Enhancing Sentiment Analysis with Supervised Machine Learning on the IMDB Dataset

GROUP 19

GROUP MEMBERS:

- Varun Kumar Kambhampati
- Harsha Vardhan Reddy Yarmareddy
- Pravallika Chidipudi

Outline

- **Abstract**
- **Introduction**
- **Methodology**
- **Tools**
- **Techniques**
- **Coding Phase**
- **Conclusion**
- **Recommendations**

ABSTRACT

- The sentiment analysis is part of natural language processing as it is needed to have the insight of the tone that is being communicated in textual data, which is from many sources like product reviews, social media posts or client responses.

- We rely on the IMDb dataset to be able to carry out evaluations of the diversity of supervised learning classes--such as Support Vector Machines, Naive Bayes, Logistic Regression and Decision Trees--and choose the best model for sentiment classification.

- This research has enhanced the field of sentimental analysis by pointing out that the supervised manner of learning models gives accurate outcomes when dealing with actual text in real time.

INTRODUCTION

- The IMDB extensive document of movie review, that are of various languages, genres, and generations will serve as a tool for this research. 5,000 rows of movie reviews and ratings is the size of the IMDb Kaggle dataset.

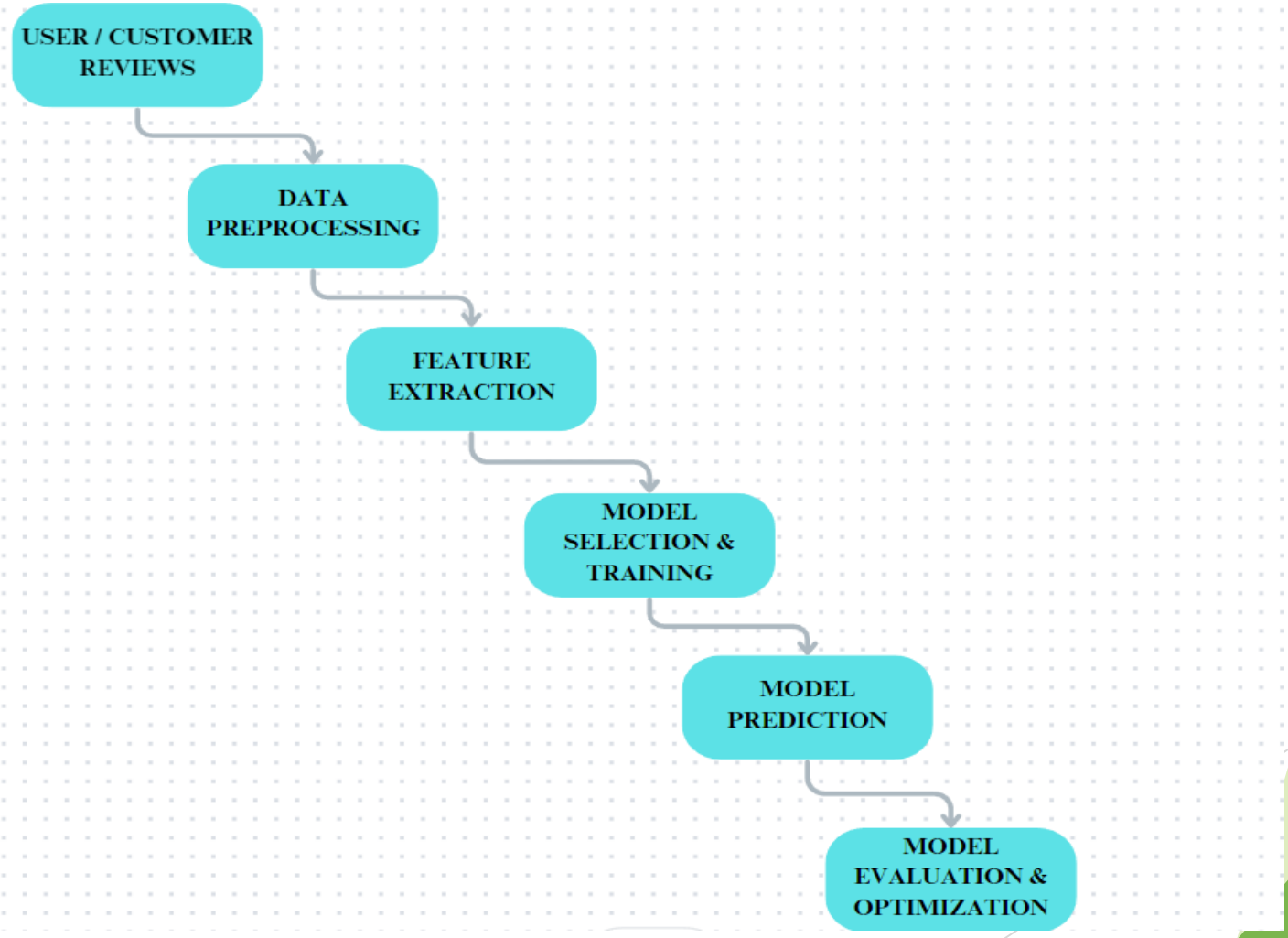
- These rows denote the review content and ratings. The film covers so many different genres and movements, and, of course, the reviews were from different eras as well.

- Getting this labeled corpus which is entirely ideal for training and testing purposes of ML supervised model models for sentiment analysis. We apply logistic regression, a decision tree, an SVM, and a naive Bayes learner performs optimally on the agenda to create the classifier for this dataset

INTRODUCTION

- The right choice of machine learning algorithm turns itself into a determining factor in developing fully ripe sentiment analysis models.
- Our paper intends to determine the sentiment analysis capacity of the SVM, Naive Bayes, Decision Tree and Logistic Regression methods when done on the IMDB dataset through rigorous testing and evaluation.
- We plan to figure out the algorithm which will make the correct movie review categorization in to positive or negative sentiments and we will evaluate the results by using the measures like, accuracy, precision, recall, and F1-score.

METHODOLOGY



TOOLS

- Python, for instance, is a widely-used language allowing tools of data exploration, research in the field of machine learning and application of NLP (Natural Language Processing) technologies.

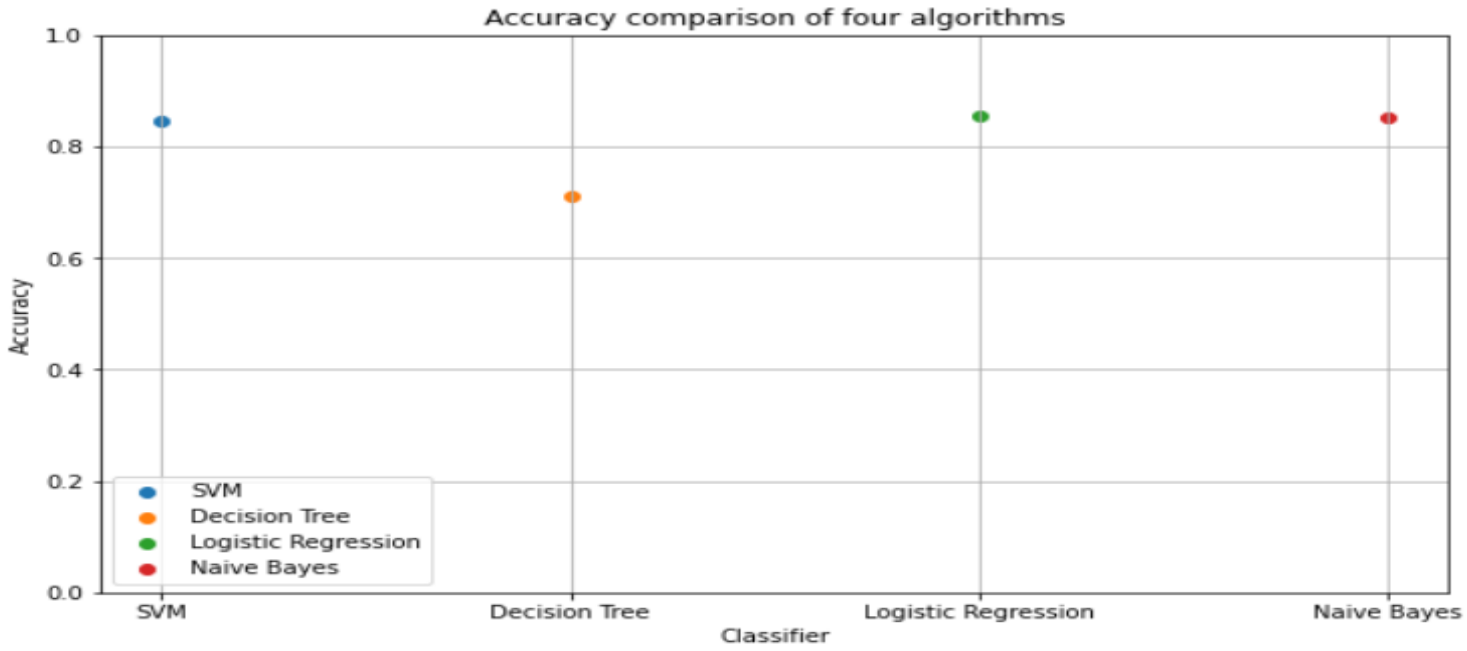
- The Python libraries having mentioned names and others like scikit-learn, NLTK (Natural Language Toolkit) are commonly used for the text extraction, the preprocessing, the training of the models and their evaluation.

- We have used some programs like Matplotlib which are used for data visualization

TECHNIQUES

- TF-IDF
- Naive Bayes
- SVM
- Logistic Regression
- Decision Tree
- Hyperparameter tuning
- Cross-Validation techniques

CODING PHASE



Accuracy Comparison Graph of Classifiers

- **Support Vector Machine Classifier has 89%**
- **Decision Tree Classifier has 71%**
- **Logistic regression has 85%**
- **Naive bayes has 87%**

CONCLUSION

- Support Vector Machine and Naïve Bayes Classifiers exhibit amazing accuracy levels of **89%** and **87%** respectively, it should be noted that the Logistic Regression comes the next best with **85%**. Despite the fact that the Decision Tree Classifier has the accuracy of **71%**, it is still an online tool worth recommending for sentiment analysis applications.

- It will be shown that sentiment analysis models capabilities of profiting and generalizing will be assessed by metrics like accuracy, precision, recall, and F1-score.

- It is clear that there are several industries that may benefit from the combination of datasets like IMDB and Machine learning algorithms, for instance, market research, feedback analysis, and the entertainment industry that is now using Machine Learning based sentiment analysis.

RECOMMENDATIONS

- We would like to proffer some useful strategies that could lead up to better efficacy of the models and their capability to the generalization.

- Elaborating the neural network architecture with sophisticated deep learning models such as transformer approach (for instance, BERT) or recurrent neural networks (RNNs) could boost the accuracy of the models and their capability to comprehend fine-grained semantic relationships in movie reviews.


thank
you