Sentiment Analysis in Urdu Language

CSE422: Artificial Intelligence

Submitted By

Team 6

Name	ID
Malika Muradi	21241057
Basit Hussain	21141064

Submitted To

Syed Zamil Hasan Shoumo, Mehran Hossain

Department of Computer Science

BRAC University Dhaka

December, 2023

CONTENTS

0.1 Introduction 1

0.2 Dataset Description 1

0.2.1 Source 1

0.2.2 Dataset Description 1

0.2.3 Imbalanced Dataset 2

0.3 Dataset Pre-Processing 3

0.3.1 Deal with Fault Values and removel of emojis 3

0.3.2 Removal of stop words, ... 3

0.3.3 Tokanization 3

0.3.4 Stemming 3

0.4 Future Scaling 4

0.5 Dataset Splitting 4

0.6 Model Training and Testing 4

0.7 Model Selection Comparison and Analysis 4

0.7.1 Accuracies comparison 4

0.7.2 Precision, recall comparison 5

0.7.3 Confusion matrix 6

0.8 Conclusion 6

0.1 INTRODUCTION

Understanding sentiment analysis is a crucial task in NLP which leads businesses to devise effective strategies in today's competitive landscape. However, sentiment analysis in languages like Urdu presents unique challenges due to the scarcity of natural language processing (NLP) solutions tailored to these languages. The study involved data collection from various social media platforms to gather a comprehensive corpus of Urdu text. The data was then preprocessed to clean, normalize, and transform it. The study presents a comprehensive annotated dataset for analyzing market demand in Urdu-speaking regions. Our study used three machine learning models, such as Support Vector Machines (SVM), Logistic Regression, and Naive Bayes. The effectiveness of the approach was evaluated through various algorithms. Logistic regression with 90.92% accuracy, SVM with 88.89% accuracy, and Naive Bayes 86.68% performed consistently well.

0.2 DATASET DESCRIPTION

0.2.1 SOURCE

We used the Instant Data Scraper tool to gather comments about laptops from a variety of social networking sites. To enable further analysis, the unsupervised data was subsequently saved in a CSV file. Additionally, we used a Python web scraper to collect data about laptop models from Wikipedia for our second dataset.

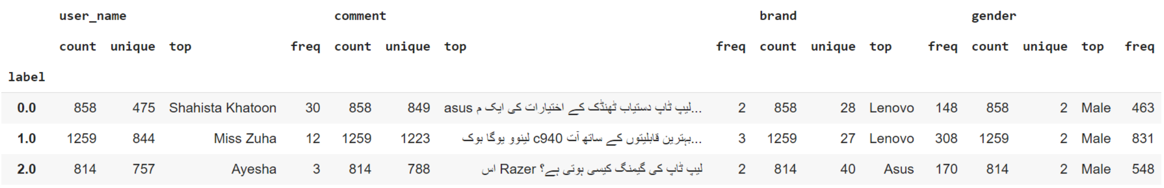
Dataset Link: www.sentiment analysis in urdu language.dataset.github

0.2.2 DATASET DESCRIPTION

We extracted raw data from social media platforms, specifically laptop-related YouTube channels and public pages on Facebook. These platforms served as valuable sources for gathering user preferences and feedback. Utilizing the Instant Data Scraper tool, we amassed a substantial dataset comprising over 37,000 raw entries, primarily consisting of comments from various social media sites.

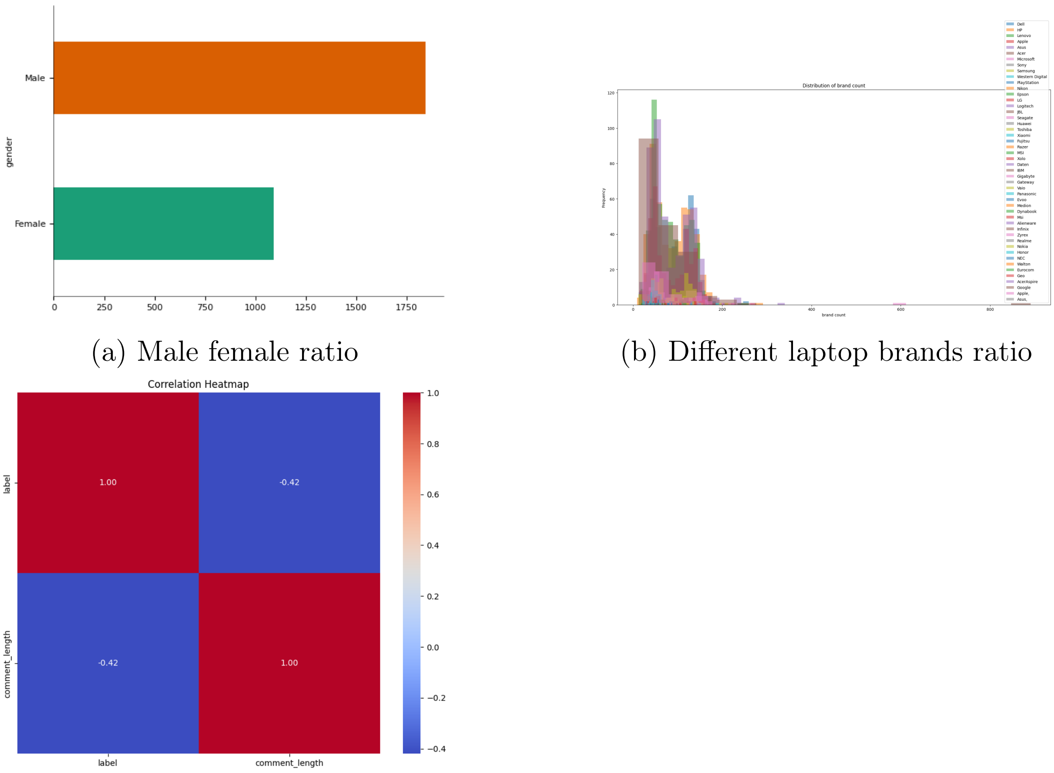
In the process of refining our dataset for sentiment analysis using Natural Language Processing (NLP) in the Urdu language, we systematically eliminated irrelevant entries. This curation served as the foundation for our investigation, focusing on a classification problem. We envision that this dataset will offer meaningful insights into consumer behavior and product preferences within the Urdu-speaking market.

Following the removal of irrelevant comments, our dataset now comprises 3,045 rows and 5 columns, encompassing information such as username, comment, brand, gender, and label. Notably, the categorical nature of the future data is recognized, emphasizing its potential for classification purposes.



	user_name				comment				brand				gender			
	count	unique	top	freq	count	unique	top	freq	count	unique	top	freq	count	unique	top	freq
label																
0.0	858	475	Shahista Khatoon	30	858	849	...لیپ ٹاپ دستیاب ٹھنڈک کے اختیارات کی ایک م asus	2	858	28	Lenovo	148	858	2	Male	463
1.0	1259	844	Miss Zuha	12	1259	1223	...بہترین قابلیتوں کے ساتھ آت c940 لینوو یوگا بوک	3	1259	27	Lenovo	308	1259	2	Male	831
2.0	814	757	Ayesha	3	814	788	لیپ ٹاپ کی گیمنگ کیسی ہوتی ہے؟ Razer اس	2	814	40	Asus	170	814	2	Male	548

Figure 1: Description of dataset



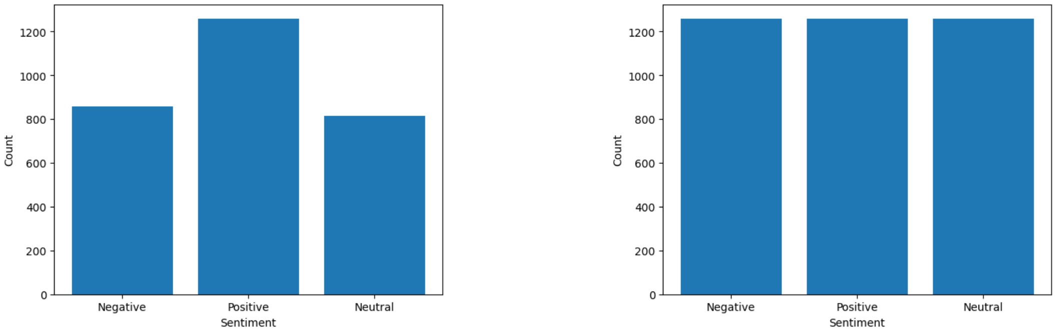
(a) Male female ratio

(b) Different laptop brands ratio

(c) Correlation between comment and label

0.2.3 IMBALANCED DATASET

In the initial class distribution of our dataset, we observed notable imbalances, with Class 1.0 having 1259 instances, Class 0.0 with 858 instances, and Class 2.0 with 814 instances as shown in figure above. Recognizing the potential impact of class imbalance on model training and performance, we implemented data balancing techniques. After balancing, the revised class distribution showcases a more equitable representation, with each class now having 1259 instances. This balancing process ensures that our model is exposed to a more uniform distribution of instances across classes during training, mitigating the risk of bias towards any particular class.



(a) Before data augmentation

(b) After data augmentation

Figure 3: Data augmentation by random oversampling technique

0.3 DATASET PRE-PROCESSING

0.3.1 DEAL WITH FAULT VALUES AND REMOVEL OF EMOJIS

To ensure the cleanliness of our text data, we implemented a process to remove punctuation marks, Null value and extra columns. Using the regular expression library in Python, we effectively eliminated punctuation and special characters from the dataset. In addition, we eliminated all emojis and null values found in the dataset for Urdu. This method was essential for keeping the text data clean and prepared for final processing and analysis

0.3.2 REMOVAL OF STOP WORDS,

Furthermore, we concentrated on deleting stop words from the data set in order to enhance the quality of our text data and enable more precise analysis. To do this, we extracted these unnecessary phrases from our dataset using a complete list of Urdu stop words that we collected from a Kaggle dataset . By taking this action, we were able to clean up the text data and make sure that the analyses that came next were built on more insightful and pertinent content.

0.3.3 TOKANIZATION

Then we separated the text into individual words or tokens. Through this method, we were able to separate the comments into their component parts for additional study. This method ensured that the text data was properly prepared for extensive analysis and future language processing activities.

0.3.4 STEMMING

Stemming reduced word variants and standardized them to their base or root forms by removing prefixes and suffixes. In conclusion, data collection, punctuation removal, stop-word removal, tokenization, and stemming were some of the crucial processes involved in our data preprocessing. By following these processes, we were able to create a polished dataset that was a useful tool for sentiment analysis using machine learning methods.

0.4 FUTURE SCALING

StandardScaler is used to standardize the features after splitting the data into training and testing sets. It's essential to fit the scaler on the training set and use the same scaler to transform both the training and testing sets to prevent data leakage.

0.5 DATASET SPLITTING

To assess the performance of our model, we divided our dataset into two essential subsets by stratify technique: the training set and the test set. The training set, constituting 70% of the data, was utilized to train and build the model, allowing it to learn patterns and relationships within the dataset. The remaining 30% of the data was reserved for the test set, which serves as an independent dataset for evaluating the model's performance. By using this distinct test set, we aim to assess how well our model generalizes to new, unseen data.

0.6 MODEL TRAINING AND TESTING

We used a variety of machine learning models for sentiment analysis. For our models, we conducted a train-test split using the train_test_split function, setting the test size to 0.3. Our models including Logistic Regression, Support Vector Machine (SVM), and Naive Bayes. we trained our models using scikit-learn library.

0.7 MODEL SELECTION COMPARISON AND ANALYSIS

0.7.1 ACCURACIES COMPARISON

In this section, we present the results of our performance evaluation for a variety of machine learning models applied to our sentiment classification task. This allowed us to assess the models' generalization performance on unseen data. Logistic regression performed well, with an accuracy rate of 90.92%, SVM Accuracy was 88.89%, and Naive Bayes Accuracy was 86.68%. These models demonstrated their proficiency in understanding and classifying sentiment patterns.

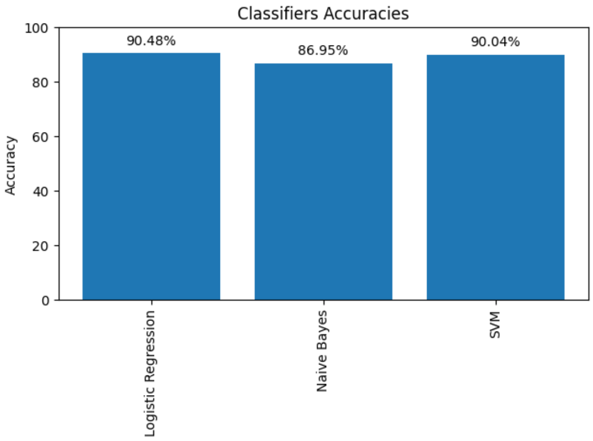


Figure 4: Accuracy comparison of the Models

0.7.2 PRECISION, RECALL COMPARISON

We provide a detailed analysis of each model's performance, including precision, recall, F1-score, and accuracy. The Fig 5 below, summarizes the comparative analysis of these models. Logistic Regression and Support Vector Machine (SVM) models achieved remarkable consistency in performance across all evaluation metrics.

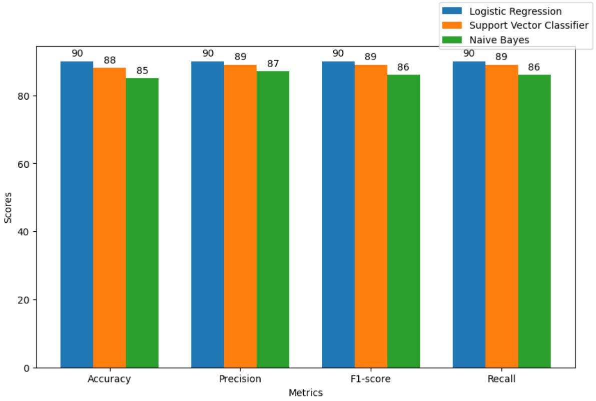
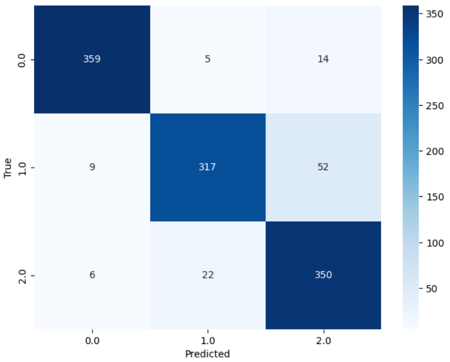
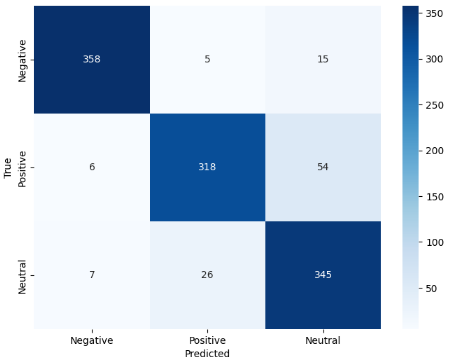


Figure 5: Precision, recall comparison of the Models

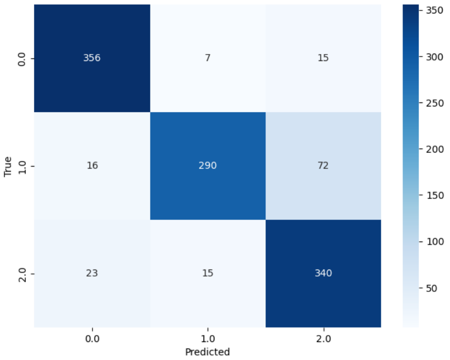
0.7.3 CONFUSION MATRIX



(a) Confusion metric for Logistic regression



(b) Confusion metric for SVM



(c) Confusion metric for Naive Bayes

0.8 CONCLUSION

In this project, the focus is on addressing the sentiment analysis challenge in the Urdu language using three machine learning models. The dataset, compiled from social networking sites and Wikipedia, undergoes a meticulous pre-processing phase, including the removal of faulty values, emojis, and punctuation, as well as the elimination of stop words, tokenization, and stemming. Feature scaling is achieved using the StandardScaler, and the dataset is split into training and test sets. Leveraging Logistic Regression, Support Vector Machine, and Naive Bayes models, the sentiment analysis is conducted. The subsequent evaluation reveals the models' performances, with Logistic Regression leading in accuracy at 90.92%, followed by SVM (88.89%) and Naive Bayes (86.68%). This comprehensive approach ensures a robust exploration of sentiment patterns in the Urdu-speaking market, offering valuable insights into consumer behavior and product preferences.