Market Demand Analysis Using NLP in Urdu Language.

Malika Muradi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
malika.muradi@g.bracu.ac.bd

Basit Hussain
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
basit.hussain@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

Sania Azhmee Bhuiyan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sania.azhmee.bhuiyan@g.bracu.ac.bd

***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. This research paper aims to bridge this gap by investigating the application of NLP techniques, including sentiment analysis and named entity recognition, 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, capturing consumer sentiment and identifying specific entities such as brand names, product features, and gender from the user name. This research also introduced the largest annotated dataset for Urdu market demand analysis, providing a rich collection of consumer reviews, expert annotations, and market trends. The findings shed light on the unique characteristics of market demand in Urdu-speaking regions, enabling businesses to make informed decisions, tailor strategies, and enhance their competitive advantage. The effectiveness of the approach was evaluated through various machine learning and deep learning models. SVM, Multinomial Logistic Regression, and Random Forest performed consistently well with 93% accuracy in machine learning algorithms, 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, Sentiment analysis, Natural Language Processing, Name Entity Recognition,Gender Prediction, Urdu language, LIME XAI.**

I. INTRODUCTION

Urdu is a language spoken by millions of people in Pakistan, India, and other countries. 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 (POS) 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 most popular 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 gathered raw text data in Urdu and from social media buy and sell groups using automated data scraping techniques. 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 on 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, accurately forecasting 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 use of sentiment analysis and natural language processing techniques to analyze social media data has gained significant attention in recent years. Researchers have explored various approaches to collect, preprocess, and analyze social media data to gain insights into consumer behavior and predict market demand for different products [5]. The impact of emerging social platforms on financial markets is significant. Research has shown that sentiment expressed through these social platforms has a more considerable and enduring influence on stock returns compared to views expressed in traditional news sources [2]. The internet has enabled the collection and analysis of vast volumes of text data, which may be used to better comprehend people's ideas and feelings. This data can be utilized for a number of purposes, including market research, customer satisfaction surveys, and political polls [7]. The researcher of [8] mentioned that the internet has made it possible to collect and analyze large amounts of text data, which can be used to understand people's opinions and emotions. This information can be used for a variety of purposes, such as market research, customer satisfaction surveys, and political polling.

The study closely related to ours is conducted by [1], where the researchers develop a model to identify popular device entities in Bangla-speaking regions by analyzing consumer preferences. The study employs sentiment analysis, gender prediction, and named entity recognition to extract valuable insights from consumer data. Through the successful identification of highly sought-after device entities, the study facilitates informed business decisions in Bangla-speaking regions, providing significant benefits to companies operating in the market. Additionally, another paper [3] proposes a novel unified model that combines CNN and LSTM networks for sentiment analysis of tweets. The unified model is skillfully designed to capture contextual and sequential information in tweets, resulting in more accurate sentiment analysis. This model's applications extend to understanding public opinion and sentiment trends on social media.

Market demand analysis plays a crucial role in understanding consumer behavior and preferences, enabling businesses to make informed decision about product development, marketing strategies,and resource allocation. while several studies have explored market demand analysis in English language, it is important to understand the unique consumer dynamics of Urdu-speaking market.

Previous literature on market demand analysis had predominantly focused on English speaking markets. These studies have employed different methodologies, techniques, and model to examine consumer behavior and demand pattern.

Studying market demand analysis specifically in Urdu Language holds significant importance due to large population of Urdu speakers worldwide.with Urdu being national language of Pakistan and widely spoken in India, this language present vast market opportunities. [6] Understanding the unique consumer preferences and cultural nuance in Urdu speaking market is crucial for the businesses to tailor their products and marketing strategies effectively.

Conducting market analysis in Urdu language present certain challenges. one major challenges is the availability of reliable up-to-date data and linguistic barriers may hinder comprehensive data collection. Additionally, cultural differences, and regional disparities within Urdu speaking market can significantly impact consumer behaviors and demand patterns. However, addressing these challenges present unique opportunities for researcher and businesses to gain a competitive advantage in these untapped markets.

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. The main components of our method include sentiment analysis gender prediction and Named Entity Recognition as depicted in Figure 1.

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 37000 raw entries.

In order to identify laptop brand names within the comments, we supplemented our data with laptop brands and models name 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 Google API to translate the device model names. For Named Entity Recognition (NER) we designed a function that employed a dictionary-based system to match the device names by comparing the laptop list with 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 user names and comments that did not mention any

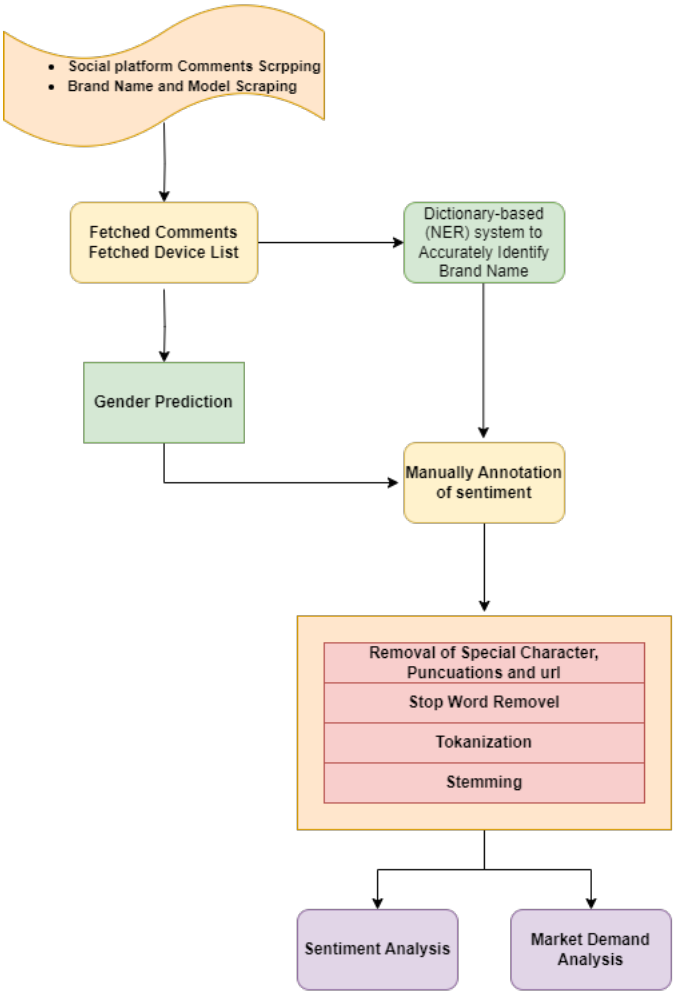

Social platform Comments Scrpping
Brand Name and Model Scraping
Fetched Comments
Fetched Device List
Dictionary-based (NER) system to Accurately Identify Brand Name
Gender Prediction
Manually Annotation of sentiment
Removal of Special Character, Puncuations and url
Stop Word Removel
Tokanization
Stemming
Sentiment Analysis
Market Demand Analysis

Figure 1. work model flowchart.

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 various machine learning and deep learning algorithms. Following this we trained our named entity classifier using Amazon Comprehend for custom named entity recognition and the results were satisfactory. Finally, based on the genders in the current market, we plotted a list of the most demanding devices.

By following this meticulous methodology, we successfully constructed robust datasets that effectively captured and analyzed market demand in Urdu languages. These datasets provided valuable insights into consumer preferences and behaviors enabling us to gain a deeper understanding of the market landscape.

IV. DATASET

For training our model, we created two datasets: one based on customer product reviews and the other containing laptop brands name list data from Wikipedia. During our research on market demand analysis in Urdu, we encountered a scarcity of existing research and organized data concerning customer product reviews. To overcome this challenge we utilized social networking platforms particularly laptop-related YouTube channels and Facebook public pages, as valuable sources for collecting consumer preferences and reviews. Through the use of the Instant Data Scraper tool, we collected a significant number of comments from various social networking platforms, resulting in a dataset comprising over 37000 raw entries. After thorough preprocessing, we obtained a labeled dataset containing 3700 entries with information on sentiment, gender, and product entities.

A. *Data collection*

In the modern era, social networking sites have become instrumental in shaping the producer-consumer relationship. Numerous laptop-related YouTube channels and Facebook pages on these platforms serve as hubs where consumers share their preferences and reviews. Recognizing the significance of these platforms for understanding product demand, we opted to gather our raw data from social networking sites. To accomplish this, we employed the Instant Data Scraper tool to collect people's comments pertaining to Laptop from various social networking sites. The unsupervised data was then stored in a CSV file to facilitate further analysis. Additionally, for our second dataset, we utilized a Python web scraper to gather laptop model data from Wikipedia.

B. *Removal of Punctuations and emojis*

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. Additionally, we removed all emojis present in Urdu datasets. This step was crucial in maintaining the cleanliness and preparedness of the text data for subsequent processing and analysis.

C. *Stop-word removal*

In order to improve the quality of our text data and facilitate more accurate analyses, we focused on removing stop words from the data set. Stop words refer to commonly used words that do not carry substantial meaning in a particular language and can potentially disrupt sentiment analysis by introducing noise. To accomplish this, we obtained a comprehensive list of Urdu stop words from a

Kaggle dataset [4], which we utilized to eliminate these non-essential words from our dataset. This step allowed us to refine the text data and ensure that the subsequent analyses were based on more meaningful and relevant content.

D. Tokenization

Tokenization, an essential step in natural language processing, was employed to split the text into individual words or tokens in our study on market demand analysis in Urdu languages. This process enabled us to break down the comments into their constituent units for further analysis. By conducting tokenization, each word within a comment was isolated and made available for individual examination and subsequent processing. This approach ensured that the text data was effectively prepared for comprehensive analysis and subsequent language processing tasks.

E. Stemming

Stemming played a vital role in our market demand analysis using NLP in Urdu languages, as it helped normalize the vocabulary extracted from tokenized comments. By eliminating prefixes and suffixes, stemming reduced word variations and standardized them to their base or root forms. To accomplish this, we applied an appropriate stemming algorithm to the tokenized words in the comments, allowing us to bring them to their base forms. This normalization process was instrumental in consolidating the vocabulary and minimizing the noise caused by word variations.

Our data pre-processing involved several key steps, including data collection, punctuation removal, stop-word removal, tokenization, and stemming. Through these steps, we successfully prepared a refined dataset that served as a valuable resource for sentiment analysis using machine learning techniques. Ultimately, this enabled us to gain meaningful insights into the dynamics of market demand.

V. NAMED ENTITY RECOGNITION

Using a dictionary-based approach for Named Entity Recognition (NER) is a practical technique to identify specific elements in text. In our research, we applied NER to extract brand-related information from written comments, an important task in Natural Language Processing (NLP). Our dataset contained laptop names in both English and Urdu. To address this, we developed a multilingual dictionary-based system that covered both languages. Our NER process, built upon this strategy, aimed to identify certain words connected to different 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.

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 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. Finally, The gender prediction component of our approach was effectively produced after thorough evaluation and improvement.

VII. SENTIMENT ANALYSIS

We used a variety 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%. We also tested the BERT model, which showed 67% accuracy. 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. However, BERT's performance was not as good, but we may look into it more. These findings are critical for selecting and refining models for understanding sentiment analysis.

Table I
Model Performance

Models	Precision	Recall	F-1 Score	Accuracy
Multinomial LR	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	94	94	94	94
CNN	92	92	92	92
LSTM	93	93	93	92
BERT	68	64	66	67

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 Random Forest Classifier 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 an example. This sentence was initially classified as having a positive sentiment. The original sentence states, ” لیپ ٹاپ کوالٹی اور ادائیگی میں acer

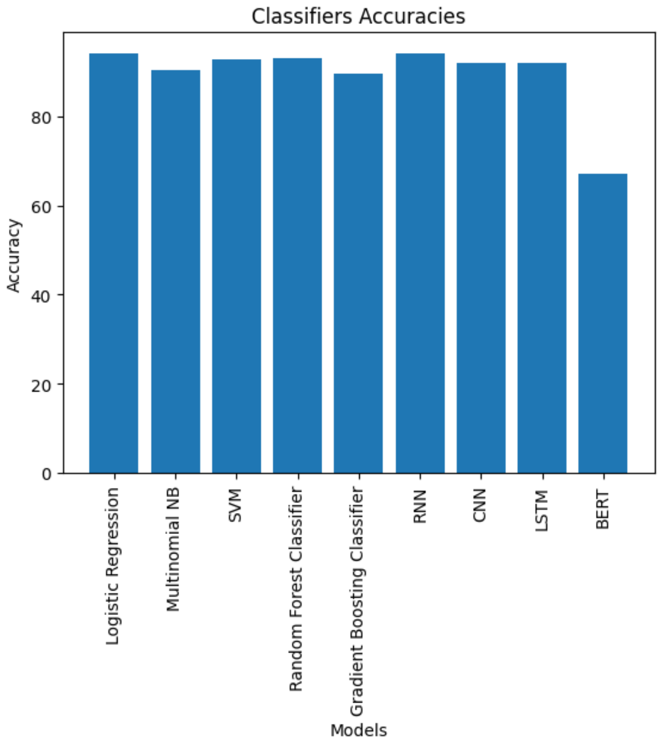


Figure 2. Classifiers Accuracies.

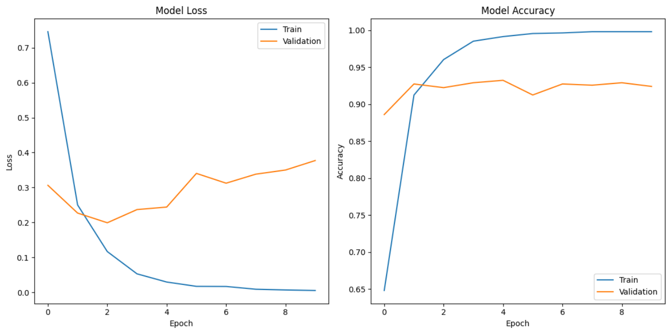
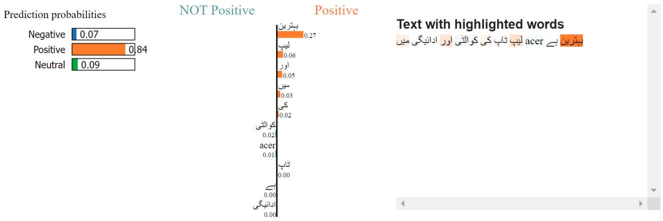


Figure 3. RNN Loss and Accuracy.

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.

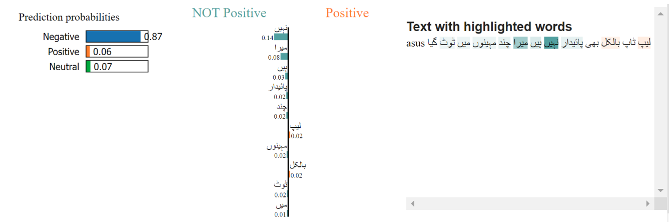


Text with highlighted words

بہترین ہے acer لیپ ٹاپ کی کوالٹی اور ادائیگی میں

Figure 4. LIME result for positive Sentiment.

Furthermore, Figure 5 provides a case study illustration that exemplifies a negative sentiment. Within the sentence ”asus 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.

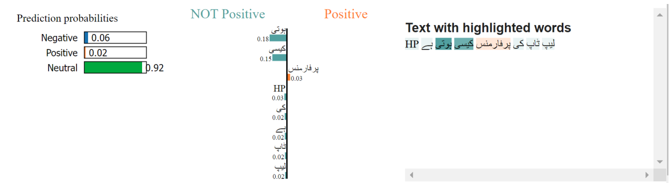


Text with highlighted words

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

Figure 5. LIME result for Negative Sentiment.

Finally, in the Figure 6 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 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 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 shows the percentage of men and women interested in the top 8 laptops in the Urdu-speaking region market. Surprisingly, the data highlights that Lenovo is the most preferred laptop brand for both genders. Below, you can find a detailed bar chart presenting our predictions for the three most preferred laptops among males and females.

X. Conclusion and Future work

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

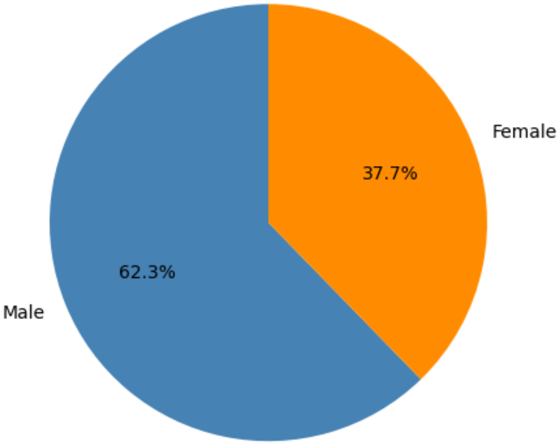


Figure 7. Male-Female Ratio Pie Chart.

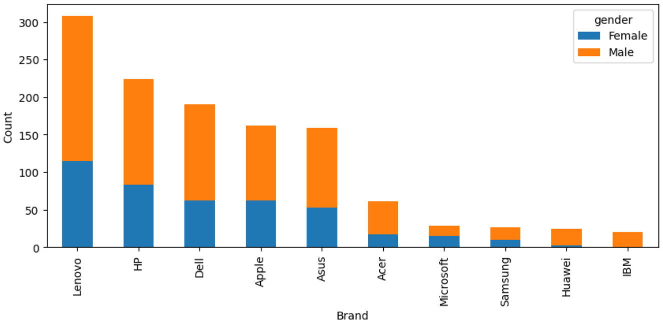


Figure 8. Demand Analysis Chart.

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 widen our vision to the market demand of the Urdu speaking region. This information can be used by the businesses to make informed decisions and enhance their business in today's competitive market.

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) [Preprint]. doi: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: 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. [Online]. Available.

[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: https://doi.org/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. Lal et al., "A Systematic Study of Urdu Language Processing its Tools and Techniques: A Review," International Journal of Engineering Research Technology, vol. 9, no. 12, Dec. 2020, doi: https://doi.org/10.17577/IJERTV9IS120031.