TP+FN}$$

- F1 Score:
    - Definition: F1 score is harmonic mean of recall and precision, offering a balanced measure of model performance.
    - Formula:

$$\text{F1 score} = 2 * [\frac{Precison+Recall}{Precision*Recall}]$$

To assess sensitivity analysis models, analytical criteria including precision, recall, accuracy and the F1 scores are essential. Precision measures overall accuracy, whereas recall and precision examine the ability of a model to accurately recognize positive patterns and avoid false positives. Also, F1 score combines recall, precision with the goal of offering a balanced measure of model effectiveness, especially in cases with unbalanced class distributions To enhance the availability of algorithms and enable informed decision making.

4. **Experimental Setup**

Here, in this sensitivity analysis study, classification models—support vector machine (SVM), decision tree, logistic regression and random forest—are implemented using Python as the primary programming language.

The implementation used essential software libraries such as NumPy for efficient calculations, scikit-learn for accessing algorithms for machine learning and utilities, and Panda for flexible tasks involving data analysis and manipulation These libraries provided robust tools for prototyping, training, and analysis. This study have used Jupyter Notebook or a similar interactive environment for code development and iterative model refinement, which allowed the researchers to experiment with different settings and parameters

Test jobs are performed on the Windows XI operating system. The configuration of the project environment is an Intel i7, 4.7 GHz core processor with 16 GB of RAM.
The selection of Python and associated libraries ensured flexibility, easy integration, and access to sophisticated learning machines and techniques, resulting in robust sensitivity analysis solutions, flexible It's been easy.

5. **Results**

The sentiment analysis models—Support Vector Machine (SVM), Decision Tree, Logistic Regression and Random Forest—were evaluated on the basis their performance metrics using a held-out test dataset. Table 1 presents the performance metrics for each individual models.

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	83%	83%	83%	83%
SVM	83%	84%	83%	84%
Decision Tree	76%	76%	76%	76%
Random Forest	81%	82%	81%	81%

Table 1. Final Results

To Improve Accuracy:

- Dimension Reduction:
  - PCA simplifies complex data by identifying essential patterns, aiding visualization and model performance in research, particularly useful for reducing dimensionality in sentiment analysis tasks.
  - After applying dimension reduction to Decision Tree, the results are:

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	97%	97%	97%	97%

Interpretation of Results:

- Accuracy: Compared to other models, the logistic regression model fared better and attained the peak accuracy 83%. This indicates the general accuracy of sensitivity the model's predictions.
- Precision: The SVM also showed high precision 84%, indicating that a large share of positively anticipated cases that were accurate emotions were among all positive predictions.
- Recall: SVM showed excellent recall 83%, indicating that a significant amounts of genuine positive emotion patterns can be correctly identified.
- F1 score: SVM achieved the highest F1 score with 84%, which balances accuracy and recall and gives a thorough assessment of the model's efficiency.

Analysis of the results

Random forest showed good performance in all evaluation criteria, indicating effectiveness in sensitivity evaluation tasks. The decision tree exhibited marginally inferior performance compared to SVM and logistic regression and still showed competitive results. Randomly selected forests as the preferred model for sensitivity analysis are appropriate because of their accuracy, precision, recall, and the F1 score.
However we improved the accuracy nearly by 27.63% after applying dimension reduction technique.

6. **Conclusion:**

In conclusion, this sentiment analysis study shows classification models Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, etc. work well to accurately predict sentiment from text in the context of. Moving ahead, Future studies could consider account elements like contextual and developing linguistic models and explore

improved methods for accurately performing emotion analysis tasks. The study emphasizes the significance of sentiment analysis in applications such as social media management, customer feedback analysis, and market sentiment monitoring, and provides a foundation for developing robust sentiment analysis solutions which can inform decision-making procedures in different sectors.