Unveiling Persian Market Dynamics: A Comprehensive Analysis of Consumer Demand Using NLP Techniques.

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*—This research paper presents a comprehensive analysis of consumer demand dynamics in the Persian market using Natural Language Processing (NLP) techniques. With the growing significance of data-driven insights for businesses, understanding consumer preferences and market trends has become paramount. Our multifaceted data collection strategy amalgamates laptop-related comments from YouTube channels and Facebook public pages, supplemented by Urdu data translated into Persian for a comprehensive dataset. The effectiveness of our approach is rigorously evaluated through extensive experiments and comparisons with existing NLP techniques, utilizing performance metrics such as accuracy, precision, and recall. This research represents a significant step towards enhancing our understanding of market dynamics in languages with limited NLP support, ultimately benefiting businesses operating in these markets.**

***Index Terms*—Natural Language Processing, Persian Language, Dari language, Sentiment Analysis, XAI, Named Entity Recognition, Gender Prediction.**

I. INTRODUCTION

Persian Language is spoken by over 110 million people around the world, making it the 18th most spoken language in the world. Persian language is the national language of Afghanistan and Iran, but little progress has been made in developing natural language processing (NLP) techniques for Persian language. This is mainly due to its lack of available annotated dataset, diversity in dialects, and its complexity in morphology and syntax making it difficult to conduct sentiment analysis. As a result of these barriers, little work has been done in Persian language using natural language processing (NLP) in comparison to languages such as the English language that has made good progress due to its value to businesses and its impact on market demand. However, more demand for this work in English language does not reduce the significance of dynamic persian market demand.

In recent years, the use of natural language processing (NLP) techniques has emerged as a promising new way to analyze consumer demand. Sentiment analysis which is the subfield of natural NLP is focused on extracting the opinion, emotion and preference of users. These advancements in sentiment analysis are opening up new possibilities for businesses. We can track customer satisfaction, identify the problem with the products or services, and target our audience effectively. Knowing the consumer's preference would inform our insight on the market demand for both old businesses and new businesses in the modern competitive market. Additionally, it is crucial in understanding the market demand for making better policy and strategic decisions.

For this purpose, We applied an automatic data scraping method to collect raw text data from different social media platforms. This method is used for exploring consumer's behavior in the context of online buy and sell groups. After completion of data collection, we used natural language processing to filter and organize our data into a well-structured dataset. Finally, by applying different machine learning techniques, mainly NER and gender prediction, could train out a dataset.

In this paper, we investigate the use of natural language processing (NLP) to understand the market demand in Persian language. We gathered data from social media and then trained it to predicate the demand for laptop brands based on consumer reviews. Our model can identify the most preferable product based on the gender and can provide valuable insight to the business for their target market.

II. RELATED WORK

abc

III. METHODOLOGY

The purpose of this study is to look at the genders of social media users and the customer demand for various laptop brands in Persian. Additionally, the study uses various Natural Language Processing (NLP) methods to determine the most popular laptop brand in the Persian-Speaking region. As shown in Figure 1, the primary components of our system are sentiment analysis, named entity recognition, and gender prediction.

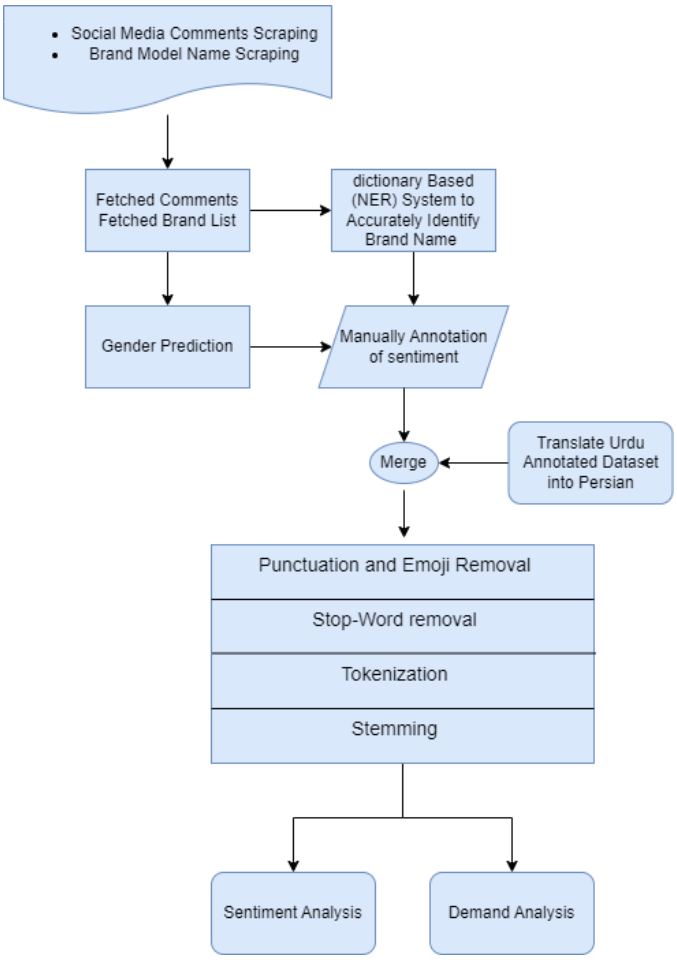

Figure 1. Flowchart of The Proposed Model.

To achieve the research objective, a multi-faceted data collection strategy was employed. Laptop-related comments from YouTube channels and Facebook public pages were utilized to gather consumer preferences and reviews. Additionally, Urdu data was collected and translated into Persian for inclusion in the project. The Instant Data Scraper tool was employed to efficiently gather comments from various social platforms, ensuring a diverse dataset for analysis. Laptop brand and model names were obtained from Wikipedia using a Python web scraper. The obtained model names were then translated using the Google API to ensure data consistency.

An important stage in the process was the creation of a function that used a dictionary-based system to match laptop brand and model names with the collected comment data. The comments were segregated into separate Persian and English segments. Translations were performed from English to Persian to ensure that both language sources were consistent. Data cleaning procedures were carried out to remove rows lacking usernames and comments that did not mention any laptop brands.

To gauge consumer sentiments, the collected comments were categorized into negative, positive, and neutral sentiments. This classification was carried out with the help of local Persian speakers to ensure proper emotion labeling. Various NLP techniques were employed for data preprocessing, including tokenization, stop word removal, and stemming. The dataset was then split into training and test sets for sentiment analysis.

The gender of social media users was predicted using the gender guesser Python library. Because Persian names were not supported, they were converted to English using the Google Cloud Translation API. Gender prediction was based on analyzing only the first names of users. This step aimed to add another layer of understanding by exploring the relationship between gender and consumer opinions.

The collected data underwent comprehensive cleaning and preprocessing using a combination of NLP techniques. Sentiment analysis was performed using both traditional machine learning algorithms and deep learning models, allowing for a thorough exploration of sentiment trends across different laptop brands.

The results of the analysis were presented in the form of insightful visualizations, which included a list of the most demanding laptop devices based on gender in the current market. These visual representations allowed for a quick grasp of market dynamics and consumer preferences.

Through the execution of this methodology, robust datasets were constructed, capturing and analyzing consumer demand for laptop brands in Persian languages. The research yielded valuable insights into consumer preferences and behaviors, providing a deeper understanding of the market landscape in Persian-speaking regions.

In short, this comprehensive methodology employed a combination of NLP techniques to analyze consumer opinions and genders related to laptop brands in Persian-speaking regions. The strategic data collection, NER, sentiment analysis, and gender prediction allowed for a holistic exploration of the market dynamics, contributing to a better understanding of consumer preferences and behavior in this context.

IV. DATASET

To facilitate a comprehensive analysis of consumer demand in Persian-speaking regions, two primary datasets were created. The first dataset consisted of customer

product reviews, providing essential textual data for sentiment analysis. The second dataset contained laptop brand names sourced from Wikipedia, ensuring accurate identification of brands for association with the sentiment analysis. This dual-dataset approach addressed the scarcity of organized data in the realm of Persian consumer demand analysis, enabling a more robust study.

A. Data collection

The data collection strategy employed in this study was multi-faceted, ensuring a diverse and representative dataset. Laptop-related comments were gathered from a variety of sources, including YouTube channels and Facebook pages, where consumer preferences and opinions were openly shared. Additionally, the study Urdu data were collected and translated into Persian for inclusion in the project, ensuring that a wide range of opinions and preferences were captured. The Instant Data Scraper tool was employed to efficiently collect comments from various social platforms, providing a rich collection of data for analysis. Moreover, laptop brand and model names were extracted from Wikipedia using a Python web scraper, adding a layer of specificity to the analysis. To guarantee data consistency, model names were translated using the Google API. Following rigorous preprocessing procedures, a meticulously labeled dataset was compiled, containing a total of 2500 entries. This dataset was enriched with sentiment, gender, and product entity information, allowing for a comprehensive analysis of consumer opinions, demographics, and their associations with specific laptop brands. This labeled dataset formed the cornerstone of the subsequent analysis, enabling insightful findings to emerge.

B. Preprocessing for Enhanced Analysis

Clean and high-quality data serves as the cornerstone of our study. In this section, we detail the meticulous steps undertaken to ensure the data's readiness for comprehensive analysis. Our approach involved the use of Python's regular expression library to systematically eliminate punctuation, special characters, and emojis from the Persian dataset. This not only enhanced the dataset's cleanliness but also provided a standardized foundation, free from unnecessary noise.

Furthermore, the process of noise reduction was complemented by the removal of common stop words. Leveraging the Persian NLP Toolkit, Hazm, allowed us to seamlessly eliminate these frequently occurring words, consequently elevating the overall quality of the text data and bolstering the accuracy of subsequent analyses.

Segmentation of text into individual tokens, known as tokenization, was another pivotal step in our preprocessing pipeline. By dividing comments into their constituent parts, we paved the way for in-depth examination and analysis of the textual content. This process proved instrumental in preparing the text data for extensive language processing.

Moreover, the tokenized language was standardized through the application of stemming techniques. This practice involved converting words to their base or root forms, effectively minimizing word variants and ensuring consistency throughout the dataset. Our systematic removal of prefixes and suffixes contributed to presenting words in their most fundamental structures.

The culmination of these preprocessing stages resulted in the creation of a meticulously refined dataset. From data collection and punctuation removal to stop-word elimination, tokenization, and stemming, each step was diligently executed to cultivate a dataset primed for insightful analysis. This polished dataset, a valuable asset, empowered our sentiment analysis efforts, enabling us to extract meaningful insights into consumer dynamics within the Persian-speaking market. Through this comprehensive data preparation process, we laid a robust foundation for exploring consumer preferences and behaviors, underscoring the significance of proper data preprocessing in our study.

V. NAMED ENTITY RECOGNITION

The vast amount of textual material on the internet, including sites like Facebook, blogs, and Wikipedia, is a true goldmine of knowledge. The constant evolution of techniques, algorithms, and tools to extract valuable insights from this ever-expanding digital landscape underscores the importance of Named Entity Recognition (NER) [2]. NER plays a pivotal role in identifying noun entities, including names, dates, times, and locations. In this study, we used a dictionary-based NER system to extract names of different laptop brands names from user comments on social media. This task holds fundamental significance within the domain of Natural Language Processing (NLP). Notably, our experiment encompassed laptop brand names in both the English and Persian languages. To address this multilingual challenge, we developed a bilingual dictionary-based system capable of effectively accommodating both linguistic contexts.

It is imperative to acknowledge that customers often employ diverse writing styles when referring to the same brand names. For instance, the brand 'HP' may be represented as 'Hp,' 'hp,' 'hP,' 'اچ پی', 'اچ-پی', and 'ایچ-پی' In recognition of this variation, our dictionary-based NER model was meticulously crafted to accurately identify different writing styles associated with each brand. This approach was thoughtfully designed to ensure precise extraction of brand names from user comments, even when expressed in various linguistic variations.

In line with this methodology, our NER algorithm was customized to recognize specific terms linked to individual brands. Additionally, we developed a function through manual coding and regular expressions. This function harnessed the dictionary-based approach to match device

names by comparing the laptop list with the content of the comments.

VI. GENDER PREDICTION

Understanding the gender base variations in product demand is essential for the business, As gender preferences exhibit significant diversity. In our research, we used three gender prediction libraries to identify the gender from the username. Initially, our approach involved the utilization of the Python gender guesser package, a tool designed to infer user genders based on given names. However, we encountered a various challenge along the way. The Python gender guesser package, while robust for many purposes, did not possess the capability to handle Persian names effectively. Recognizing the importance of inclusivity, we found a solution. To bridge this linguistic gap, we turned to the Google Cloud Translation API. This powerful tool allowed us to translate Persian names into English, rendering them compatible with the gender guesser package.

To further enhance our gender prediction accuracy, we conducted a comparative analysis. We leveraged two additional libraries, Genderizer and gender_guess, to assess the performance of our prediction model. This meticulous evaluation allowed us to fine-tune our approach and achieve improved results. Despite our diligent efforts to overcome this language barrier, we encountered persistent inaccuracies in gender predictions. These discrepancies underscored the complexity of the task at hand. In response, we adopted a refined strategy. We decided to exclusively rely on individuals' first names for gender prediction, simplifying the process while maintaining a degree of accuracy.

Through this evaluation and iterative improvements, we honed the gender prediction component of our approach. The result was an effective and reliable method for predicting gender based on first names, enhancing the precision of our analysis.

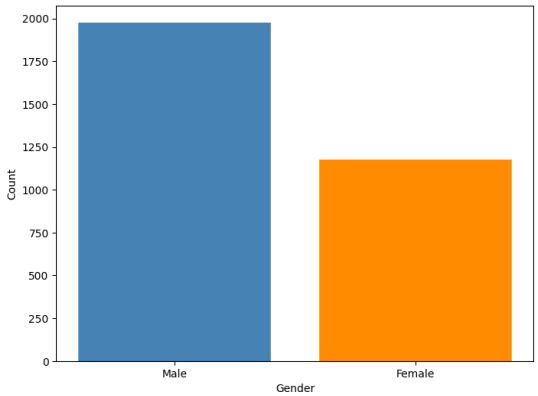

Figure 2. Distribution of Gender in the Dataset.

VII. SENTIMENT ANALYSIS

In the field of sentiment analysis we evaluated the effectiveness of machine learning and deep learning models using four performance metrics; Precision, Recall, F 1 Score and Accuracy. Table 1, below, presents a summary of the results obtained from evaluating these models on the task of sentiment analysis. The table demonstrates how each model performs across these metrics. Notably, both the Multinomial LR and SVM models show precision, recall, F 1 score and accuracy scores ranging from 91% to 92%. The Random Forest model also exhibits performance with a Score of 91% and an accuracy score of 91%. Although the Gradient Boosting model slightly falls behind with a score and accuracy score of 88% it still maintains results. Moreover, the Multinomial NB model delivers slightly lower scores across all metrics at 89%. The RNN CNN and LSTM models consistently perform well with precision, recall, F 1 score and accuracy values all at 91%. These findings highlight the competence of machine learning and deep learning approaches in sentiment analysis. They also provide a foundation for discussions on selecting models, for sentiment analysis applications.

Table I
MODEL PERFORMANCE

Models	Precision	Recall	F-1 Score	Accuracy
Multinomial LR	91	91	92	92
SVM	92	91	91	91
Multinomial NB	89	89	89	89
Random Forest	93	91	91	91
Gradient Boosting	89	88	88	88
RNN	91	91	91	91
CNN	92	92	92	92
LSTM	91	91	91	91

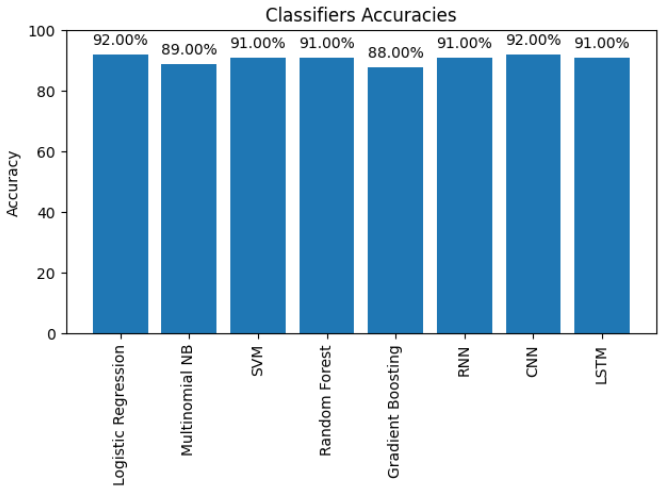

Figure 3. Classifiers Accuracies.

VIII. APPLYING LIME XAI METHOD

asd

IX. DEMAND ANALYSIS

In the pursuit of a comprehensive understanding of consumer demand in the laptop market, we conducted an in-depth analysis that shed light on the preferences of both male and female consumers. Our analysis revealed intriguing insights into the popularity of laptop brands and specific laptop models in Persian-speaking region.

We initiated our analysis by examining the popularity of laptop brands among male and female consumers. To visualize this, we generated a stacked bar chart illustrating the top ten laptop brands based on demand. The results indicated that Lenovo emerged as the most preferred laptop brand, followed closely by Dell, HP, Apple, Asus, and Acer, in both male and female consumer segments. This consistent ranking across genders highlights the universal appeal of these brands in the laptop market.

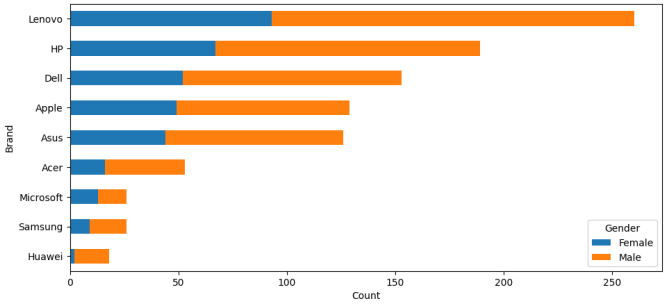

Figure 4. Top 9 laptop brands among male and female.

The demand analysis not only underlines the dominance of certain laptop brands but also highlights the significance of understanding the finer nuances in consumer preferences. Such insights are invaluable for businesses and policymakers looking to tailor their products and strategies to better meet the needs and desires of their target audience.

In conclusion, our demand analysis provides a comprehensive view of the laptop market, revealing the hierarchy of popular brands and specific laptop models among male and female consumers. These findings can serve as a foundation for informed decision-making, helping businesses navigate the dynamic landscape of consumer demand effectively.

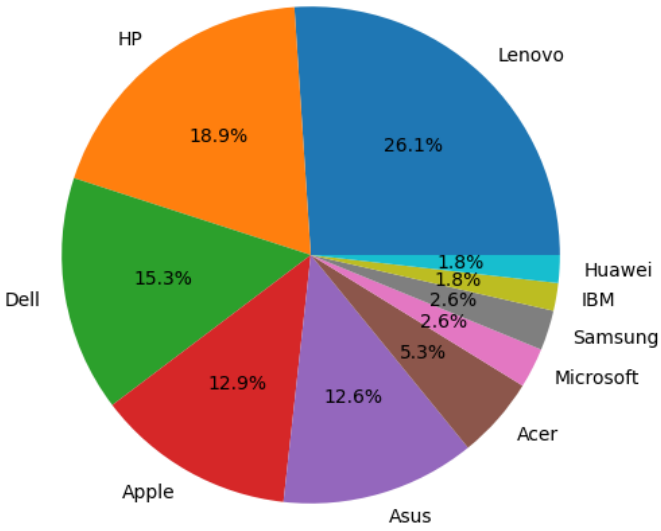

Figure 5. Top Ten Brands for positive sentiment.

X. CONCLUSION

In this paper we unveiled the market demand among Persian native speakers around the world by using NLP techniques which is greatly affecting business in the market. And by using NLP and machine learning techniques mainly gender prediction and NER model we could have a well structured and trained dataset. After we applied our models which were unable to give us valuable insights about the most popular products based on gender and sentiment analysis.

XI. FEATURE WORK

abc

REFERENCES

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

[2] S. Naseer, M. M. Ghafoor, S. bin K. Alvi, A. Kiran, G. M. Shafique Ur Rahmand, and G. Murtaza, "Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance.," Pakistan Journal of Multidisciplinary Research, vol. 2, no. 2, pp. 293–308, 2021. http://pjmr.org/pjmr/article/view/150