Market Demand Analysis Using NLP in Urdu Language.

***Abstract*—Understanding market demand is crucial for businesses to devise effective strategies in today's competitive landscape. However, analyzing market demand 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 market demand analysis in Urdu. 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. Various NLP models and algorithms were then employed to extract insights, capture consumer sentiment, and identify specific entities such as brand names and gender from the username. The research presents a comprehensive annotated dataset for analyzing market demand in Urdu-speaking regions. The effectiveness of the approach was evaluated through various algorithms. Among machine learning models, SVM, multinomial logistic regression, and random forest performed consistently well with 93% accuracy, while RNN excelled among deep learning models with 94% accuracy. Finally, this research also used the LIME XAI method to improve understanding of sentiment classification.**

***Index Terms*—Market Demand Analysis, Urdu Language, Sentiment analysis, Natural Language Processing, LIME XAI, Name Entity Recognition,Gender Prediction.**

I. INTRODUCTION

Million of the people in Pakistan, India, and other countries speak Urdu as language. It is a mix of Indo-European, Indo-Iranian, and Indo-Aryan languages, with influences from Persian and Arabic. 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. NLP can now perform sentiment analysis, parts of speech tagging, unnecessary word deletion, parsing, name entity recognition (NER), and many more. NLP has made a lot of progress in English, but there is still more work to be done in Urdu. 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. Social media is becoming a significant industry, allowing businesses to engage with customers and track their activity.

To achieve this, we utilize automated data scraping techniques to gathered raw text data in Urdu from various social media buy and sell groups. After filtering and organizing this data with natural language processing, we created a well-structured dataset. We used machine learning algorithms, particularly named entity recognition (NER) and gender prediction, to train the dataset for sentiment analysis. The accuracy of our models was validated using a comprehensive test dataset, providing actionable insights regarding the most sought-after products based on gender and sentiment analysis.

This paper intends to thoroughly investigate the methods and tools used to understand market demand in Urdu using natural language processing (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.

II. 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 demand for certain items, researchers have looked at a variety of techniques for collecting, processing, and analyzing online social media data [5]. 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 [18]. 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 [7]. In another publication, [8], the researchers 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.

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 (MSE) values, showcased through its successful application in analyzing time-series data encompassing stock prices and returns, as mentioned in [13]. 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 [14], 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 Standalone Fuzzy Neural Network (SOFNN) algorithm, expounded in [15], standing out due to its exceptional accuracy in performing sentiment analysis tasks.

Moreover, 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 [16]. Within the realm of related research, the work conducted by [1] 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 research paper authored by [3] introduces an innovative unified model that synergistically merges convolutional neural networks (CNN) and LSTM networks. 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. This application extends beyond sentiment analysis, facilitating a deeper comprehension of prevailing public opinions and sentiment trends across diverse social media platforms, essentially opening avenues for a more insightful understanding of contemporary digital discourse. In summation, the various research works cited in this compilation collectively highlight the dynamic interplay between cutting-edge technological methodologies, sentiment analysis, and the profound impacts these factors collectively exert on text summarization, financial forecasting, sentiment quantification, and public sentiment analysis across digital domains.

Market demand analysis plays a crucial role in understanding consumer behavior and preferences, enabling businesses to make informed decisions about product development, marketing strategies, and resource allocation. While several studies have explored market demand analysis in the English language, it is important to understand the unique consumer dynamics of the Urdu-speaking market. The study focuses on addressing the limitations of existing natural language processing (NLP) solutions for sentiment analysis in Roman Urdu, an informal language. The research investigates various machine learning algorithms, including a support vector machine (SVM) enhanced with a Roman Urdu steamer, to analyze sentiments and proposes a new dataset for Roman Urdu sentiment analysis [17]. The literature on market demand analysis has predominantly focused on English-speaking markets. These studies have employed different methodologies, techniques, and models to examine consumer behavior and demand patterns. Studying market demand analysis specifically in Urdu holds significant importance due to the large population of Urdu speakers worldwide. With Urdu being the national language of Pakistan and widely spoken in India, this language presents vast market opportunities [6]. Understanding the unique consumer preferences and a cultural nuance of the Urdu-speaking market is crucial for businesses to tailor their products and marketing strategies effectively. Conducting market analysis in the Urdu language presents certain challenges. One major challenge is the availability of reliable, up-to-date data, and linguistic barriers may hinder comprehensive data collection. Additionally, cultural differences and regional disparities within the Urdu-speaking market can significantly impact consumer behaviors and demand patterns. However, addressing these challenges presents unique opportunities for researchers and businesses to gain a competitive advantage in these 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.

III. 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 Figure 1, the primary components of our system are sentiment analysis, named entity recognition, and gender prediction.

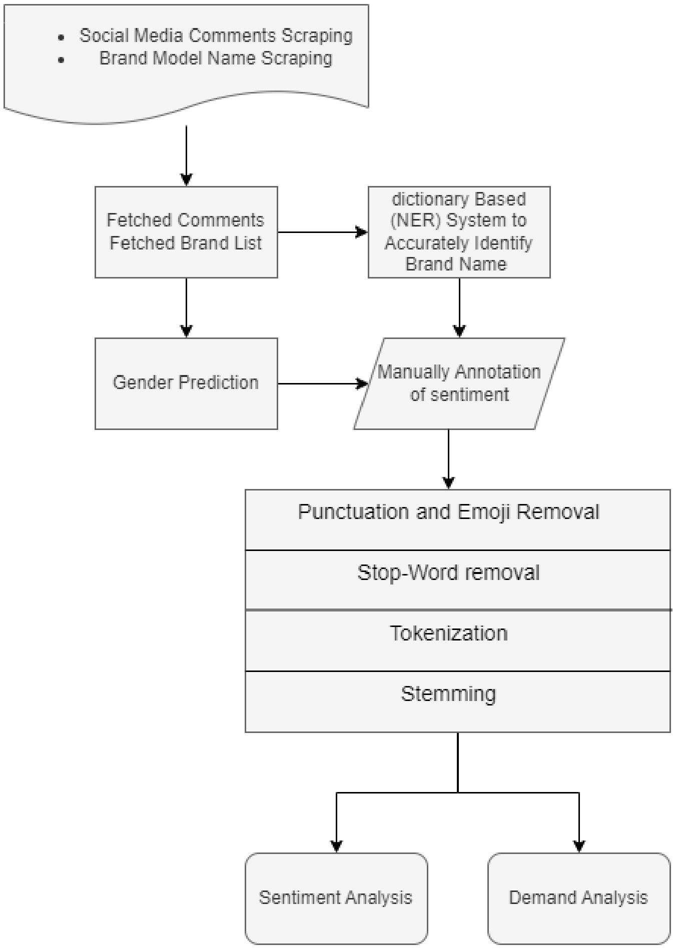

Social Media Comments Scraping
Brand Model Name Scraping
Fetched Comments
Fetched Brand List
dictionary Based (NER) System to Accurately Identify Brand Name
Gender Prediction
Manually Annotation of sentiment
Punctuation and Emoji Removal
Stop-Word removal
Tokenization
Stemming
Sentiment Analysis
Demand Analysis

Figure 1. Flowchart of The Proposed Model.

To conduct market demand analysis in Urdu, we faced a lack of existing research and organized 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 named entity recognition (NER), we devised a function that utilized a dictionary-based system to identify device names through comparison with the laptop list and the comments' data.

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 user genders, 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 couldn't 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 and test sets. 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.

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.

IV. DATASET

We built two datasets for training our model: one based on consumer product reviews from different social media platforms, and the other contained laptop brand name data from Wikipedia. We discovered a shortage of current research and organized data concerning client product reviews throughout our investigation on market demand 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 dropped all those rows that do not have usernames, brand names, or noises. Finally, with the help of native speakers, we labeled the dataset with 3700 entries comprising information on sentiment, gender, and product entities after extensive preprocessing.

A. Data collection

In the modern era, social networking sites have become instrumental in shaping the producer-consumer relationship. Customers may express their preferences and reviews on a variety of Facebook and YouTube pages and channels related to laptops. We made the decision to collect our raw data from social media platforms since we were aware of the importance of these platforms for comprehending product demand. To do this, 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.

B. 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.

C. 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 [4]. 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.

D. Tokenization

Tokenization is the process of breaking down text into individual tokens [9]. It is a crucial stage in natural language processing. In our study on market demand 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.

E. 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 [10]. 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.

V. Named Entity Recognition

Named Entity Recognition (NER) using a dictionary-based method is a useful tool for identifying certain items in text [11]. We used NER to extract brand-related information from written comments in our study, which is an essential task in natural language processing (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. 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 expressions that utilized this dictionary-based approach to match device names by comparing the laptop list with comments' content.

VI. Gender Prediction

Understanding gender variances in product demand is crucial, since gender preferences can vary greatly. 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 utilized 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. Finally, the gender prediction component of our approach was effectively produced after thorough evaluation and improvement.

VII. Sentiment Analysis

We used a various of machine learning and deep learning models for sentiment analysis. Compared to other machine learning models, support vector machine (SVM), multinomial logistic regression, and random forest performed better, with an accuracy rate of 93%, while gradient boosting had an accuracy of 89% and multinomial NB Accuracy was 90%. On the other hand, RNN performed exceptionally well in deep learning models with high accuracies of 94%. The CNN and LSTM model had an accuracy rate of 92%. In conclusion, machine learning models such as SVM, Multinomial Logistic Regression, and Random Forest were consistent and performed well. Deep learning models (RNN, CNN, and LSTM), on the other hand, were good at interpreting complicated patterns, resulting in high accuracy. These findings have critical implications for model selection and development in the domain of sentiment analysis.

Table I
Model Performance

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 Classifier	93	93	93	93
Gradient Boosting Classicier	89	89	89	89
RNN	94	94	94	94
CNN	92	92	92	92
LSTM	93	93	93	92

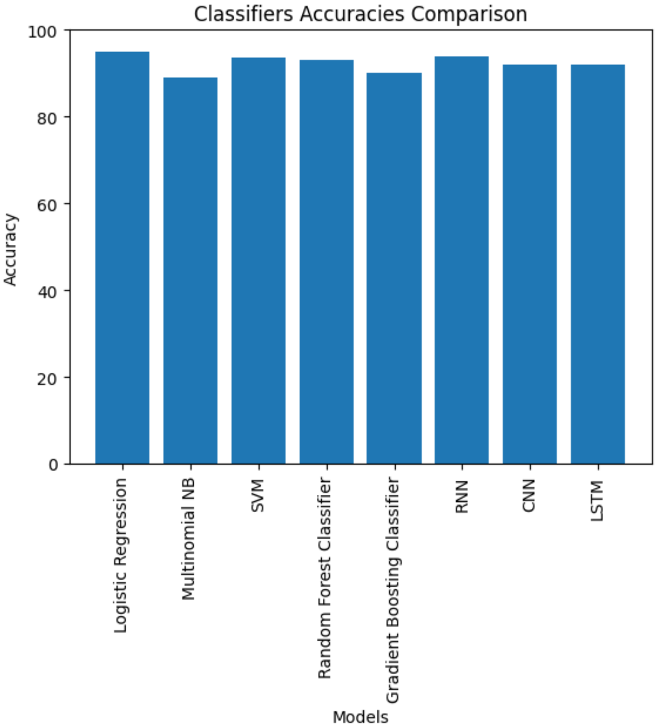


Figure 2. Accuracy comparison graph of Models.

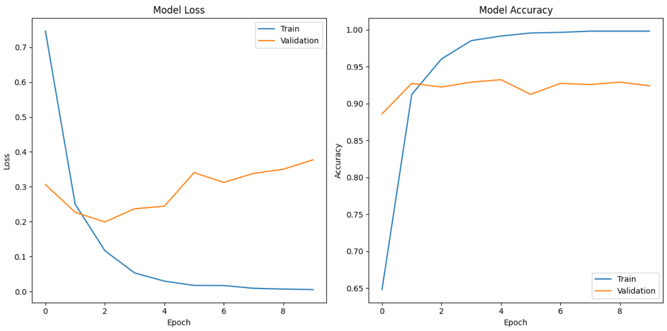
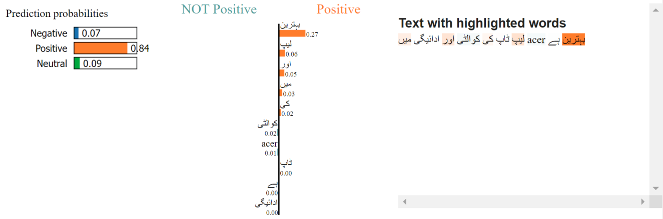


Figure 3. RNN Loss and Accuracy.

VIII. Applying LIME XAI Method

The accuracy of the SVM, Multinomial Logistic, and Random Forest Classifiers was higher in earlier studies. We applied the LIME XAI technique, which includes randomly choosing samples from the Logistic Regression model, to improve comprehension for non-native speakers. This clarifies the rationale for the categorization of particular sentiments. We wanted to ensure that even after translation, sentiment representations were simple to understand. Take a look at Figure 4 for example. This sentence was initially classified as having a positive sentiment. The original sentence states, ” acer لیپ ٹاپ کوالٹی اور ادائیگی میں بہترین ہے ” (Acer laptops are the best in quality and price). In the illustration below, you can see that the word ”بہترین” (best) was categorized as positive.

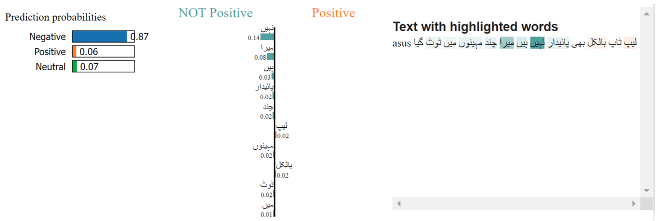


Text with highlighted words

بہترین ہے acer لیپ ٹاپ کی کوالٹی اور ادائیگی میں

Figure 4. LIME XAI result for positive Sentiment.

Furthermore, Figure 5 provides a case study illustration that exemplifies a negative sentiment. Within the sentence ”susa لیپ ٹاپ بالکل بھی پائیدار نہیں ھیں میرا چند مہینوں میں ٹوٹ گیا” (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. This case study vividly demonstrates the intricate nature of sentiment analysis and underscores the significance of contextual comprehension.

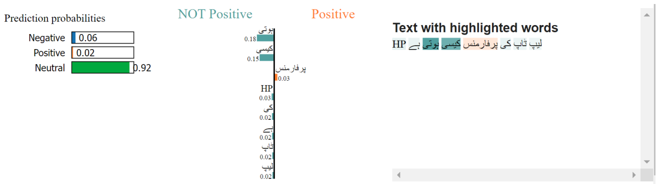


Text with highlighted words

لیپ ٹاپ بالکل بھی پائیدار نہیں ہیں میرا چند مہینوں میں ٹوٹ گیا asus

Figure 5. LIME XAI result for Negative Sentiment.

Finally, in the Figure 6 as we can see the sentiment ”PH 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.



Text with highlighted words

لیپ ٹاپ کی پرفارمنس کیسی ہوتی ہے HP

Figure 6. LIME XAI result for Neutral Sentiment.

IX. Demand Analysis

After conducting extensive research, we eagerly awaited the outcomes of a substantial workload, Our team devised a visually appealing pie chart to present the usage patterns of social media among different genders, as well as their inclination towards gadgets. It is evident from the data that approximately 62.3% of Men and 37.7% of women exhibit an interest in this particular sector. Moreover, we created a bar chart that visualizes the percentage of men and women interested in the top 10 laptops in the Urdu-speaking region market. Surprisingly, the data highlights that Lenovo is the most preferred laptop brand among both genders. Below, you can find a detailed bar chart presenting our predictions for the three most preferred laptops among males and females.

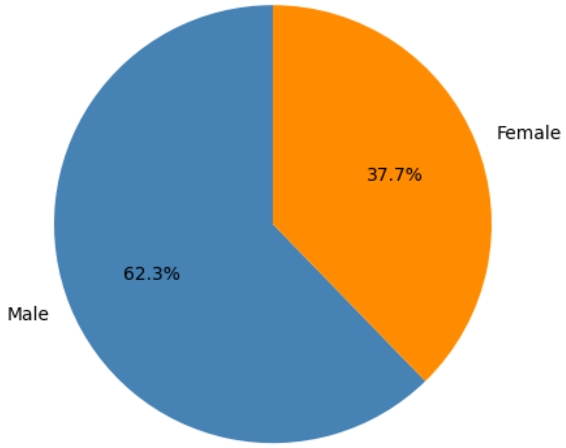


Figure 7. Male-Female Ratio.

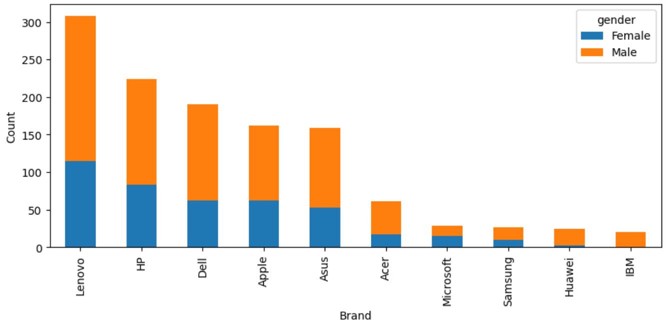


Figure 8. Top Ten Laptop Brands.

X. Conclusion

In this paper, we inquired about the application of NLP techniques in market demand 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 demand 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.

XI. Future Work

Currently, our focus is only on the Urdu language; however, our trajectory will extend to encompass multitude of languages. Although our paper focuses on demand forecasting for different laptop brand names, we have plans to expand this analysis to include specific model names of different laptop brands. Additionally, our future efforts include exploring other variables such as a desired price, battery timing, condition, and performance within the demand analysis domain, with the aim of improving the accuracy of our model's demand predictions.

References

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

[2] H. Chen, P. De, Y. Hu, and B.-H. Hwang, "Sentiment revealed in social media and its effect on the stock market," pp. 25–28, 2011. doi: https://doi.org/10.1109/SSP.2011.5967675.

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

[4] R. Tatman, "Urdu Stopwords List," Kaggle, 2016.

[5] K. Chaudhary, M. Alam, M. S. Al-Rakhami, and A. Gumaei, "Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics," Journal of Big Data, vol. 8, no. 1, May 2021, doi:10.1186/s40537-021-00466-2.

[6] Ahmad, W. (2022). "Urdu speech and text analyzer". https://arxiv.org/pdf/2207.09163v1.pdf

[7] V. Sathya, A. Venkataramanan, A. Tiwari, and D. D. P.S., “Ascertaining public opinion through sentiment analysis,” pp. 1139–1143, 2019. doi: 10 . 1109/ICCMC.2019.8819738

[8] M. A. Shafin, M. M. Hasan, M. R. Alam, M. A. Mithu, A. U. Nur and M. O. Faruk, ”Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language,” 2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020, pp. 1-5, doi: 10.1109/ICCIT51783.2020.9392733.

[9] “Speech and Language Processing,” Stanford.edu, 2018. https://web.stanford.edu/ jurafsky/slp3/

[10] C. D. Paice, “Stemming,” Encyclopedia of Database Systems, pp. 1–5, 2016, doi:10.1007/978-1-4899-7993-3_942-2.

[11] H. Jiang, Y. Hua, D. Beeferman, and D. Roy, “Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis,” Jan. 2022, doi:10.48550/arXiv.2201.07281

[12] R. Boorugu and G. Ramesh, “A Survey on NLP based Text Summarization for Summarizing Product Reviews,” IEEE Xplore, Jul. 01, 2020. https://ieeexplore.ieee.org/document/9183355.

[13] Guo, Y. (2020) ‘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). doi:10.1109/icemme51517.2020.00206.

[14] V. Sathya, A. Venkataramanan, A. Tiwari and D. D. P.S., ”Ascertaining Public Opinion Through Sentiment Analysis,” 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1139-1143, doi: 10.1109/ICCMC.2019.8819738.

[15] Sen, J. and Mehtab, S. (2021) A robust predictive model for stock price prediction using Deep Learning and Natural Language Processing. doi:10.36227/techrxiv.15023361.v1.

[16] V. Sathya, A. Venkataramanan, A. Tiwari and D. D. P.S., ”Ascertaining Public Opinion Through Sentiment Analysis,” 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1139-1143, doi: 10.1109/ICCMC.2019.8819738

[17] B. Chandio et al., “Sentiment Analysis of Roman Urdu on E-Commerce Reviews Using Machine Learning,” Computer Modeling in Engineering & Sciences, vol. 131, no. 3, pp. 1263–1287, 2022, doi: 10.32604/cmes.2022.019535

[18] P. Jiao, A. Veiga, and A. Walther, “Social media, news media and the stock market,” Journal of Economic Behavior & Organization, vol. 176, pp. 63–90, Aug. 2020, doi: 10.1016/j.jebo.2020.03.002.