NATIONAL UNIVERSITY
of Computer & Emerging Sciences

Data Analysis and Visualization

Project Documentation

Nareen Waseem	21L-6270
Omamah Shahzad	21L-5638
Fiza Shahzad	21L-7693

Project Title

Sentiment Analysis of Smartphone Reviews on Daraz

Problem Statement:

In the context of e-commerce platforms like Daraz, people are uncertain about the authenticity of products. As the marketplace offers a vast variety of products from various sellers, users often face uncertainty regarding the legitimacy of items they intend to purchase. This uncertainty is a significant barrier to customer trust and poses risks.

Purpose of the Project:

The purpose of this project is to conduct sentiment analysis on smartphone reviews from the Daraz platform, addressing the issue of uncertainty surrounding product authenticity. By employing Natural Language Processing (NLP) techniques, the project aims to extract meaningful insights from user reviews, contributing to a more secure and trustworthy online shopping experience.

Scope of the Project:

The scope includes collecting a comprehensive dataset of smartphone reviews, performing data preprocessing and cleaning, employing NLP techniques for sentiment analysis, categorising sentiments into negative, neutral, or positive, identifying patterns in user reviews, and providing insights related to product authenticity.

Target Audience:

Analysts, developers, and stakeholders interested in understanding customer sentiments and improving the trustworthiness of e-commerce platforms, particularly Daraz.

Technologies Used:

- Python

- Selenium library
- Natural Language Processing (NLP) libraries (e.g, NLTK)
- Machine learning libraries (e.g, scikit-learn)
- Deep learning libraries (e.g, TensorFlow, Keras)
- Plotly library

Methodology:

1. Data Scraping:

We begin by scraping a diverse dataset of smartphone reviews from the Daraz platform. The dataset covers various brands (samsung, vivo, infinix, realme, and tecno), and user sentiments to ensure the model's robustness.

2. Data Wrangling:

Our approach involves a series of preprocessing steps tailored to the nature of the algorithm. The dataset undergoes the removal of stopwords, elimination of unwanted characters and symbols, removal of emojis, conversion of roman urdu comments to english and normalisation and standardisation procedures. Following this preprocessing, the cleaned data is tokenized and padded to facilitate the learning process.

3. Feature Extraction:

We aim to use the Countvectorizer, Bag of Words (BOW) and tf-idf (Term Frequency - Inverse Document Frequency) technique in our project. Using CountVectorizer, Bag of Words, and TF-IDF in our project offers simplicity, interpretability, and efficiency in representing text data, serving as effective baseline models for comparison with more advanced techniques.

4. Modelling:

Finalising the model for sentiment analysis is done using comparative analysis. To determine which model is suitable for our project, we did a comparison between; three Machine learning models (Naive Bayes Classifier, Logistic Regression and Support Vector Machine) and two Deep learning models (Multi Layer Perceptron Classifier and Recurrent Neural Network).

Naive Bayes Classifier (NB Classifier) is a probabilistic Machine Learning model based on Bayes theorem. It is particularly suited for text classification tasks, including Sentiment Analysis.

Logistic Regression is a versatile statistical method used in machine learning for binary and multiclass classification tasks, leveraging the logistic function to model the probability of an instance belonging to a particular class.

Support Vector Machine (SVM) is a supervised machine learning model used for classification and regression tasks. It locates the hyperplane in the feature space that best splits the data into different categories.

The Multilayer Perceptron (MLP) classifier serves as a robust tool for sentiment analysis tasks, leveraging its capacity to learn intricate patterns and relationships within data. With multiple layers of neurons, it can discern complex structures in textual information, making it well-suited for extracting nuanced sentiment features from diverse sources like news headlines.

Recurrent Neural Network (RNN) is particularly effective in capturing sequential dependencies in data. Given that product reviews often have inherent sequential structures, RNN can be beneficial. It processes input sequentially, maintaining a hidden state that retains information about previous inputs.

5. Evaluation:

After implementing various machine learning and deep learning models for Sentiment Analysis, it's crucial to assess their performance using appropriate evaluation metrics. The chosen metric is accuracy, which provides a comprehensive view of the model's effectiveness.

6. Comparison:

In the analysis of these models, Multi Layer Perceptron Classifier (MLP) and Recurrent Neural Network (RNN) have been concluded as an efficient and better Machine learning and Deep learning model to analyse user sentiments on product reviews.

Results:

The results extracted from the models are shown.

Models	Accuracy
Naive Bayes Classifier	80.60
Logistic Regression	86.91
Support Vector Machine (SVM)	86.07
Multi Layer Perceptron Classifier (MLP)	89
Recurrent Neural Network (RNN)	88

Conclusion:

In conclusion, this project addresses the uncertainty regarding product authenticity on Daraz by conducting sentiment analysis using NLP techniques. By providing insights into customer sentiments, this research contributes to enhancing trust in Daraz, making the online shopping experience more reliable and trustworthy for both customers and sellers.