Machine Learning And Natural Language Processing Applications In Urdu Product Market Sentiment Analysis

Malika Muradi[a], Basit Hussain[a], Md Humaion Kabir Mahedi*[a], Sania Azhmee Bhuiyan[a], Ehsanur Rahman Rhythm[a], Annajiat Alim Rasel[a]

[a]*Department of Computer Science and Engineering, School of Data and Sciences, BRAC University, 66 Mohakhali, 1212, Bangladesh*

Abstract

Understanding product market sentiment is crucial for businesses to devise effective strategies in today's competitive landscape. However, analyzing market sentiment in languages like Urdu presents unique challenges due to the scarcity of natural language processing (NLP) solutions tailored to these languages. The main objective of this research paper is to address this gap by investigating the application of NLP techniques, including sentiment analysis, named entity recognition, and gender prediction, for product market sentiment analysis in Urdu. Our study involves utilizing five machine learning models, such as support vector machines (SVM), logistic regression, multinomial naïve bayes, random forest, and gradient boosting, as well as three deep learning models: recurrent neural network (RNN), convolutional neural network (CNN), and bidirectional long short-term memory (Bi-LSTM). We evaluate the effectiveness of the approach through various algorithms. Among the machine learning models, SVM, logistic regression, and random forest perform consistently well with 93% accuracy, while RNN excels among deep learning models with 93% accuracy. Finally, this research also implements the local interpretable model-agnostic explanations (LIME) XAI method

*Corresponding author

Email addresses: malika.muradi@g.bracu.ac.bd (Malika Muradi), basit.hussian@g.bracu.ac.bd (Basit Hussain), humaion.kabir.mehedi@g.bracu.ac.bd (Md Humaion Kabir Mahedi*), sania.azhmee.bhuiyan@g.bracu.ac.bd (Sania Azhmee Bhuiyan), ehsanur.rahman.rhythm@g.bracu.ac.bd (Ehsanur Rahman Rhythm), annajiat@gmail.com (Annajiat Alim Rasel)

to improve understanding of sentiment classification, name entity recognition, and gender prediction.

Keywords: Market sentiment Analysis, Urdu Text, Natural Language Processing, Explainable Artificial Intelligence, Name Entity Recognition, Sentiment Analysis.

1. Introduction

Urdu is a language that is spoken by millions of people worldwide [1]. It is a mix of Indo-European, IndoIranian, and IndoAryan languages, with influences from Persian and Arabic [2]. Urdu is a rich and complex language with a long history, but it has been less studied than European languages. It is an important language with a rich history and culture, and it deserves more attention from researchers [3]. A notable aspect of Urdu is its popularity among speakers, who prefer to communicate in their native tongue, sharing opinions and feedback on various social media platforms and websites. In the contemporary era, with advancements in technology and evolving business strategies, more customers are turning to online platforms for purchases and subsequently sharing their feedback through various social media channels. Consequently, a vast trove of Urdu text data now exists online, presenting both opportunities and challenges for businesses as they seek to harness and preprocess this data for their strategic endeavors.

NLP can now perform sentiment analysis, parts of speech tagging, unnecessary word deletion, parsing, name entity recognition, and many more [4]. NLP has made a lot of progress in English, but there is still more work to be done in Urdu [5]. Because the business environment is changing continuously, it is very important for organizations to have a well-defined strategy in order to thrive in an increasingly competitive marketplace. A solid business plan not only provides guidance, but it also determines how new businesses are regarded. Entrepreneurs must grasp the top-selling gender-specific items on the market in order to develop a profitable business in a highly competitive sector. A good plan is necessary to sell your product to target buyers and understand their interests. Opinion analysis is a strong tool that assists entrepreneurs in gathering, measuring, and analyzing consumer opinion in order to develop effective business strategies [6]. Social media is becoming a significant industry, allowing businesses to engage with customers and track their activity.

This paper intends to thoroughly investigate the methods and tools used to understand market sentiment in Urdu using NLP. We introduce a predictive model using social media data and NLP algorithms that accurately forecasts demand for laptop brands based on consumer reviews. Our objective is to equip business owners with a deep understanding of competitive dynamics, aiding informed decision-making. Moreover, our model identifies popular products via gender and sentiment analysis, offering valuable insights for companies seeking to enhance their offerings to meet customer needs.

To achieve this,

- We will utilize automated data scraping techniques to gather raw text data in Urdu from various social media buy and sell groups. After filtering and organizing this data with NLP, we will create a well-structured dataset.
- We will use machine learning algorithms, particularly named entity recognition and gender prediction, and sentiment analysis.
- The accuracy of our models will validate using a comprehensive test dataset, will be providing actionable insights regarding the most sought-after products based on gender and sentiment analysis.
- For deeper understanding of the models, we will implement the LIME XAI method for sentiment analysis, NER, and gender prediction.

In this paper, we organize our paper as follows: Firstly, section 1 introduces the problem and study objectives. Subsequently, in section 2, we delve into a comprehensive review of related work in the field. Moving on to Section 3, we briefly explore the methodology and the paper's model implementation. and we also elaborate on the datasets we employed. in Section 4, we delve into the sentiment analysis. Section 5, meanwhile, is dedicated to the presentation of our results and discussion. Furthermore, in Section 6, we delve into the implementation of Explainable AI. Finally, in Section 7, we conclude our findings and offering insights into future directions.

2. Related Work

The analysis of social media data using sentiment analysis and natural language processing techniques have drawn a lot of attention in recent years. In order to comprehend consumer behavior and predict market sentiment for certain items, researchers have looked at a variety of techniques for collecting, processing, and analyzing online social media data [7]. Developing digital platforms has a significant impact on the stock market. Studies show that sentiment shared on social media has a greater and more influence on stock returns compared to the sentiment found in traditional news sources [8]. Huge amounts of text data can now be collected and analyzed by the internet, which may help us better understand the opinions and ideas of people. This information might be used for a number of reasons, such as market research, customer satisfaction surveys, and political polling [9]. The researchers in another publication by [10] stated that machine learning has the potential to enhance customers' online shopping experiences by enabling them to find product reviews sorted by the proportion of both positive and negative comments provided by other clients. Moreover, according to the research by [11] mention that NLP has a significant impact on businesses by enabling applications and software that address language barriers in international trade, handle customer service inquiries, and provide commercial artificial intelligence assistants to enhance operational efficiency.

The study referenced as [12] emphasizes the transformative potential of text summarization as a tool to condense extensive reviews into concise sentences, all the while ensuring the retention of crucial concepts within the content. This effectiveness is further amplified through the strategic integration of the seq2seq model, LSTM, and attention mechanisms. Shifting the focus to the domain of finance, the application of machine learning techniques is exemplified by the LSTM model's adeptness in predicting mean squared error values, showcased through its successful application in analyzing time-series data encompassing stock prices and returns, as mentioned in [13]. similarly, in the research paper [14] addressed the increasing significance of sentiment analysis in regional languages, particularly Urdu, on the internet, and introduced a framework that leveraged deep learning techniques to achieve significant accuracy improvements, with the BiLSTM-ATT model outperforming others with an accuracy of 77.9% and an F1 score of 72.7%. In another paper [15], the authors delved into the realm of NLP to address the complexities of contextual sequence labeling for low-resource languages, with a particular emphasis on NER in Roman Urdu. Despite the limited prior research in NLP applications for Roman Urdu, the paper conducted a comprehensive comparative analysis of deep learning-based models, showcasing the remarkable performance of Bi-LSTM with an F1-score of 82.7%. This study underscored the potential for achieving sophisticated contextual understanding in processing morphologically rich, low-resource languages like Roman Urdu. In addition, in this research [16], sentiment analysis in Urdu, the national language of Pakistan, was explored using machine learning techniques on Twitter data in both Urdu Script (Nastaleeq) and Roman Urdu. The social media optimization was identified as the best algorithm for Urdu (Nastaleeq) tweets, while random forest performed well for Roman Urdu tweets, addressing a gap in Urdu sentiment analysis research amid the growing digital landscape.

The intricate nature of stock markets, driven by an intricate interplay of countless variables, presents a substantial challenge in terms of achieving accurate forecasting and comprehensive understanding. Expanding the purview to encompass the burgeoning influence of social media platforms on financial markets, as discussed in [17], underscores the pivotal role of sentiment analysis harnessed from an array of digital platforms. This sentiment analysis emerges as a potent input for constructing robust forecasting frameworks, with the stand-alone fuzzy neural network (SOFNN) algorithm, expounded in [18], standing out due to its exceptional accuracy in performing sentiment analysis tasks.

Furthermore, the convergence of diverse natural language processing elements with sentiment analysis not only contributes significantly to shaping prevailing emotional tones, attitudes, and opinions across digital spaces but also plays a pivotal role in quantifying the proportions of positive, negative, and neutral sentiments within ongoing and trending discussions on various social media platforms, as discussed in [19]. Within the realm of related research, the work conducted by [20] gains prominence as it underscores the practicality of employing sentiment analysis to identify popular entities and consumer preferences, thereby providing a foundation for informed decision-making within distinct linguistic contexts. Similarly, the another work by [21] introduces an innovative unified model that synergistically merges convolutional neural networks and long short term memory. This model serves as an advanced approach to sentiment analysis in the context of tweets, effectively capturing the intricate contextual nuances and sequential intricacies inherent to such short-form content. Furthermore, this application goes beyond sentiment analysis, providing a deeper grasp of public opinions and trends on social media. This expands the understanding of modern digital discussions. To conclude, the cited research collectively underscores the interplay between advanced technology, sentiment analysis, and their substantial impact on text summarization, financial predictions, sentiment measurement, and public sentiment analysis across digital realms.

Understanding market sentiment is crucial for shaping product development, marketing strategies, and resource allocation. While English markets are well-explored, comprehending Urdu-speaking dynamics is vital. This research addresses NLP limitations in sentiment analysis for Roman Urdu, proposing a new dataset and exploring machine learning algorithms [22]. Existing research primarily focuses on English markets, neglecting Urdu's vast global reach. Urdu, spoken in Pakistan and India, offers significant market potential [23]. Unveiling unique preferences and cultural nuances is key for effective strategies. Despite challenges, delving into Urdu market analysis presents opportunities for researchers and businesses in untapped markets.

Beyond the stock market prediction, the preceding research highlights, the expanding importance of natural language processing in analyzing consumer behavior, forecasting market trends, and assessing public opinion using social media data. The studies mentioned above illustrate an importance of social media sentiment on stock markets as well as the promise of machine learning, and deep learning models in applications such as stock prediction and sentiment analysis. In digital conversations, NLP integration shapes emotional tones and quantifies sentiment, including in specific linguistic contexts like Urdu-speaking regions.

3. Methodology

In this study, we present an approach to analyze opinions and genders of social media users regarding various laptop brands in Urdu-speaking regions. The primary objective is to identify the most popular and sought-after laptop brand in these regions. As shown in Fig. 1, the primary components of our system are sentiment analysis, named entity recognition, and gender prediction.

To conduct product market sentiment analysis in Urdu, we faced a lack of existing research and organize data related to customer product reviews. To address this challenge, we turned to social networking platforms, specifically laptop-related YouTube channels and Facebook public pages, as valuable sources for collecting consumer preferences and reviews. Using the Instant Data Scraper tool, we gathered a substantial volume of comments from diverse social networking platforms, resulting in a dataset comprising over 37,000 raw entries. In order to identify laptop brand names within the comments, we supplemented our data with laptop brand and model names obtained from Wikipedia using a Python web scraper. Furthermore, to ensure that we have data on laptop model names in both Urdu and English, we utilized the Google API to translate device model names. For the purpose of NER, we devise a function that used a dictionary-based system to identify device names through comparison with the laptop list and the comments' data.

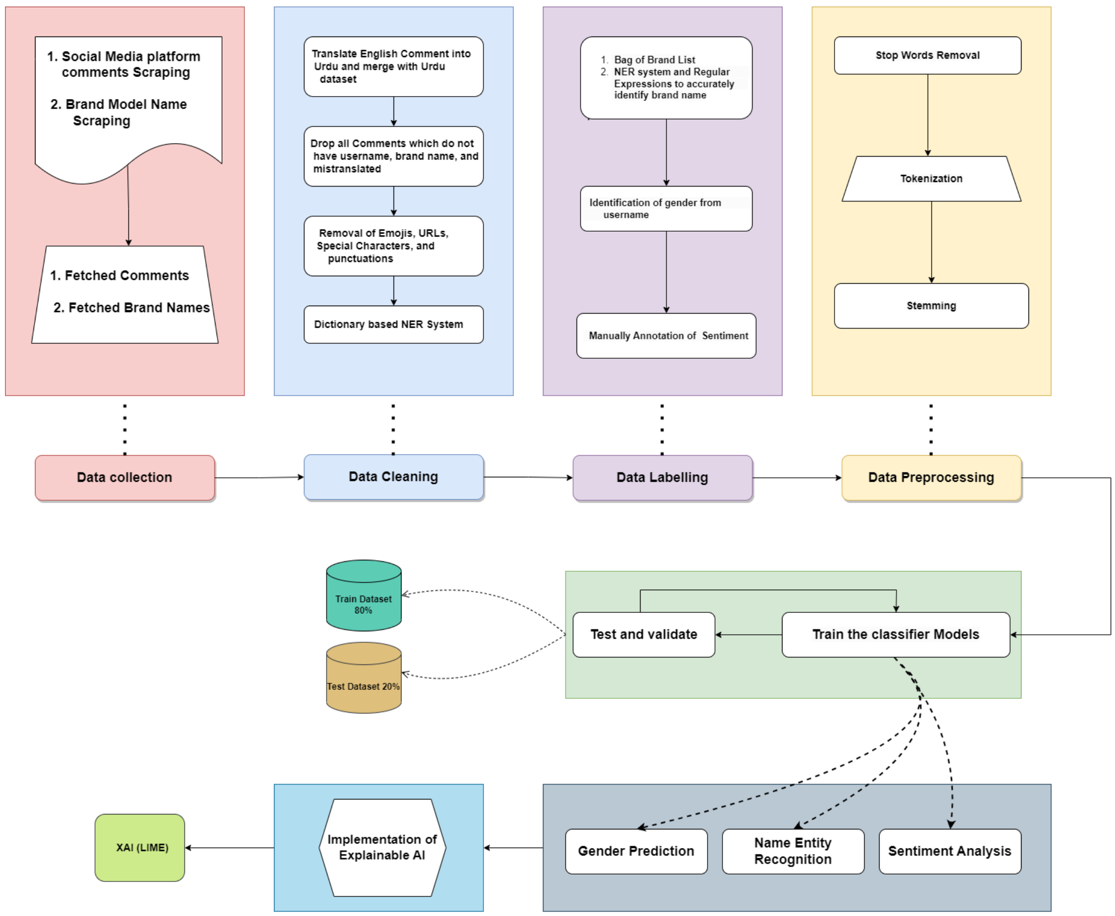

Figure 1: A Visual Work Plan for Analyzing Product Sentiment in Urdu Language by utilizing natural language processing techniques.

To facilitate comprehensive analysis, we categorized the comments into separate Urdu and English segments. This involved translating English comments into Urdu.

We also performed data cleaning by excluding rows without usernames and comments that did not mention any brand names. Subsequently, the comments were categorized into negative, positive, and neutral sentiments with the assistance of native Urdu speakers. For predicting gender from username, we utilized the gender guesser Python library; however, as the predefined model did not support Urdu names, we found a viable solution by translating Urdu names to English using the Google Cloud Translation API. Nevertheless, the library still could not predict genders accurately, so we relied on using only the first names. To further process the data, we applied various NLP techniques for cleaning and then split it into training, test, and validation sets test. We conducted sentiment analysis using different machine learning and deep learning algorithms. Finally, on the bases of genders in the current market, we plotted a list of the most demanding devices.

By employing a rigorous and systematic approach, we dedicated substantial effort to curate extensive Urdu datasets. These datasets have been carefully crafted to unlock valuable insights into the intricate realm of consumer preferences. By delving into the nuances of these datasets, we aim to gain a profound understanding of market dynamics, allowing us to discern trends, preferences, and behaviors that shape consumer choices. In Fig. 2, we visualized the intricate process of model implementation.

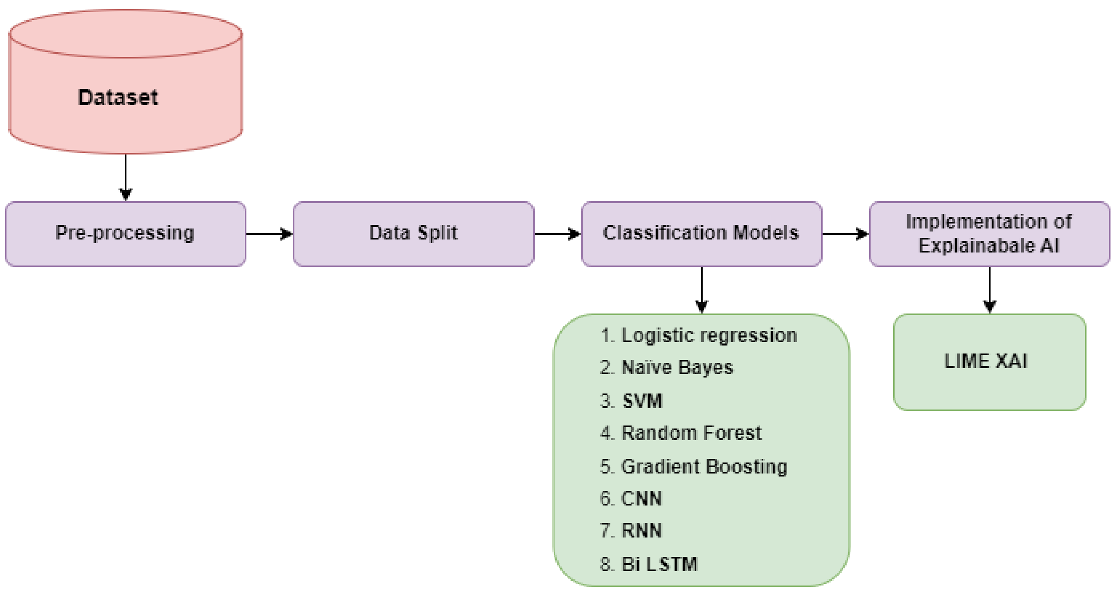

Figure 2: Model Implementation: From Data Collection to Model Implementation and Explainable AI for In-Depth Market Sentiment Analysis.

3.1. *Data Cleaning and Preprocessing Techniques*

3.1.1. *Punctuation and emoji removal*

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

3.2. *Data Cleaning and Preprocessing Techniques*

3.2.1. *Punctuation and emoji removal*

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

3.2.2. *Stop-word removal*

Stop words are words that are often used but have little to no real significance in a language. They may interfere with sentiment analysis by adding noise. 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 [24]. 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.

3.2.3. *Tokenization*

Tokenization is the process of breaking down text into individual tokens. It is a crucial stage in natural language processing. In our study on market sentiment analysis in Urdu, 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.

3.2.4. *Stemming*

In our research, stemming was essential since it helped standardize the language that was taken from tokenized comments. Stemming reduced word variants and standardized them to their base or root forms by removing prefixes and suffixes [25]. To do this, we used a suitable stemming technique to revert tokenized words in the comments to their original forms.

In conclusion, data collection, punctuation removal, stop-word removal, tokenization, and stemming were some of the crucial processes involved in our data pre-processing. By following these processes, we were able to create a polished dataset that was a useful tool for sentiment analysis using machine learning methods. In the end, this allowed us to learn important things about the dynamics of market demand.

3.2.5. Named Entity Recognition

Named entity recognition using a dictionary-based method is a useful tool for identifying certain items in text [26]. We used a dictionary-based NER system to extract brand-related information from written comments in our study, which is an essential task in NLP. The laptop names in the experiment were in both English and Urdu. In order to solve this, we created a bilingual dictionary-based system that supported both languages. Moreover, it is important to acknowledge that customers often employ varying writing styles when referring to the same brand names. For instance, for the brand DELL, diverse renditions such as 'Dell', 'dell', 'ڈیل', and 'دیل' might be used. In light of this phenomenon, our dictionary-based NER model was intentionally designed to diligently identify writing variations associated with each brand. This approach was adopted to ensure that our system accurately extracted brand names from user comments, even when expressed in different writing styles. Based on this method, our NER algorithm sought to find specific terms associated with certain brands. Additionally, we created a function by manual coding with the help of regular expression that utilized this dictionary-based approach to match device names by comparing the laptop list with the comments' content.

3.2.6. GENDER PREDICTION

Understanding gender variances in product sentiment is crucial, since gender preferences can vary greatly. In our research, we used three gender prediction libraries to identify the gender from the username. At first, we tried to use the Python gender guesser package to determine user genders. We ran into a problem, though, because the library did not support Urdu names. We used the Google Cloud Translation API to translate Urdu names into English and make them compliant with the library in order to get around this restriction. Sadly, despite this strategy, the gender guesser library continued to make incorrect predictions. As a result, we decided to just use first names when predicting gender. This endeavor was further complicated by the presence of titles like ‘Dr’, ‘Sr’, ‘Pro’, ‘Eng’, ‘PhD’, and similar prefixes, preceding certain first names, leading to potential anomalies. To address this complexity, we implemented manual coding involving regular expressions to refine our approach. additionally, we must note that within this paper, we focused solely on the ‘Male’ and ‘Female’ labels, disregarding entries falling outside these categories, including ‘undefined’ and ‘other’.

In order to improve the precision of our gender prediction, we carried out a detailed analysis. We utilized two more gender prediction libraries, namely Genderizer and gender_guess, to evaluate how well our prediction model was performing. This thorough assessment enabled us to refine our methodology and achieve better outcomes. Despite our dedicated attempts to address this linguistic challenge, we consistently faced inaccuracies in our gender predictions. These disparities highlighted the intricacies involved in this task. After thoroughly evaluating and improving our gender prediction approach, we successfully created it. Then, we made a visually appealing pie chart in Fig. 3 to show how different genders use social media and their preferences for gadgets. The data clearly reveals that about 62.3% of men and 37.7% of women are interested in this area.

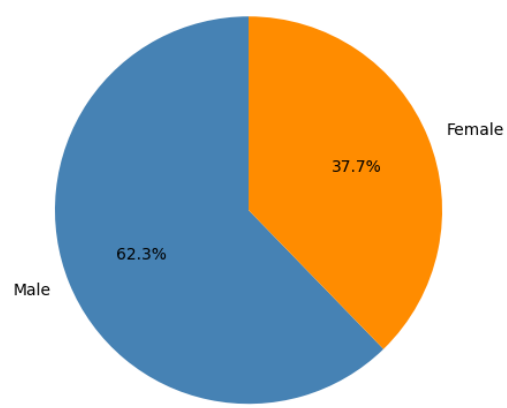

Figure 3: Male-Female Ratio: A Visual Representation of Gender Distribution with 62.3% Male and 37.7% Female Ratio.

3.2.7. SENTIMENT ANALYSIS

We used a variety of machine learning and deep learning models for sentiment analysis. For machine learning models, we conducted a train-test split using the train_test_split function, setting the test size to 0.2.

3.2.8. Dataset

We created two datasets to train and test our model. One dataset is composed of consumer product reviews from various social media platforms, while the other contains data on laptop brand names sourced from Wikipedia. We discovered a shortage of current research and organized data concerning client product reviews throughout our investigation on market sentiment analysis in Urdu. To address this issue, we turned to social media sites, namely laptop-related YouTube channels and Facebook public pages, as an excellent resource for gathering user preferences and feedback. We collected a large number of comments from several social media sites using the Instant Data Scraper tool, resulting in a dataset with over 37,000 raw entries. We removed all rows lacking usernames, brand names, or containing extraneous noise. Subsequently, with the assistance of native speakers, we meticulously labeled the dataset, resulting in a final dataset comprising 2931 rows and 5 columns. These columns encompass a range of attributes, including username, comment, gender, brand, and class, rendering it a comprehensive resource for our research.

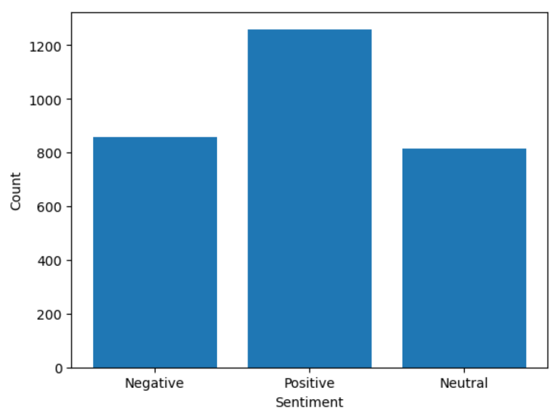

Figure 4: Sentiment polarity distribution before augmentation.

As our dataset was imbalanced, the polarity distribution in Fig. 4 revels that there are more positive sentiments than negative and neutral sentiments. To address the issue of class imbalance within the dataset, we applied an oversampling technique known as Up Sampling. This method serves to rectify unequal class distributions by augmenting the size of the underrepresented samples. Before balancing, the class distribution stood at 1259 samples for positive, 858 samples for Class negative, and 814 samples for neutral comments. Post-balancing, we achieved a balanced dataset Fig. 5 with each class containing 1259 samples. This balanced dataset is a pivotal component of our research, as it ensures that each class is equally represented and mitigates the risk of bias in our machine learning models.

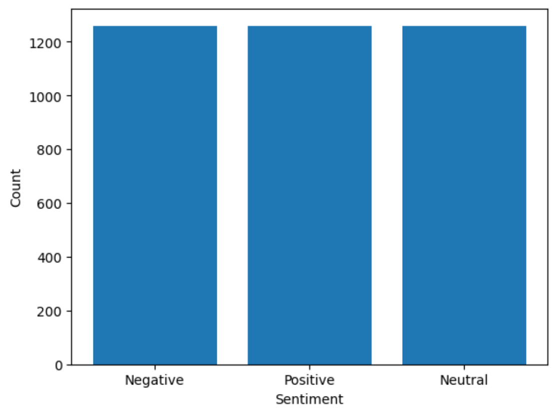

Figure 5: Sentiment polarity distribution after augmentation.

This dataset forms the cornerstone of our investigation into market sentiment analysis using NLP in the Urdu language, and we anticipate that it will provide valuable insights into consumer behavior and product preferences within the Urdu-speaking market.

4. Sentiment ANALYSIS

After conducting extensive research, we eagerly awaited the outcomes of a substantial workload, Our team generated a bar chart in below Figures that shows the percentage of men and women interested in the top laptops brands in the Urdu-speaking region market. Below, in the Fig. 6, a detailed stack bar chart presenting our predictions for the ten most preferred laptops among males and females. Surprisingly, the data highlights that Lenovo is the most preferred laptop brand for both genders. This unexpected discovery emphasizes the significant popularity and trust that Lenovo enjoys among consumers in the Urdu-speaking region. The data we have gathered not only underscores Lenovo's dominance but also highlights the nuanced preferences of men and women when it comes to their choice of laptops. To provide a more comprehensive understanding of these preferences, we have prepared a detailed bar charts in Fig. 7 and 8, that delineates our predictions for the five most favored laptop brands among both males and females in the Urdu-speaking region market. This chart offers a clear and informative visual representation of our findings, enabling stakeholders to grasp the intricacies of consumer choices in this market segment.

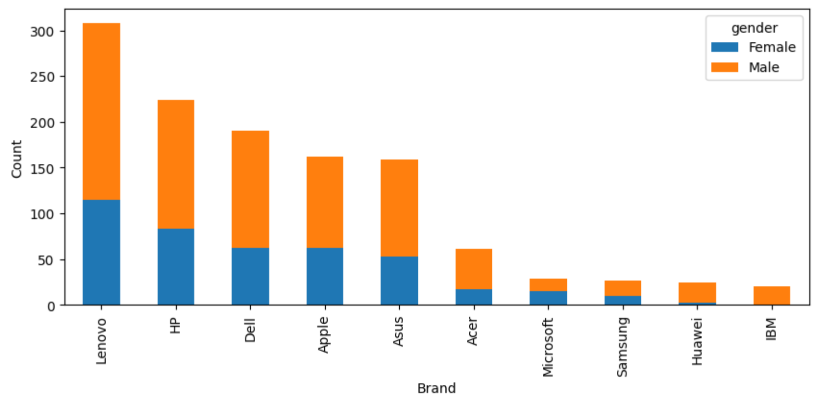

Figure 6: Ranking of top ten Laptop Brands Preferred by Both Males and Females.

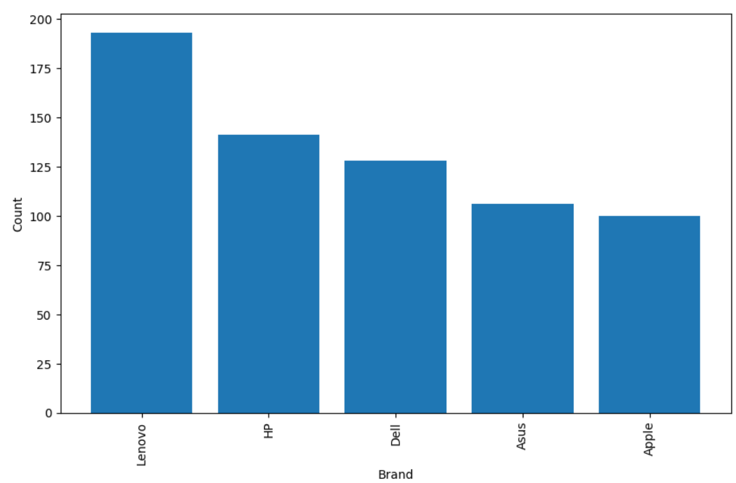

Figure 7: Ranking of top five Laptop Brands Preferred by Males.

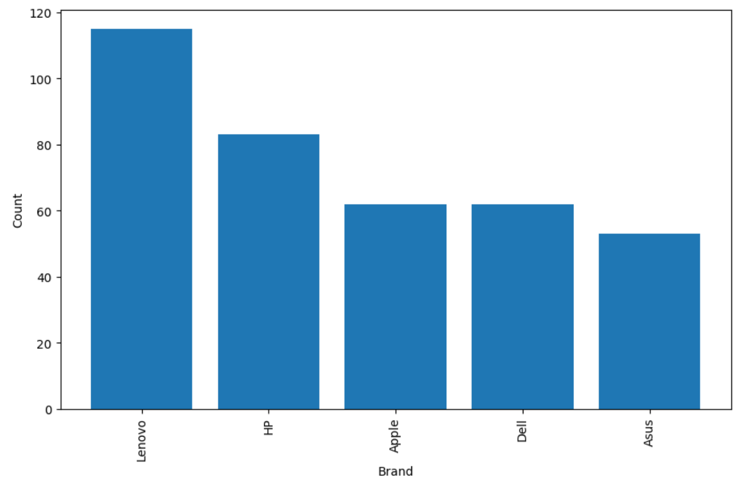

Figure 8: Ranking of top five Laptop Brands Preferred by Females.

5. Result and Discussion

Now, we present the results of our performance evaluation for a variety of machine learning and deep learning models applied to our sentiment classification task. Compared to other machine learning models, SVM, logistic regression, and random forest performed better, with an accuracy rate of 93%, while gradient boosting had an accuracy of 89% and multinomial naïve bayes Accuracy was 90%. These models demonstrated their proficiency in understanding and classifying sentiment patterns. On the other hand, the RNN performed exceptionally well in deep learning models, with high accuracies of 93%. The CNN and Bi-LSTM model had an accuracy rate of 92%. These deep learning models have a remarkable ability to interpret intricate patterns within text data, which contributes to their consistently high accuracy. These findings have critical implications for model selection and development in the domain of sentiment analysis. The accuracy comparison of all the models before and after augmentation are shown in Fig. 9, along with an illustrative word cloud displaying positive Urdu text sentiments in Fig. 10.

For machine learning models, we conducted a train-test split using the train_test_split function, setting the test size to 20% of the dataset. This allowed us to assess the models' generalization performance on unseen data. On the other hand, we trained the deep learning models for 10 epochs with a batch size of 32 and a validation split of 20%. This setup ensured that the deep learning models underwent multiple

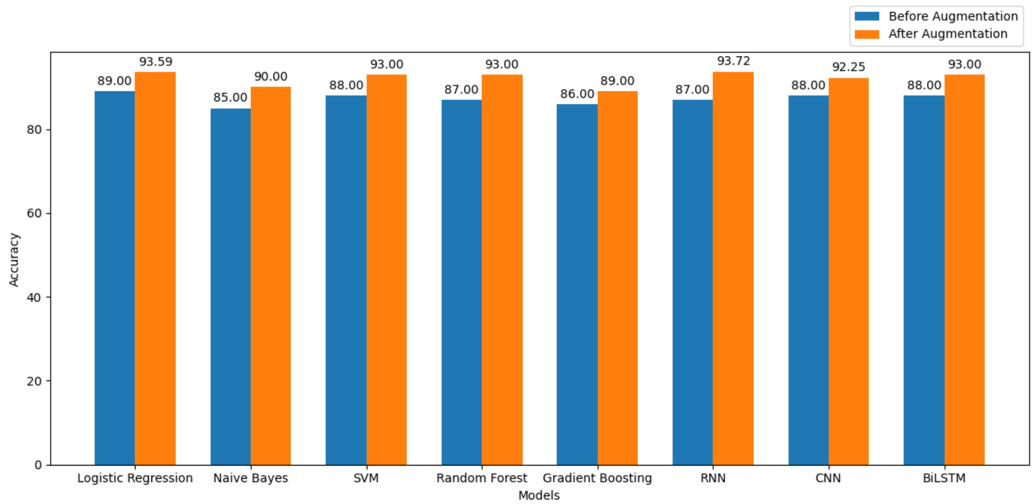

Figure 9: A Comprehensive Accuracy Comparison Across Multiple Sentiment Analysis Models – Analyzing the Performance of Logistic Regression, Naïve Bayes, SVM, Random Forest, Gradient Boosting, CNN, RNN, and Bi-LSTM before and after augmentation.

Figure 10: An Illustrative Word Cloud Displaying Positive Urdu Text Sentiments of Sentiment Analysis Study.

iterations through the training data, gradually improving their weights and representations. The validation split helped us monitor the models' performance during training and identify potential overfitting issues. We provide a detailed analysis of each model's performance, including precision, recall, F1-score, and accuracy. The table 1 below, summarizes the comparative analysis of these models.

Table 1: Model Performance (%) Evaluations After Augmentation.

Models	Precision	Recall	F1-Score	Accuracy
Logistic Regression	93	94	93	93
SVM	93	93	93	93
Multinomial NB	91	90	90	90
Random Forest	93	93	93	93
Gradient Boosting	89	89	89	89
RNN	93	93	93	93
CNN	93	93	93	92
BiLSTM	93	93	93	92

The logistic regression and SVM models achieved remarkable consistency in performance across all evaluation metrics. Both models exhibited precision, recall, F1-scores, and accuracy of 93%, making them solid contenders for our sentiment classification task. The balanced performance of these models suggests their robustness in handling a wide range of sentiment patterns. The multinomial naive bayes model performed slightly below logistic regression and SVM in terms of precision, recall, F1-score, and accuracy, with values hovering around 90%. While this model might not be the top choice for the task, its performance remains commendable, especially considering its simplicity and efficiency. The random forest and gradient boosting models exhibited precision, recall, F1-scores, and accuracy at the 93% and 89% levels, respectively. These ensemble learning methods demonstrate good overall performance, with random forest showing particularly strong results. However, the performance gap between these models and logistic regression/SVM should be noted.

Moreover, the RNN based model achieved results consistent with logistic Regression and SVM, boasting precision, recall, F1-scores, and accuracy of 93%. The RNN is known for their ability to capture sequential patterns, making them a suitable choice for text classification tasks. Both CNN and Bi-LSTM models delivered competitive results with precision, recall, and F1-scores at 93%. However, their accuracy slightly dipped to 92%, indicating that they may have struggled with some specific instances in the dataset. Despite this, their performance is noteworthy, showcasing the potential of deep learning approaches for sentiment analysis. In the Fig. below 11, 12, 13, 14, 15, 16, 17, and 18, we have provided a comprehensive visualization of the performance evaluation for each of the machine learning models and deep learning models we employed in our analysis. These visual representations include confusion matrices for the ML models and the DL models. Our objective is to allow for a thorough assessment of each model's performance and facilitate a meaningful comparison between them.

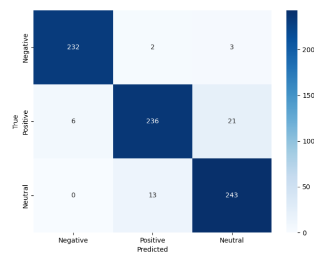

Figure 11: Confusion Matrix of Logistic Regression.

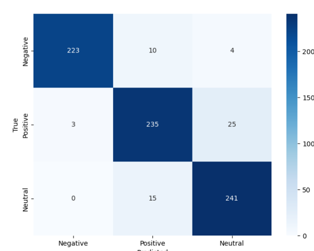

Figure 12: Confusion Matrix of SVM.

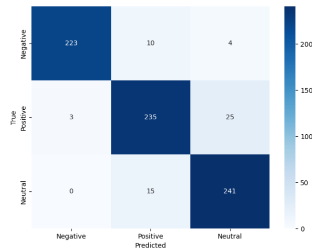

Figure 13: Confusion Matrix of Naive Bayes.

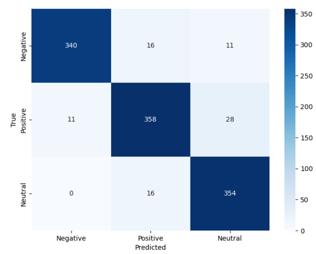

Figure 14: Confusion Matrix of Random Forest.

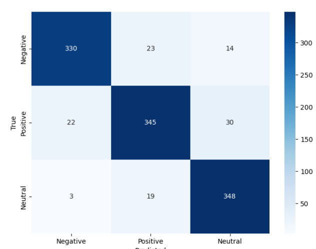

Figure 15: Confusion Matrix of Gradient Boosting.

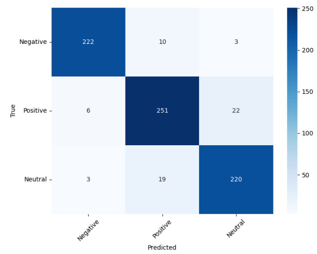

Figure 16: Confusion Matrix of Recurrent neural Network (RNN).

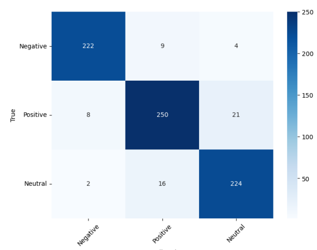

Figure 17: Confusion Matrix of Convolutional Neural Network (CNN).

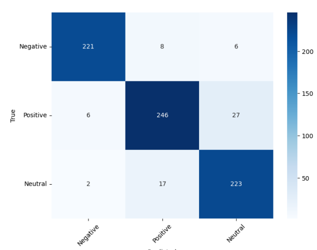

Figure 18: Confusion Matrix of Bidirectional Long Short-Term Memory (Bi-LSTM).

Our comprehensive evaluation of various machine learning and deep learning models revealed several insights. logistic regression and SVM, despite their simplicity, displayed consistent and robust performance, making them reliable choices for sentiment classification tasks. While Multinomial Naive Bayes performed slightly below the top-performing models, its efficiency and interpretability make it a viable option for scenarios where computational resources are limited. Ensemble methods like random forest and gradient boosting offered competitive results, with random forest standing out as an effective choice among ensemble techniques. Deep learning models, including recurrent neural network, convolutional neural network, and bidirectional long short term memory, demonstrated strong performance, especially in terms of precision, recall, and F1-scores. These models leverage the inherent sequential and contextual information in text data, making them suitable for sentiment analysis tasks. However, slight variations in accuracy suggest potential areas for further improvement.

6. Implementation of Explainable AI

The logistic regression is chosen for XAI due to its simplicity and interpretability, making it an ideal choice for providing insights into classification decisions and serving as a baseline model. We apply local interpretable model-agnostic explanations (LIME) XAI technique, which includes randomly choosing samples from the logistic regression model, to improve comprehension for non-native speakers. and to improve our understanding of both sentiment classification and named entity recognition. We utilize LIME XAI to accurately evaluate the contribution of specific words to sentiment polarity classification and the extraction of model names from text. LIME XAI also validates these insights by comparing word counts with probability weights.

6.1. LIME XAI Method for Sentiment Analysis

This method clarifies the rationale for the categorization of particular sentiments. We want to ensure that even after translation, sentiment representations are simple to understand. Take a look at Fig. 19 as an example. Our model classified this sentence as positive sentiment. The original sentence states, “لیپ ٹاپ کوالٹی اور ادائیگی میں acer بہترین ہے ” (Acer laptops are the best in quality and price). As illustrated below, you can observe that the word “بہترین” (best) is the positive word in Urdu, which means ’best’. Therefore, our model correctly categorized it as positive.

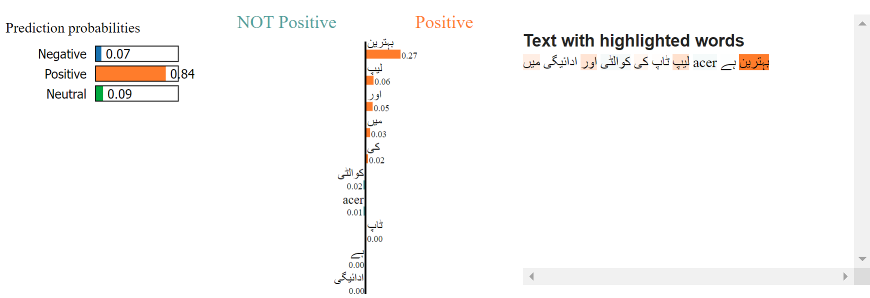

Figure 19: Illuminating Positive Sentiment – Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for Enhanced Understanding of Sentiment Using Logistic Regression Model.

Furthermore, Fig. 20 provides a case study illustration that exemplifies a negative sentiment. Within the sentence, “لیپ ٹاپ بالکل بھی پائیدار نہیں ہیں میرا چند مہینوں میں ٹوٹ گی asus” (ASUS laptops are not durable at all, mine broke down within months) the sentiment is conveyed negatively. The terms “نہیں” (Not) and “ٹوٹ” (Broken Down) act as indicators of this negative sentiment. The LIME XAI method serves as a compelling illustration of the intricate nature of sentiment analysis within the Urdu language, emphasizing the significance of contextual comprehension in accurately assessing emotions. This complexity is particularly pronounced in languages as rich and nuanced as Urdu, where the ability to discern subtle nuances and understand the impact of individual words within their context becomes paramount.

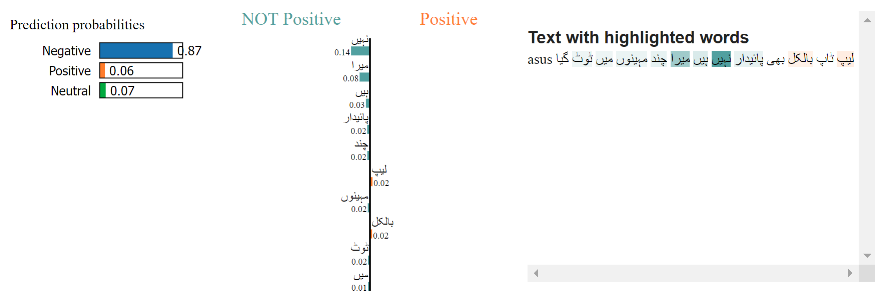

Figure 20: Illuminating Negative Sentiment – Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for Enhanced Understanding of Sentiment Using Logistic Regression Model.

Finally, in the Fig. 21 as we can see the sentiment “HP لیپ ٹاپ کی پرفارمنس کیسی ہوتی ہے” (How does HP laptop perform?) is categorized as neutral. In this example, the term “کیسی” (How) have the highest probability to be a Neutral. This segmentation is important because it explains how potential context changes may impact sentiment analysis results. The LIME method provides us with a more full grasp of how words and context interact, providing us with a more comprehensive view of categorization.

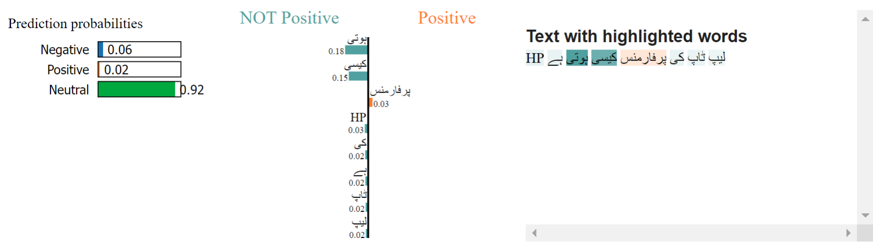

Figure 21: Illuminating Neutral Sentiment – Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for Enhanced Understanding of Sentiment Using Logistic Regression Model.

6.2. LIME XAI Method for NER

We implement the Local Interpretable Model-agnostic Explanations (LIME) technique to enhance understanding of brand name identification within comments. This approach involves randomly selecting samples from the Logistic Regression model to improve clarity. This process aids in identifying specific brands mentioned in the provided comments. The illustration is provided below in the figures. In Fig 22, we can observe the sentiment expressed as, “کا لوک کیسا ہوتا ہے VivoBook S15” (What does the Asus VivoBook S15 look like?). This sentiment specifically pertains to the ASUS laptop model known as the Asus VivoBook S15.

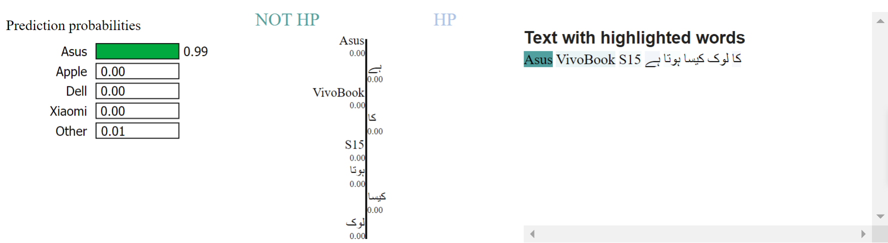

Figure 22: Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for NER Using Logistic Regression Model.

Similarly, in and Fig 23 “ایپل میک بوکس میں تخلیقی سافٹ ویئر کا ایک جامع سویٹ ہے فوٹو ایڈیٹنگ سے لے کر میوزک پروڈکشن تک وہ تخلیق کاروں کے لئے طاقتور ٹولز پیش کرتے ہیں” (Apple MacBooks have a comprehensive suite of creative software, from photo editing to music production, they offer powerful tools for creators.) and in Fig 24 “Apple لیپ ٹاپ کے سپیکر کی کوالٹی کیسی ہے” (How is the speaker quality of Apple laptops) are related to the laptop brand Apple. In the above examples, it’s notable that the term “ایپل میک” and “Apple” carries the highest probability of denoting the Apple laptop brand. This identification holds significant importance as it elucidates how names are distinguished within the comment.

It allows us to pinpoint and recognize the specific brand, such as ASUS, Apple, Dell, etc. from the context provided in the comment. This capability to discern and attribute specific brand mentions in comments or text data is particularly valuable for various applications. For instance, in the context of NER or market research, being able to accurately identify and extract brand names can provide insights into consumer opinions, preferences, and trends.

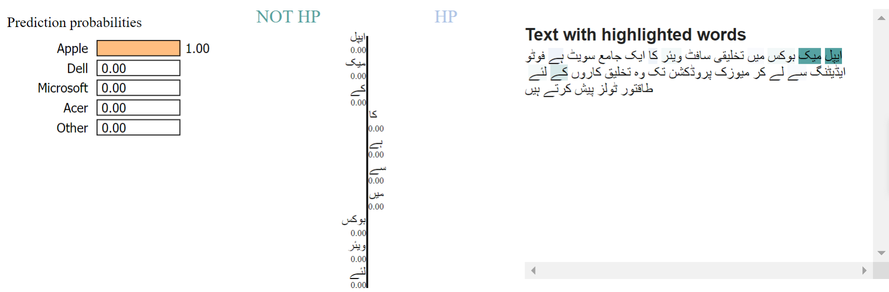

Figure 23: Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for NER Using Logistic Regression Model.

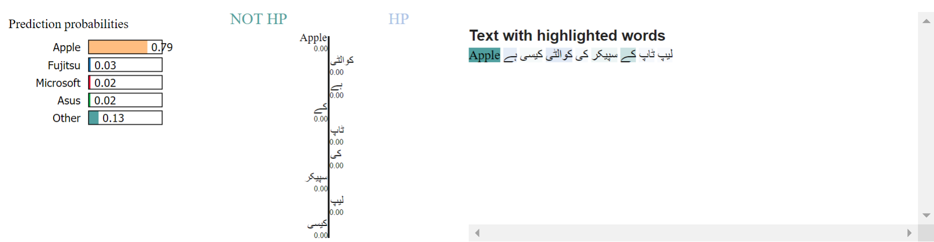

Figure 24: Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for NER Using Logistic Regression Model.

6.3. LIME XAI Method for Gender Prediction

Predicting gender is a key component of our project. Our labeled dataset contains usernames and their associated gender labels. To achieve this prediction, we apply a logistic regression model using our training dataset. In order to gain deeper insights into how our gender prediction model functions, we implement the LIME (Local Interpretable Model-agnostic Explanations) method. LIME is an explainable artificial intelligence (XAI) technique that randomly selects instances from our logistic regression model to help us understand its decision-making process. Using LIME, we randomly sampled examples from the logistic regression model that we constructed and applied this sampled data to predict gender based on usernames. This method provided us with valuable insights into the underlying mechanisms and features influencing the model's gender predictions. To visually illustrate this process, we have included Fig. 27, 25, and 26. below that depict the LIME-based analysis of our gender prediction model, highlighting the interpretability and transparency of our approach.

Figure 25: Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for Gender Prediction Using Logistic Regression Model.

Figure 26: Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for Gender Prediction Using Logistic Regression Model.

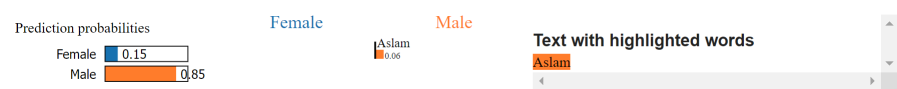

Figure 27: Leveraging Local Interpretable Model-agnostic Explanations (LIME) XAI Technique for Gender Prediction Using Logistic Regression Model.

7. Conclusion and Future Work

In this paper, we inquired about the application of NLP techniques in market sentiment analysis in the Urdu language. Thus, figured out the effectiveness and accuracy of NLP techniques. Then our model identified the most demanded product in the market based on gender by using social media data, such as consumer sentiment and their preferences in the product features, which widened our vision to the market sentiment of the Urdu-speaking region. This information can be used by businesses to make informed decisions and enhance their business in today's competitive market. For future work, we will expand our language coverage beyond Urdu, incorporate specific laptop model names into our demand forecasting, and further investigate factors such as price, battery life, condition, and performance to enhance the accuracy of our demand predictions.

Availability of supporting data

The data will be available on request.
drive link

References

[1] E. Britannica, Urdu language — history, script, & words, https://www.britannica.com/topic/Urdu-language (2023).

[2] G. Cardona, Indo-iranian languages — history, characteristics & classification (2023).
URL `https://www.britannica.com/topic/Indo-Iranian-languages`

[3] Lingua Education, The most spoken languages in the world, https://lingua.edu/the-most-spoken-languages-in-the-world/ (2023).

[4] D. Khurana, A. Koli, k. Khatter, S. Singh, Natural language processing: State of the art, current trends and challenges, Multimedia Tools and Applications 82 (2022) 3713–3744. `doi:10.1007/s11042-022-13428-4`.
URL `https://doi.org/10.1007/s11042-022-13428-4`

[5] U. Khan, M. Ahmad, F. Shafiq, M. Sarim, Urdu natural language processing issues and challenges: A review study, in: Proceedings of the International Conference on Future Computational Technologies and Applications (ICFCTA), 2019, p. 461–470. `doi:10.1007/978-981-15-5232-8\_39`.

[6] I. Farida, D. Setiawan, Business strategies and competitive advantage: The role of performance and innovation, Journal of Open Innovation: Technology, Market, and Complexity 8 (3) (2022) 163. `doi:10.3390/joitmc8030163`.
URL `https://doi.org/10.3390/joitmc8030163`

[7] K. Chaudhary, M. Alam, M. S. Al-Rakhami, A. Gumaei, Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics, Journal of Big Data 8 (1) (May 2021). doi:10.1186/s40537-021-00466-2.

[8] P. Jiao, A. Veiga, A. Walther, Social media, news media and the stock market, Journal of Economic Behavior & Organization 176 (2020) 63–90. doi:10.1016/j.jebo.2020.03.002.

[9] V. Sathya, A. Venkataramanan, A. Tiwari, D. D. P.S., Ascertaining public opinion through sentiment analysis, in: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1139–1143. doi:10.1109/ICCMC.2019.8819738.

[10] M. A. Shafin, M. M. Hasan, M. R. Alam, M. A. Mithu, A. U. Nur, M. O. Faruk, Product review sentiment analysis by using nlp and machine learning in bangla language, in: 23rd International Conference on Computer and Information Technology (ICCIT), 2020, pp. 1–5. doi:10.1109/ICCIT51783.2020.9392733.

[11] IvyPanda, Natural language processing in business - 1419 words — research paper example (2022).
URL https://ivypanda.com/essays/natural-language-processing-in-business/

[12] R. Boorugu, G. Ramesh, A survey on nlp-based text summarization for summarizing product reviews, IEEE Xplore (Jul. 2020). doi:10.1109/ICIRCA48905.2020.9183355.
URL https://ieeexplore.ieee.org/document/9183355

[13] Y. Guo, Stock price prediction based on lstm neural network: The effectiveness of news sentiment analysis, 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) (2020). doi:10.1109/icemme51517.2020.00206.

[14] U. Naqvi, A. Majid, S. A. Abbas, Utsa: Urdu text sentiment analysis using deep learning methods, IEEE Access 9 (2021) 114085–114094. doi:10.1109/access.

[15] M. A. Nadeem, K. Irfan, K. Atiq, M. O. Beg, M. U. Arshad, Sequence-driven neural network models for ner tagging in roman urdu, 2022 International

Conference on Frontiers of Information Technology (FIT) (2022). doi:10.1109/fit57066.2022.00040.

[16] I. Rehman, T. R. Soomro, Urdu sentiment analysis, Applied Computer Systems 27 (1) (2022) 30–42. doi:10.2478/acss-2022-0004.

[17] V. Sathya, A. Venkataramanan, A. Tiwari, D. D. P.S., Ascertaining public opinion through sentiment analysis, in: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1139–1143. doi:10.1109/ICCMC.2019.8819738.

[18] J. Sen, S. Mehtab, A robust predictive model for stock price prediction using deep learning and natural language processing, TechRxiv (2021). doi:10.36227/techrxiv.15023361.v1.

[19] V. Sathya, A. Venkataramanan, A. Tiwari, D. D. P.S., Ascertaining public opinion through sentiment analysis, in: 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1139–1143. doi:10.1109/ICCMC.2019.8819738.

[20] M. Hossain, N. Nayla, A. Rassel, Product market demand analysis using nlp in banglish text with sentiment analysis and named entity recognition, 56th Annual Conference on Information Sciences and Systems (CISS) (2022). doi:10.1109/ciss53076.2022.9751188.

[21] M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah, G. Sang Choi, Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model, Computational Intelligence 37 (2021) 409–434. doi:10.1111/coin.12415.

[22] B. Chandio, et al., Sentiment analysis of roman urdu on e-commerce reviews using machine learning, Computer Modeling in Engineering & Sciences 131 (3) (2022) 1263–1287. doi:10.32604/cmes.2022.019535.

[23] W. Ahmad, M. Edalati, Urdu speech and text based sentiment analyzer (2022). doi:10.48550/arXiv.2207.09163.

[24] R. Tatman, Urdu stopwords list (2016).
URL https://www.kaggle.com/rtatman/urdu-stopwords-list

[25] C. D. Paice, Stemming (2016). doi:10.1007/978-1-4899-7993-3_942-2.
URL https://doi.org/10.1007/978-1-4899-7993-3_942-2

[26] H. Jiang, Y. Hua, D. Beeferman, D. Roy, Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis, arXiv preprint arXiv:2201.07281 (2022).