Product Demand Prediction In Current Market Using Twitter Feeds

Sai Samyuktha N
School of Computer Science and Engineering (SCOPE)
Vellore Institute of Technology
Chennai, India
saisamyuktha.n@gmail.com

Kaviya Sree R M
School of Computer Science and Engineering (SCOPE)
Vellore Institute of Technology
Chennai, India
kaviyasree.rm2020@vitstudent.ac.in

Harshitha Devina Anto
School of Computer Science and Engineering (SCOPE)
Vellore Institute of Technology
Chennai, India
harshithadevina.anto2020@vitstudent.ac.in

***Abstract*—This study explores the potential of using Twitter feeds to predict product demand in the current market. With the increasing use of social media, Twitter has become a valuable source of information for businesses to understand consumer behavior and preferences. The study analyzes Twitter feeds related to specific products and uses machine learning algorithms to predict future demand. The results show that Twitter feeds can be a reliable source for predicting product demand in the current market. This approach can help businesses make informed decisions about inventory management, marketing strategies, and product development. Overall, this study highlights the importance of social media in understanding consumer behavior and its potential for improving business operations.**

Keywords—social media analytics, machine learning, demand forecasting

I. Introduction

One out of every three consumers use social media to discover new products, services, or brands, 82% of which use such handles to make a purchase. The pandemic has drastically altered the way of shopping, shoppers tend to engage with different social media handles and digitize their shopping experience. Consumers are more likely to purchase a product with positive reviews and ratings, as it gives them confidence that the product has been tried and tested by others. Tweets reflect a diverse group of customer opinions & preferences, providing information about which products are in major demand. Analyzing Twitter feeds is a dynamic way to solicit customer feedback as it gives an insight on the customer's fondness towards a particular product, which can help companies identify areas for improvement according to public demand and make update their services to meet customer needs. Sentiment analysis of Twitter data will help forecast the demand for products in the current market and gain a better understanding of customer's needs, preferences, and concerns. This information is valuable to ensure effective targeted marketing strategies, plan the right amount of time and resources spent, stay competitive in the market, and shoot up the company's sales.

II. Literature Survey

A. Tweets Scraping

According to the paper [1] Latest tweets posted by influential users are retrieved by the data collector that uses the Twitter timeline API. Traditional text summarising techniques are no longer appropriate since tweets contain properties that are substantially distinct from those of conventional articles. There are two primary difficulties:

1. Twitter information is rather casual and random, with lots of acronyms and colloquial language; and

2. It is challenging to estimate the size of a Twitter data document. Twitter context trees can be built to group numerous tweets into documents of the right size. For every issue that was initially suggested by one of the significant users, a related context tree will be created. The original tweet sent by the important person serves as the root of this Twitter context tree, which is built based on the relationship between replies and retweets.

This study [2] examines the online demand for electronic goods using big data technology and neural network modelling. The findings demonstrate that all of the study's factors are effective predictors of product sales. The findings also demonstrate that variables from online reviews and online promotional marketing studies can be used to build a model that forecasts online sales of electronic goods. Since none of the predictors were eliminated during the sensitivity analysis, our model likewise shown that all of them are significant. This study has a number of significant ramifications. There are several limitations to this study. This study only initially considers electrical gadgets. Future study should examine the applicability of our model to diverse product types as well as additional factors. Second, our sample only contains a small number of records—roughly 30.000. However, as noted in the paper, this study coupled the usage of neural networks with big data architecture, providing a foundation for future researchers who wish to undertake studies with larger sample sizes. Finally, we focused only on Amazon.com in our investigation. Future research may compare results and assess generalizability by examining more online markets. The same is shown in figure 1.

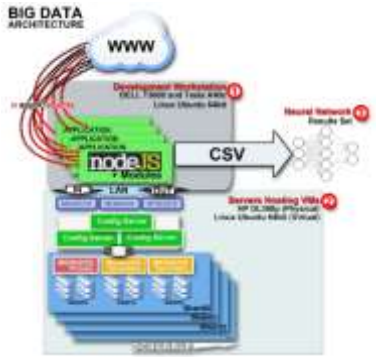

figure 1 : Big Data Architecture

This paper [3] aimed to predict the box office movie performance and opening weekend box office collection of any Bollywood movie. It used two components: a set of movie tweets from Twitter web sites and an actor/actress rating. The research investigated the important problem of mining opinions and sentiments from movie reviews, which had predicted the box office movie performance by classifying the movie into three categories: hit, flop and average. The proposed system initially selected any Bollywood movie for prediction before release, and the project work was broadly classified into different modules for application development.

The inability of the current study to crawl the data due to a lack of computational resources. There is incorrect information about ticket prices, the total number of seats per screen, and the total number of shows per day on all screens that can be used to anticipate box office receipts. In their upcoming work, the authors can include a number of additional inputs that will enhance the quality and accuracy of the prediction, such as the movie's budget, Central Board of Film Certification rating, genre, and target audience. Figures 2-6 depict the process flows and Figures 7,8 depict the results obtained and the corresponding limitation

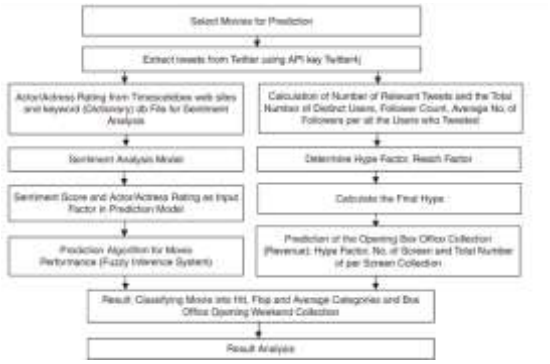

figure 2 : Block diagram of Proposed System

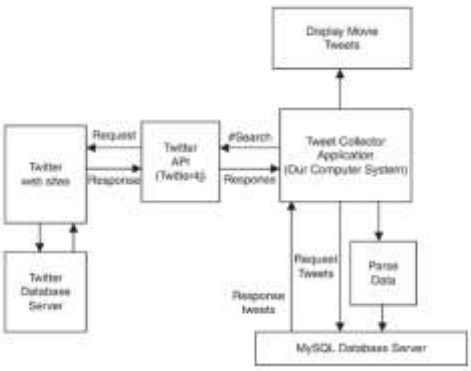

figure 3 : Fetching Data from Twitter

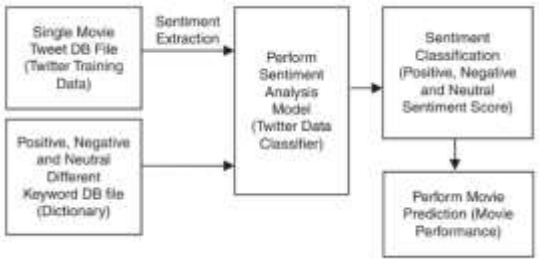

figure 4 : Tweets Sentiment Analysis

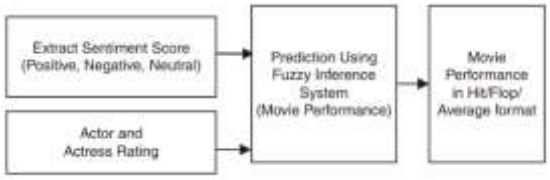

figure 5 : Prediction using Fuzzy Inference System

figure 6 : Movie box office prediction using hype factor (Box Office Revenue)

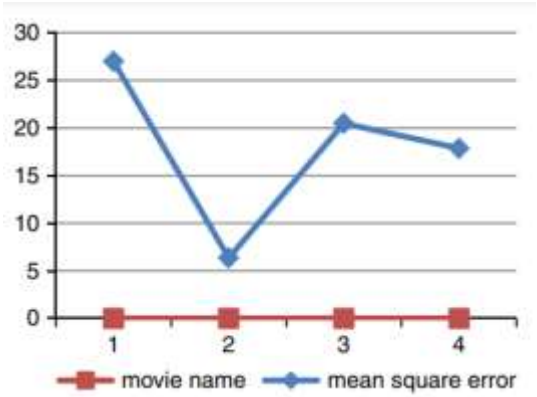

figure 7 : Means square error of final movies box of weekend collection using prediction method

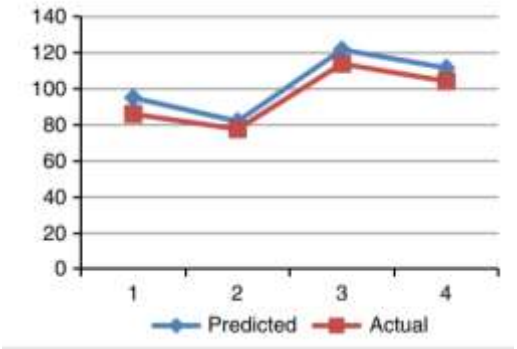

figure 8 : Actual vs predicted movie collection graph

B. Sentiment Analysis

Hossain [4] et.al have proposed a model that gathers Banglish text data and uses sentiment analysis and named entity identification to assess Bangladeshi market demand for smartphones in order to determine the most popular smartphones and classify their customers based on gender. Banglish text-based input was chosen as there are close to about 228 million native Bengali speakers, the majority of whom use Banglish text to interact with others on social media platforms. The purpose of their model was to determine the demand of products in the market in order to apply efficient

business strategies to stay competitive in the market. They have used social media sites for data as consumers evaluate and share their experiences on such platforms. The model focuses on the sentimental analysis of public comments data and gender prediction for detecting the most demanding device entities. The method involves data scraping of valid product name entities from tweets, and preprocessing data extracted from social media sites using Natural Language Tool kits and Regular Expressions. They trained the model to translate Banglish to Bangla text using Google Cloud Translation API for predicting the gender of a Banglish name. A TensorFlow sequential model was used as a sentiment analysis classifier and used Amazon Comprehend for custom-named entity recognition. They analyzed the tweets positive and negative demand and tagged those with an appropriate entity based on gender and sorted and plotted the most demanding devices based on gender in the current market. Their model had 95.51% in Amazon Comprehend Custom NER, 87.02% in the Sequential and 87.99% accuracy in Spacy Custom Named Entity recognition, for demand analysis, after which they managed to correct 80% of mistakes related to misspelled words using a mix of Levenshtein distance and ratio algorithms. The work model flowchart of the same is depicted in figure 9.

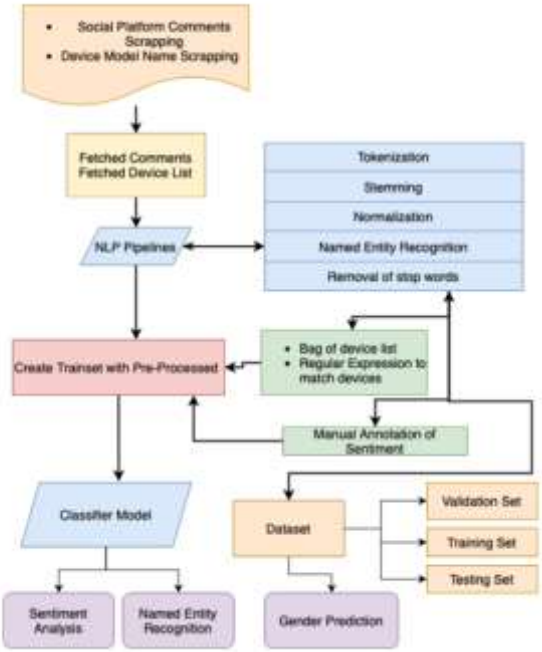

figure 9 : Work model flowchart

Social media platforms tend to have a huge pool of user-generated content on a particular topic, which keeps updating dynamically every day. Cossu [5] et.al proposed an automatic summarization model to provide guided automatically generated summaries of Micro-Blogs conversation, given a pool of tweets and a user-generated query, it provides a focused view of the pool. Their main goal was to verify if Automatic Summarizers are efficient enough to provide a reduced set of tweets that is informative enough regarding a given query. Tweet-selector summarizers provide the most representative tweet and answer to the query with information from the pool. These summaries are generated using key-word queries or sample tweet and offer a focused view of the whole Micro-Blog network. Their generic summarization system includes a set of eleven independent metrics combined by a Decision Algorithm. Query-based summaries were generated by using a modification of the scoring method. Cortex summarizer uses an optimal decision algorithm for processing statistical and informational algorithms on the document vector space representation. The Cortex system is applied to each document of a topic and the summary is generated by concatenating higher score sentences. Artex computes the score of each sentence, the summary is then generated concatenating the sentences with the highest scores. Their intuition was to guess the expected queries and summarize the tweets stream to ease a later clustering or classification stage. Instead of working at a single tweet granularity, the systems would be able to handle a summarized cluster.

The methodology used here is a combination of text processing, sentiment analysis, machine learning, and regression analysis to predict product sales based on the sentiment of tweets. The authors used this methodology to demonstrate the potential of sentiment analysis of Twitter data for improving sales predictions.

Identification and analysis of strawberries' consumer opinions [7] on twitter for marketing purposes focuses on using data mining techniques to identify and analyze consumer opinions about strawberries on Twitter. The aim is to gain a better understanding of consumer preferences and attitudes towards the fruit, which can then be used by marketers to inform their marketing strategies.
To conduct the research, the authors collected a large number of tweets about strawberries and used text mining techniques to extract information about consumer opinions. The tweets were analyzed for sentiment (positive, negative, or neutral), topics, and demographics (age, gender, location, etc.). The results of the analysis were then used to provide insights into consumer preferences and attitudes towards strawberries. For example, the authors found that consumers were more likely to express positive opinions about the taste and health benefits of strawberries, while negative opinions were more likely to focus on price and availability. These insights can help marketers to understand consumer preferences and develop targeted marketing strategies that address consumer concerns. Overall, the paper provides valuable insights into consumer opinions about strawberries on Twitter and demonstrates the importance of using data mining techniques to inform marketing strategies. By understanding consumer opinions, marketers can create more effective campaigns and build stronger relationships with their target audience. An example process flow is depicted in figure 10.

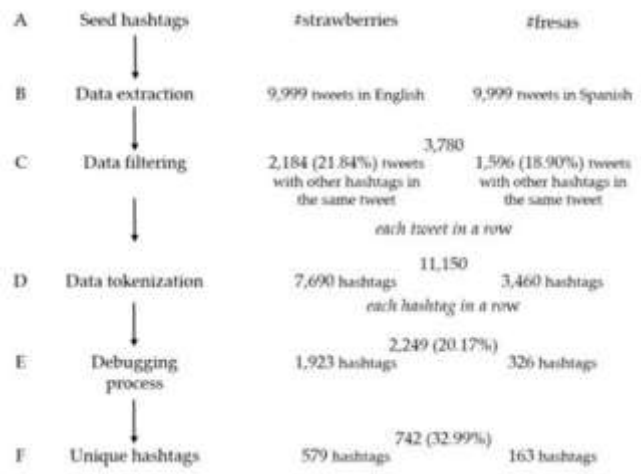

figure 10 : Sample Process flow

III. PROPOSED METHODOLOGY

Initially the user is prompted to enter the name of the product to be analyzed. This search query is now fed into the Twitter Data Scraper which uses the given search query to scrape all related tweets containing the particular tag or keyword. The scrapped tweets are then sent as inputs to the Text summarizer engine as well as the Sentiment Analysis engine.

The text summarizer would process all the tweets into a proper summary denoting the overall public opinion in few sentences. The Sentiment Analysis engine processes each tweet, categorizing it as negative, neutral and positive. Later using the output from both the text summarizer as well as the sentiment analysis engine, the overall product performance review is given. The same is depicted in the Architecture diagram in figure 11.

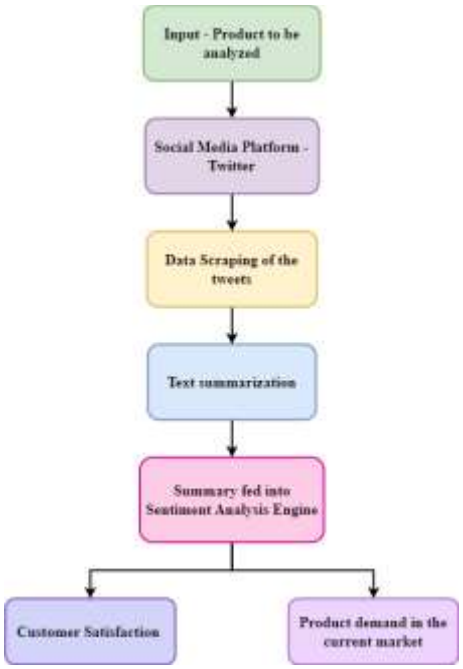

figure 11 : Architecture Diagram

A. *Tweets Scraping*

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Twitter data scraping can be done using several methods. Few identified methods include:

- Tweepy and Snscrape - using API keys
- Snscrape twitter module
- Twint API
- Selenium

In this paper, a model using Snscrape is implemented. snscrape is a scraper for social networking services (SNS). It scrapes things like user profiles, hashtags, or searches and returns the discovered items, e.g. the relevant posts. The following services are currently supported:

- Facebook: user profiles, groups, and communities (aka visitor posts)
- Instagram: user profiles, hashtags, and locations
- Mastodon: user profiles and toots (single or thread)
- Reddit: users, subreddits, and searches (via Pushshift)
- Telegram: channels
- Twitter: users, user profiles, hashtags, searches (live tweets, top tweets, and users), tweets (single or surrounding thread), list posts, communities, and trends
- VKontakte: user profiles
- Weibo (Sina Weibo): user profiles

B. *Text Summarization*

Text summarization is a process that involves condensing large volumes of information into a shorter format while still retaining the most important information. It is a valuable tool for making sense of lengthy documents, such as news articles, research papers, and legal documents. Text summarization can be achieved through two main methods:

1. Extractive

Extractive text summarization involves selecting and combining key sentences or phrases from the original text to create a summary. This approach is popular because it preserves the original language and structure of the text, making it easier for readers to understand the context and meaning of the summary. The process of extractive summarization begins by identifying important ideas and phrases within the text. This can be done manually or using automated techniques, such as natural language processing algorithms. Once the key phrases and sentences have been identified, they are combined to create a summary that captures the main ideas of the document.

Extractive summarization has several advantages over other summarization techniques. For example, it is more straightforward and easier to implement, as it does not require generating new language or understanding the underlying

meaning of the text. Extractive summarization also tends to produce summaries that are more accurate and representative of the original text, since it relies on selecting actual sentences and phrases rather than generating new ones.

However, there are also some limitations to extractive summarization. Because it relies solely on selecting sentences and phrases from the original text, it may not capture the overall meaning or tone of the document. It can also be difficult to identify which sentences or phrases are most important, particularly in longer documents.

2. Abstractive

Abstractive text summarization involves generating new sentences that capture the main ideas of the original document rather than just combining key sentences or phrases. This approach requires more sophisticated natural language processing techniques than extractive summarization, as it involves understanding the meaning and context of the original text and using that understanding to generate new language. The process of abstractive summarization begins by analyzing the original text to identify key concepts, themes, and ideas. This is done using a variety of techniques, including machine learning algorithms and statistical models. Once the key ideas have been identified, the summarization algorithm generates new sentences that convey the same information in a more concise format.

Abstractive summarization has several advantages over extractive summarization. It can produce summaries that are more concise and easier to understand, as the algorithm is able to generate new language that accurately conveys the meaning of the original text. Abstractive summarization is also more flexible than extractive summarization, as it can be used to summarize texts of any length or complexity.

But again, there are also some limitations to abstractive summarization. Because it involves generating new language, the summaries it produces may not always accurately reflect the tone or style of the original text. Abstractive summarization also requires more advanced natural language processing techniques than extractive summarization, which can make it more difficult to implement and less accurate in some cases.

Text-to-Text Transfer Transformer model

The Text-to-Text Transfer Transformer model (T5) is a state-of-the-art and consistent deep learning model for abstractive text summarization. It is a variant of the Transformer architecture, which was originally proposed for machine translation tasks. The T5 model is trained using a large corpus of text data and a technique called unsupervised pre-training where the model predicts missing words or phrases within a given text. By doing this, the model learns to understand the structure and meaning of language, which can be applied to a wide range of natural language processing tasks.

To use the T5 model for abstractive text summarization, the input text is first encoded into a fixed-length vector representation using the encoder component of the model. The decoder component then generates a summary from this vector representation, using a technique called beam search.

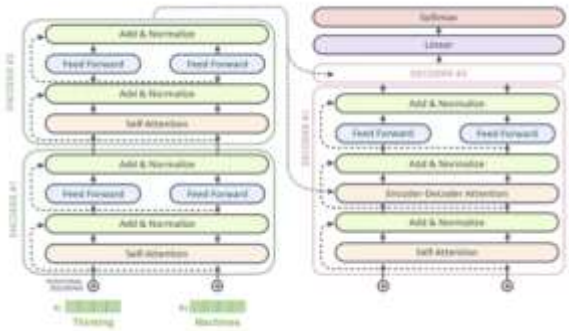

figure 12 : Model Structure

One of the key strengths of the T5 model is its ability to generate high-quality summaries that are both concise and informative. Because it has been pre-trained on a large corpus of text data, it is able to generate summaries that capture the meaning and tone of the original text in a way that is both accurate and engaging. However, it requires a large amount of training data and computational resources to achieve optimal performance. It can also produce summaries that are overly generic or do not accurately capture the nuances of the original text.

The T5 model is a powerful tool for abstractive text summarization and has shown great promise in a wide range of applications. As natural language processing techniques continue to evolve, it is likely that the T5 model will remain a key tool for making sense of large volumes of text data. Text summarization can be used in product review reports in companies to extract key insights and opinions from large volumes of customer feedback. This can help businesses quickly identify common themes and issues that customers are experiencing with a particular product, as well as areas where the product is performing well. With heavy competitions between products in the market, companies can quickly identify trends and issues that might be impacting the customer experience.

Text summarization can also be used to generate short summaries of customer reviews that highlight the most important points. For example, a company might use a summarization algorithm to extract key phrases or sentences from customer reviews that mention specific product features, such as ease of use, durability, or price. These summaries can be useful for quickly identifying areas where the product is performing well or where improvements are needed.

C. Sentiment Analysis

Sentiment analysis of Twitter feeds is done next to get a sense of how customers are talking about their products and services and get insights to drive business decisions. This step classifies the scrapped Twitter feeds according to their polarity, such as positive, negative, and neutral.

This project does sentiment analysis using a powerful NLP library, Flair, which allows one to apply state-of-the-art natural language processing (NLP) models to the scrapped tweets from Twitter. The library is model-based and supports Flair embeddings which is used in Sequence labelling based on contextual string embeddings. It is an NLP framework that builds directly on Pytorch, enhancing models with new,

different approaches to combine Flair embeddings and classes in order to perform various NLP tasks. The pre-trained model comprises of two entities namely, the sentiment classifier output and the confidence score that ranges from 0 to 1. A confidence score of 1 is given when the model is very confident and perfectly sure that the result is accurate and with 0 being otherwise. Before predicting sentiment of a particular tweet, the input has to be tokenized by Sentence() which is passed into the classifier that is used to predict the sentiment. The predicted label is saved as a value and the prediction confidence is saved as a score.

Flair's interface allows to combine different word embeddings, takes contextual string embeddings, takes account of the context of whole sentence and not just individual words. This embedding-based model supports a variety of languages, comprises of python packages that use text representation to predict text sentiments and also permits combining different word embeddings together to yield better performance thus, Flair tends to be an advantageous model for sentiment analysis in these ways.

IV. Results and Discussion

The model had been tested on 5 products. The results are as follows:

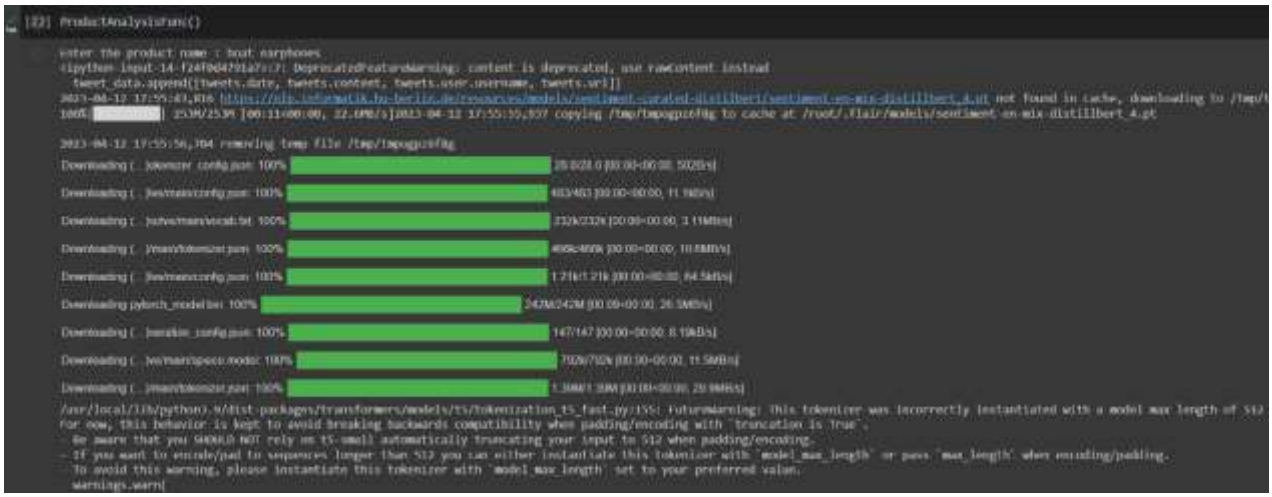

figure 13 : Boat earphones (A)

figure 14 : Boat earphones (B)

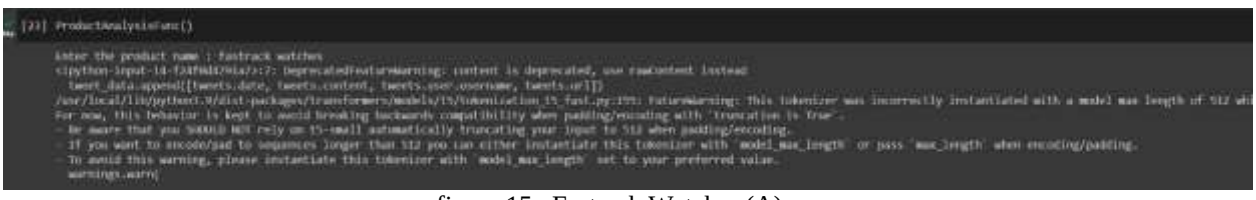

figure 15 : Fastrack Watches (A)

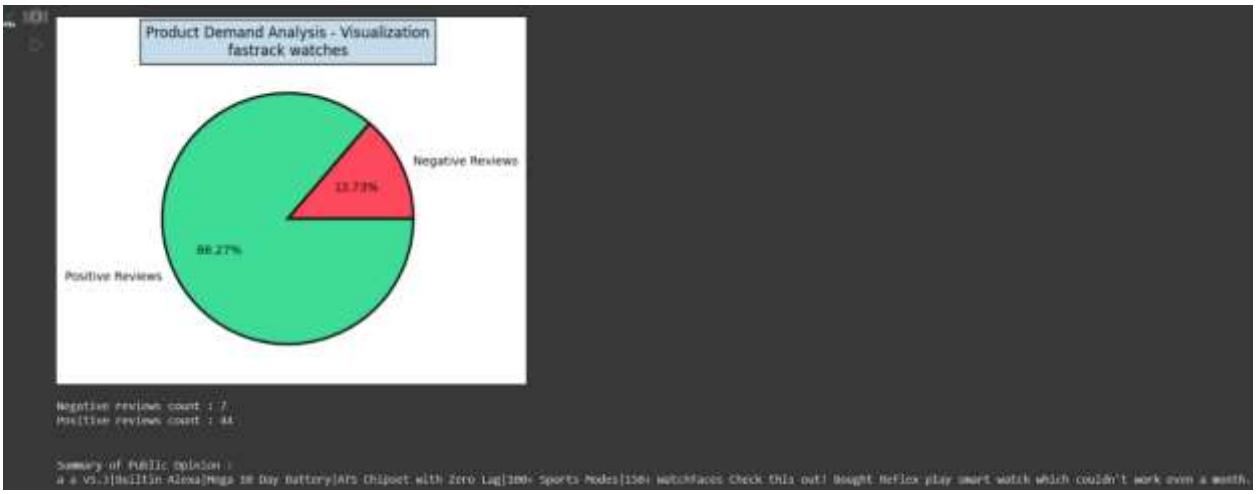

figure 16 : Fastrack Watches (B)

figure 17 : Mamaearth lotion (A)

figure 18 : Mamaearth lotion (B)

figure 19 : American Tourister bags (A)

figure 20 : American Tourister bags (B)

figure 21 : Adidas Shoes (A)

figure 22 : Adidas Shoes (B)

TABLE 1 COMPARATIVE ANALYSIS OF PRODUCTS

Product	Positive Reviews %	Negative Reviews %
Boat earphones	74.51	25.49
Mamaearth Lotion	60.78	39.22
American Tourister Bags	72.55	27.45
Adidas Shoes	49.02	50.98
Fastrack Watches	86.27	13.73

Based on the above comparison of various product reviews from Twitter feed based on the customers, Fastrack watches are thriving the most comparatively than the other four products in the current market.

V. Comparision of Models

Two Sentiment Analysis models have been used in the analysis of product reviews – VADER and Flair. A comparison of the same is as below.

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analyzer tool. Vader is optimized for social media data and can yield good results when used with data from Twitter, Facebook, etc. This model uses a list of lexical features which are labelled as positive or negative according to their semantic orientation to calculate the text sentiment. Vader sentiment returns the probability of a given input sentence to be [Positive, Negative, Neutral] These 3 probabilities will add up to 100%. The Vader model is a rule-based sentiment analysis model, which adheres to rules known as lexicons. Based on these lexicons, the words from sentences fed into the model are classified are either positive or negative along with their corresponding intensity measure. The main drawback with this rule-based approach for sentiment analysis is that the method only cares about individual words and completely ignores the context in which it is used. Thus, if negative and positive words are used in a sentence together, it would cancel each other's effect and give a neutral sentiment output, instead of actually considering the context of the words used in the particular sentence. Hence the Vader model is comparatively inaccurate in predicting the sentiment for our tweets classification based on sentiment analysis.

Flair is an embedding based model; it comprises of python packages that use text representation to predict text sentiments. This leads to better text representation in NLP and yields better model performance. Flair's interface allows to combine different word embeddings, takes contextual string embeddings, taking account of the context of whole sentence and not just individual words. Thus, Flair model is able to accurately predict the tweet's sentiment by analyzing the context of words that are used in sentences in the tweet product reviews. Hence, Flair model is advantageous over the Vader model in predicting the sentiment for Twitter review feeds.

VI. Conclusion and Future Work

In this project, tweets from Twitter are extracted using Snscrape, a python-based scraping library that allows users to extract data from social media platforms like Twitter. This scraped data is then fed to the sentiment analysis engine that uses the model flair to identify how well the product is doing in the market. The extracted data is also summarized in order to generate a short summary of customer reviews that highlight the likability of a product and scope of improvement in terms of customer service for the company. These processes are carried out for five products including – fastrack watches, adidas shoes, boat earphones, american tourister bag, mamaearth lotion to identify how much they thrive in the market and also aid the estimation of their demand.

In future, with introduction of even better models, we'll be able to better scrape data selectively – remove all unwanted tweets and summarize the scraped data properly with proper semantic sense.

Acknowledgment

We wish to express our sincere thanks and deep sense of gratitude to our project guide, Dr. G. Anushiya Rachel, SCOPE, for her consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

References

[1] Fu, K., Lu, Y. C., & Lu, C. T. (2014, November). Treads: A safe route recommender using social media mining and text summarization. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 557-560).

[2] Chong, A. Y. L., Ch'ng, E., Liu, M. J., & Li, B. (2017). Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews. *International Journal of Production Research*, *55(17)*, 5142-5156.

[3] Gaikar, D. D., Marakarkandy, B., & Dasgupta, C. (2015). Using Twitter data to predict the performance of Bollywood movies. *Industrial Management & Data Systems*.

[4] Sabbir Hossain, M., Nayla, N., & Alim Rasel, A. (2022). Product Market Demand Analysis Using NLP in Banglish Text with Sentiment Analysis and Named Entity Recognition. *arXiv e-prints*, arXiv-2204.

[5] Cossu, J. V., Torres-Moreno, J. M., SanJuan, E., & El-Bèze, M. (2020). Intweetive Text Summarization. *arXiv preprint arXiv:2001.11382*.

[6] Gaikar, D., & Marakarkandy, B. (2015). Product sales prediction based on sentiment analysis using twitter data. *International Journal of Computer Science and Information Technologies*, *6*(3), 2303-2313.

[7] Borrero, J. D., & Zabalo, A. (2021). Identification and analysis of strawberries' consumer opinions on twitter for marketing purposes. Agronomy, 11(4), 809.