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 Persian and 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 Persian and Urdu. The study involves data collection from various social media platforms to gather a comprehensive corpus of Persian and Urdu text, followed by preprocessing steps to clean, normalize, and transform the data. State-of-the-art NLP models and algorithms are employed to extract insights, capturing consumer sentiment and identifying specific entities such as brand names and product features.Additionally, the research introduces the largest annotated dataset for Persian and 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 Persian and Urdu-speaking regions, enabling businesses to make informed decisions, tailor strategies, and enhance their competitive advantage. Finally The effectiveness of the approach is evaluated through extensive experiments and comparisons with existing NLP techniques, utilizing performance metrics such as accuracy, precision, and recall.**

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

I. Introduction

The economic climate is changing rapidly, which makes it essential for businesses to have a well-defined plan to succeed in a highly competitive market. A strong business plan not only provides direction but also establishes a perception for start-ups. To launch a new business venture in a highly competitive industry, entrepreneurs must have a thorough understanding of the most in-demand gender-specific products in the current market. To achieve this, a sophisticated approach is required to engage target customers and understand their interests. Sentiment analysis is a powerful tool that can help entrepreneurs collect, quantify, and analyze consumer perceptions to develop effective business strategies. Social media has emerged as one of the most significant online marketplaces, providing entrepreneurs with opportunities to engage with consumers and study their behavior. Social media sentiments play a crucial role in influencing the stock market demand and price, and various business sectors are significantly impacted by social platforms.

II. Literature Review

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. 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

We propose a method for analyzing the opinions and genders of social media users on different laptop brand. Our goal is to identify the most popular and demanding laptop brand in Persian and Urdu speaking regions. We describe the main components of our method, which include sentiment analysis, gender prediction, and named entity recognition, as shown in Figure [1]

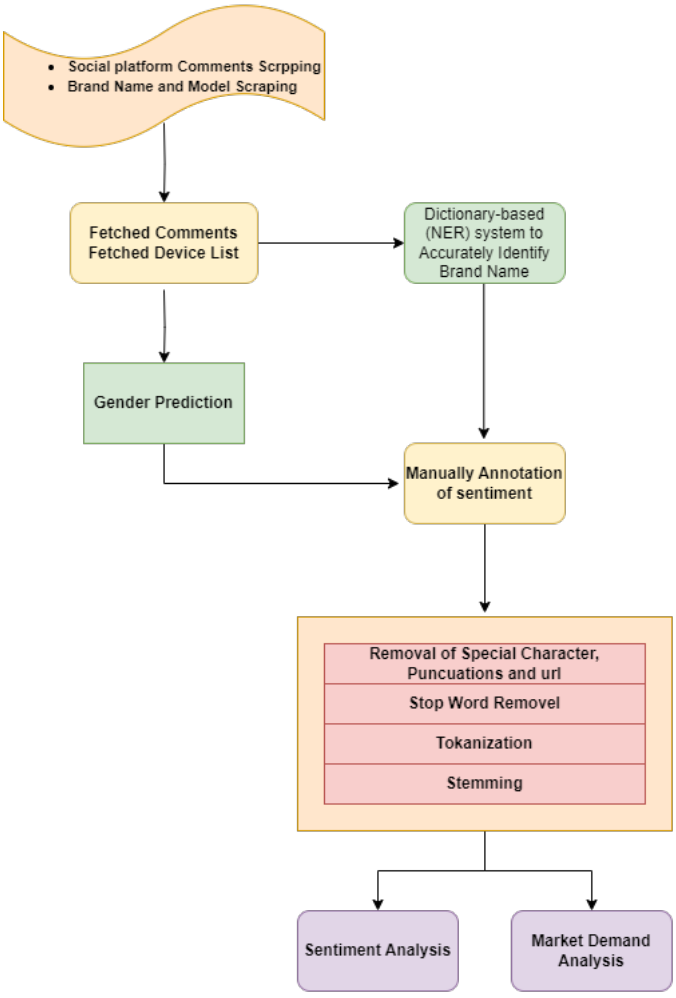

Fig. 1. work model flowchart.

To begin with our approach, we focused on the following steps:

A. Data Collection:

In our study we focused on market demand analysis in Urdu languages, we encountered a scarcity of existing research and organized data specifically related to customer product reviews. To address this challenge, we turned to social networking platforms, specifically laptop-related YouTube channels and public pages, as valuable sources for collecting consumer preferences and reviews. Using the Instant Data Scraper tool, we collected a significant volume of comments from diverse social networking platforms, resulting in a dataset comprising over 37,000 raw entries. To identify laptop brand names within the comments, we supplemented our data with laptop model information obtained from Wikipedia using a Python web scraper, and developed a dictionary-based Named Entity Recognition (NER) system.

Furthermore, we categorized the comments into separate Urdu and English segments and undertook the translation of English comments into Urdu. This enabled us to merge both segments and facilitate comprehensive analysis. To enhance our analysis, we incorporated the gender guesser Python library to determine the genders associated with user names. In addition, we conducted data cleaning by excluding rows with no user names and comments that did not mention any brand names. Subsequently, we categorized the comments into negative, positive, and neutral categories with the assistance of native Urdu and Persian speakers.

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

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 both the Urdu and Persian 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 in both the Urdu and Persian languages. 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 [1], 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 Persian and 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 Persian and 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 noise caused by word variations.

Our methodology 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.

IV. Demand Analysis

After conducting extensive research, we eagerly awaited the outcomes of a substantial workload.

References

[1] R. Tatman, "Urdu Stopwords List," Kaggle, 2016. [Online]. Available.