PROJECT REPORT DATA SCIENCE FALL 2023

Daraz Sentimental Analysis Report

Yousha Saibi 22L-7482 Ibtehaj Ali 22L-7476

Department of Data Science, National University of Computer and Emerging Sciences, Lahore

December 3, 2023

Abstract

- Problem Statement
- Methodology
- Results
- Conclusion

1 Introduction

Problem Statement

Design and implement a robust sentiment classifier for **online reviews** to automatically analyze and categorize user sentiments as positive or negative.

In the era of online shopping and information overload, user reviews play a crucial role in influencing purchasing decisions. The objective of this project is to develop a **Natural Language Processing (NLP)** model capable of accurately classifying the sentiment expressed in user reviews. The system should be able to distinguish between **positive** and **negative** sentiments to provide valuable insights into customer opinions.

The project aims to deliver a **sentiment classifier** that accurately analyzes and categorizes user reviews, providing businesses and users with valuable insights into the sentiment landscape surrounding products, services, or content.

Background

In today's digital landscape, online reviews wield substantial influence over consumer choices. The manual analysis of this vast content is impractical, prompting the need for automated solutions. Companies across various sectors, such as **Amazon** and **Netflix**, employ sentiment analysis to swiftly grasp customer sentiments. For instance, Amazon utilizes sentiment analysis to enhance product recommendations, while Netflix gauges audience reactions to refine its content offerings.

Traditional rule-based systems struggle with the nuanced language of user reviews. Advanced machine learning techniques, including deep learning and transformer models, show promise in overcoming these challenges. This project aims to leverage these techniques, providing businesses with actionable insights for quick responses to customer concerns and empowering users to make informed decisions.

The significance of this project extends beyond specific industries. As users continue to contribute opinions online, an automated sentiment classifier becomes a crucial tool for distilling meaningful insights and improving user experiences.

2 Methodology

Design and Implementation of Sentiment Classifier

1. Data Collection:
Source Selection: Collect a diverse dataset of online reviews from Kaggle, ensuring representation of positive and negative.

2. Data Preprocessing:
Text Cleaning: Remove noise(emojis, null values,special char) and irrelevant information(date).
Merging: Merged two datasets
Tokenization and Lemmatization: Break down text into tokens and reduce words to their base form for consistent representation.
Handling Imbalances: Address class imbalances, ensuring fair representation of sentiments.

3. Feature Engineering:
Text Vectorization: Utilize techniques like TF-IDF for traditional machine learning models.

4. Model Selection:
Baseline Models: Implement many baseline models like SVM, KNN, Decision Tree, NBM, MNBM for benchmarking.

5. Training and Testing:
Splitting Dataset: Divide the dataset into training and test sets.
Model Training: Train models on the training set and fine-tune hyperparameters for optimal performance.

6. Evaluation Metrics:
Performance Metrics: Assess model performance using accuracy, precision, recall, F1-score, and confusion matrices.

7. Hyperparameter Tuning:
Grid Search: Explore hyperparameter combinations for optimal model tuning.

8. Deployment:
Integration: Develop a user-friendly interface or integrate the model into existing platforms using suitable deployment tools (e.g. Streamlit).

3 Experiments

i) Models Used(Accuracy):

svcModel:	0.9534457478005866
knnModel:	0.7085777126099707
DEcisonTreeClassifierModel:	0.908724340175953
naiveBayesModel:	0.7342375366568915
LogisticRegressionModel:	0.9527126099706745
RandomForestModel:	0.9406158357771262
MultiNomialNaiveBayesbModel:	0.9420821114369502
xgbModel:	0.9266862170087976

ii) Finding Best parameters of best Model(SVC):

Using GridSearchCV

Results: {'C': 1, 'gamma': 1, 'kernel': 'rbf'}

4 Results & Discussion

Classification Report:

Accuracy:	0.9542
Hinge Loss:	0.2090

Sentiments	precision	recall	f1-score	support
Negative	0.94	0.97	0.96	1378
Positive	0.97	0.94	0.95	1350

ROC/AUC Score	0.9882

Confusion Matrix:

0 [[1335, 43]

Actual

1 [82, 1268]]

0 **Predicted** **1**

The Confusion matrix shows that 43 of sentiments that were negative were predicted postive and 82 of the positives where predicted negative.

ROC/AUC Score tends to be near 0.5 if the prediction of the model is a mere random guess. The closer it is to 1 means that it has a good measure of separability. Our model has .9882 ROC/AUC Score

Hinge loss is a function popularly used in support vector machine algorithms to measure the distance of data points from the decision boundary. The closer it is to 0 the better. Our model has 0.209 hinge loss

Precision by definition mean the ratio of correctly predicted postive instances over total predicted positive instances. In our case the ratio is .94 for positive and .97 for negative which is good.

Recall means ratio of correctly predicted instances over actual number of instanes.

Conclusion and Future Work

What are the long-term implications of your findings? Wrap up your discussion succinctly while

pointing out the significance of your work as well as it what it means for the fields you examined

as much as possible. Lastly, suggest ideas for future studies that could build on your work, and

justify why they might be useful. Otherwise, you're all done!

References

[1] Aurelien Geron. Hands on Machine Learning with Scikit-Learn, Keras & TensorFlow.

[2] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning