Finding Purchase Intentions using Social Media Implementation

Prof. Sonu Khapekar, Rokade Jayesh, Shaikh Irfan, Patil Vaishnavi

Computer Engineering Department
Nutan Maharashtra Institute of Engineering and Technology, Pune, Maharashtra

Abstract—Recently, there has been a significant increase in the ecommerce industry, specifically in people purchasing goods online. A lot of research has been conducted to determine a user's purchasing patterns and, more importantly, the factors that determine whether or not the user will purchase the product. In this study, we will investigate whether it is possible to identify and predict a user's purchase intention for a product, and then target that user with a personalized advertisement or deal. Furthermore, we hope to create software that will assist businesses in identifying potential customers for their products by estimating their purchase intention in measurable terms based on their tweets and user profiles on Twitter. We have discovered that it is possible to predict whether or not a user has expressed a desire to buy a product after applying various text analytical models to tweet data. Additionally, our analysis has shown that the majority of users who had initially expressed a desire to buy the product have also purchased it.

Keywords—*Natural Language Processing, Product, Purchase Intention, Tweets, Twitter*

INTRODUCTION

Numerous studies have been conducted to analyze the purchasing patterns of internet customers. Few, though, have addressed the intention of customers to purchase products. Our goal is to create a machine learning method that can identify possible buyers of a product by quantifying the intention to buy based on tweets. Although text analytics can be done manually, it is inefficient, so we have used a machine learning approach based on text analysis. Finding patterns and trends will be much faster and more effective when text mining and natural language processing algorithms are used. We can say that the task of detecting purchase intentions is somewhat similar to the task of determining desires in product reviews.[1] There are numerous recommendation systems out there right now that present the user with various product recommendations; however, the majority of them are ineffective. There isn't a model that works well for businesses to find potential clients. Additionally, a number of research studies have been conducted to analyze the purchasing patterns of internet users.[2] Few, though, have addressed the intention of customers to purchase products.

I. LITERATURE SURVEY

Many studies have been carried out to analyze the insights of online consumer purchasing behavior. Few, though, have talked about the intentions of customers to purchase particular products. Research on the The task of identifying "buy" desires from text has been the main focus of wishes identification from product reviews, particularly in the works of Ramanand Bhavsar and Pednekar. These desires might be to purchase the product or to make suggestions for it. To identify these two types of wishes, they employed linguistic riles. While rule-based methods for purchasing or determining the wishes work well, they have limited coverage and are difficult to expand. The task of detecting purchase intentions is similar to the task of determining wishes from product reviews. Instead of using a rule-based strategy in this case, we offer a machine learning strategy that uses generic features taken from the Tweets. Previous research has demonstrated that Named Entity Recognition (NER) and Natural Language Processing (NLP) can be applied to tweets. Applying NER to tweets, however, is very challenging because users frequently use acronyms, misspell words, and make grammatical mistakes. However, Finn et al. used crowdsourcing to identify the annotated entities in tweets. Sentiment analysis is applied to tweets in other studies. Because product or movie reviews can be either positive or negative, they were used in the initial studies. Wang et al. and Anta et al. looked over the viewpoints conveyed through tweets that had been filtered using a specific hashtag—words or phrases that denigrate the tweet's main subject. These studies only examine the tone of a tweet that an author sends after purchasing a product. The sentiment 140API, which lets you find the attitude towards a particular product, brand, or subject on Twitter, is one preprocessing method frequently used for Twitter data. Speech tagger with tokenizer from Twitter NLP library. hierarchical word clusters or structures, as well as a tweet dependency parser. throughput.

II. ARCHITECTURE

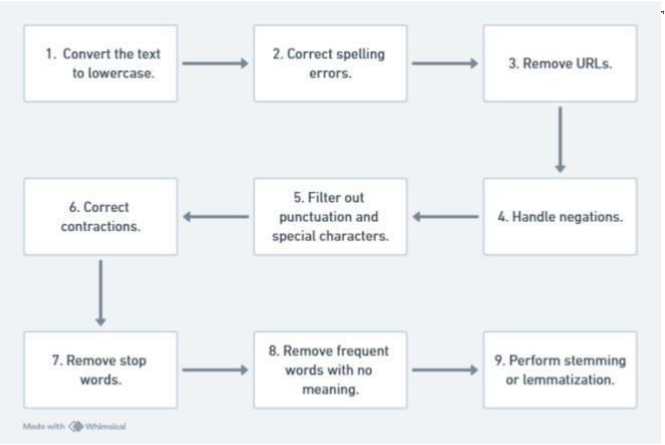

1. Convert the text to lowercase.
2. Correct spelling errors.
3. Remove URLs.
4. Handle negations.
5. Filter out punctuation and special characters.
6. Correct contractions.
7. Remove stop words.
8. Remove frequent words with no meaning.
9. Perform stemming or lemmatization.
Made with Whimsical

Figure 1 System Architecture

Steps of Machine Learning Modelling

1.Dataset - The majority of our dataset comes from Twitter. Regardless of whether a review is favourable or negative, people who tweet about a product using the Twitter app are taken into account.[16]

2.Data Preprocessing - This step involves getting the unprocessed data ready so that a machine learning model can use it. It is the initial and most important stage in the development of a machine learning model.[17] Cleaning the data and preparing it for a machine learning model are necessary steps in the data preprocessing process, which also improves the machine learning model's accuracy and efficiency.[18]

3.Train Test Split - After training and testing the data, split conditions are applied, which are then used to apply various machine learning models.

4.Application of ML models - 1. Logistic regression 2. The Tree of Decisions 3. Ignorant Bayes 4. Vector Machine Support

5.Plotting the results - The best algorithm that provides the greatest degree of precision is accuracy is taken into account when calculating the results.

6.Data Visualization - For effective analysis and outcome prediction, data can be visualized as two-dimensional rows and columns, pie diagrams, or histograms.

7.Display on website - Setting up the website so that visitors can go there to obtain pertinent data about the consumer's anticipated intentions with respect to that specific product.

A. Data Preprocessing (Text Preprocessing)

1) LOWERCASE: To achieve case uniformity, we began our foundational work by converting the text to lower case.
2) REMOVE PUNC: After that, we sent the lowercase text to functions that remove punctuations and special characters. Unwanted characters, spaces, tabs, and other elements that have no purpose in text classification may be present in the text.
3) STOPWORDSREMOVAL: Text may include words that are routinely included in sentences but serve no purpose or add anything to the sentence's meaning. The words "a," "an," and "in" are mentioned.
4) REMOVAL OF COMMON WORDS: In addition, the sentence contains a huge amount of number of words that are repeated but do not add to its meaning.[19]
5) RARE WORDS REMOVAL: We also got rid of a few uncommon words, like names and brand names that weren't enclosed in html tags. These are the special terms that don't really add anything to the interpretation model.
6) SPELLING CORRECTIONS: Spelling errors abound in social media data. It is our responsibility to eliminate errors and provide accurate words for our model.
7) STEMMING: After that, we went back to the words' etymologies. Words that have had their ends or beginnings removed are known as stemming words.[21]
8) LEMMATIZATION: Next, we subjected our text to lemmatization. The order in which this analysis is done is morphological. One can trace a word back to its lemma.
Following preprocessing, we are left with 1300 tweets for testing and training our model.

B. Text Visualization

The process of extracting information from raw text and applying various analyses to extract significant structure and pattern is known as text visualization. Tools for natural language processing (NLP) can help with this. They are able to identify popular attitudes toward a specific subject or item (sentiment analysis). To visualize the subjectivity and sentiment polarity of the tweets in the dataset, we can utilize the Seaborn library.

Positive Tweets :

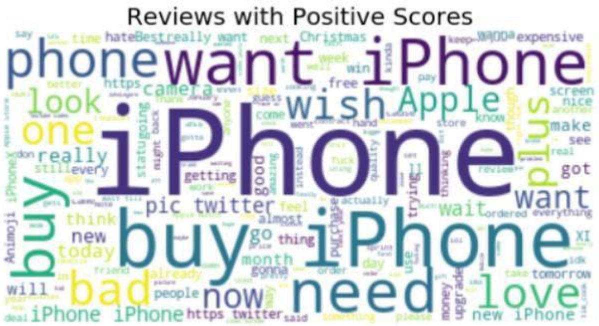


Figure 2 Positive Tweets

Negative Tweets :

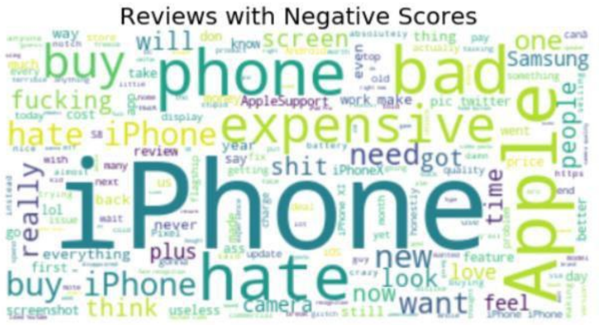


Figure 3 Negative Tweets

C. ML Models

Once the corpus was ready, we used different text analytical models to test which one gives the best result. We use the following models

Logistic Regression :

One of the most widely used machine learning algorithms, under the category of supervised learning, is logistic regression. With a given set of independent variables, it is used to predict the categorical dependent variable.

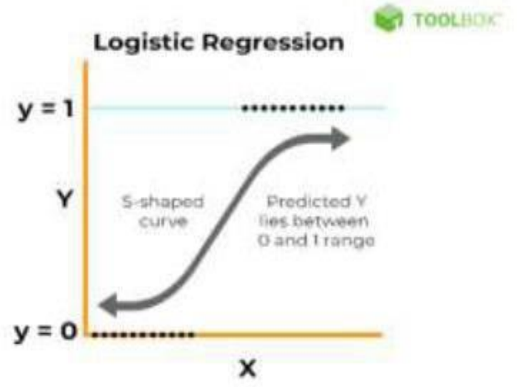


Figure 4 Logistic Regression

IV. ALORITHM

Decision Tree :

Although decision trees are a supervised learning technique, they are primarily used to solve classification problems. However, they can also be used to solve regression problems. It is a true-structured classifier, with internal nodes standing in for dataset features, branches for decision rules, and leaf nodes for each result.

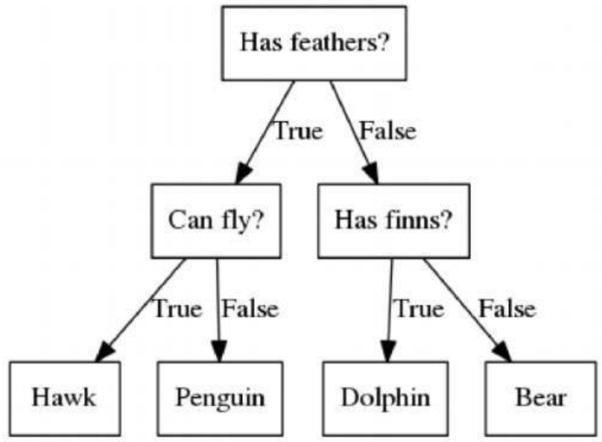


Figure 5 Decision Tree

SVM :

In order to make it simple to classify new data points in the future, the SVM algorithm seeks to identify the optimal line or decision boundary that can divide n-dimensional space into classes. We refer to this optimal decision boundary as a hyperplane.

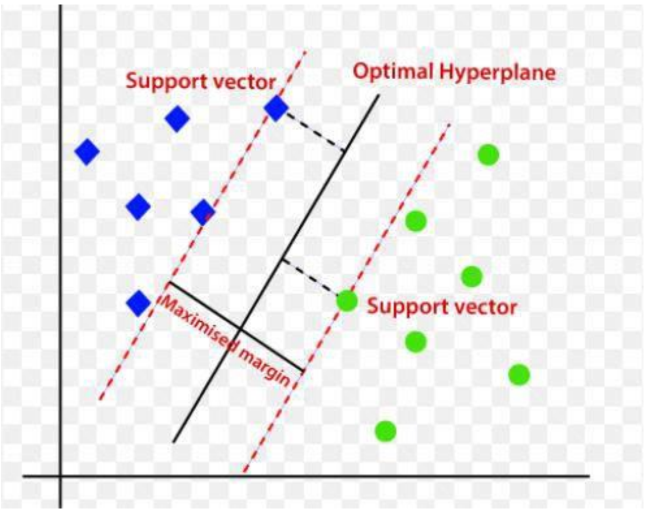


Figure 6 SVM

V. MATHEMATICAL MODEL

Naive Bayes :

One of the most straightforward and efficient classification algorithms, the Naïve Bayes classifier aids in the rapid development of machine learning models with rapid prediction capabilities. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur.

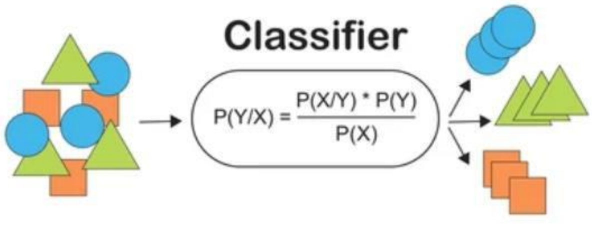


Figure 7 Naïve Bayes

Neural Network :

a technique in artificial intelligence that trains machines to handle data in a manner modeled after the human brain. Neural networks comprise of propagation functions, weights, biases, connections, propagation functions, and a learning rule.

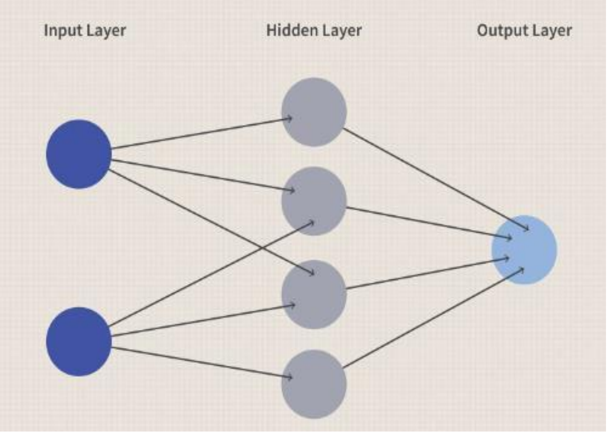
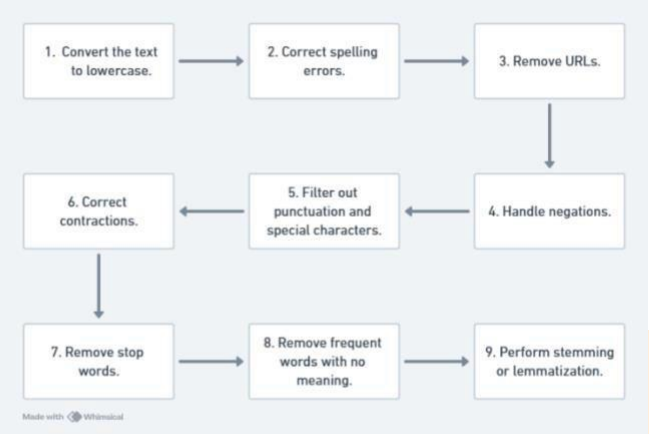


Figure 8


1. Convert the text to lowercase.
2. Correct spelling errors.
3. Remove URLs.
4. Handle negations.
5. Filter out punctuation and special characters.
6. Correct contractions.
7. Remove stop words.
8. Remove frequent words with no meaning.
9. Perform stemming or lemmatization.
Made with Whimsical

Neural Network

VI. RESULTS AND DISCUSSIONS

Model evaluation is the method of analyzing a machine learning model's performance and strengths and weaknesses using various evaluation metrics.
To assess our models, we employ the subsequent methods:

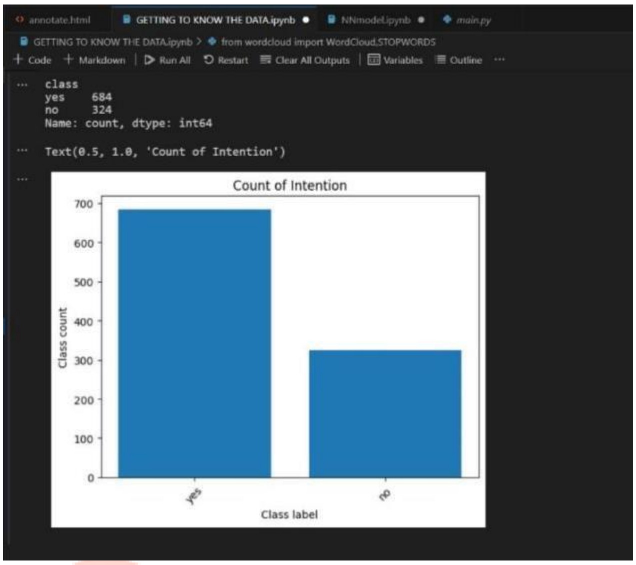
1. Confusion Matrix: An evaluation tool for machine learning classification problems where multiple classes may be produced as output. There are five distinct combinations of the expected and actual values in the table.

2. Precision: Precision Quantity of accurate forecasts total count of forecasts. Accuracy for binary classification can also be computed as follows in terms of positives and negatives: T P + T N T P + T N + F P + F N equals accuracy. where FN = False Negatives, FP = False Positives, TN = True Negatives, and TP = True Positives.

3. Precision and Recall: In machine learning, precision and recall are performance metrics used for classification and pattern recognition. To create the ideal machine learning model that produces more exact and accurate results, these ideas are necessary.

4. F-measure: Since the F-measures do not account for true negatives, it may be preferable to use metrics like Cohen's kappa, the Matthews correlation coefficient, or informedness to evaluate a binary classifier's performance.

5. True-Negative Rate: An outcome in which the model forecasts the positive class with accuracy is referred to as a true positive. A true negative, on the other hand, is a result in which the model accurately predicts the negative class. When the model predicts the positive class incorrectly, the result is called a false positive.



```
class
yes    684
no     324
Name: count, dtype: int64

Text(0.5, 1.0, 'Count of Intention')
```

Figure 9

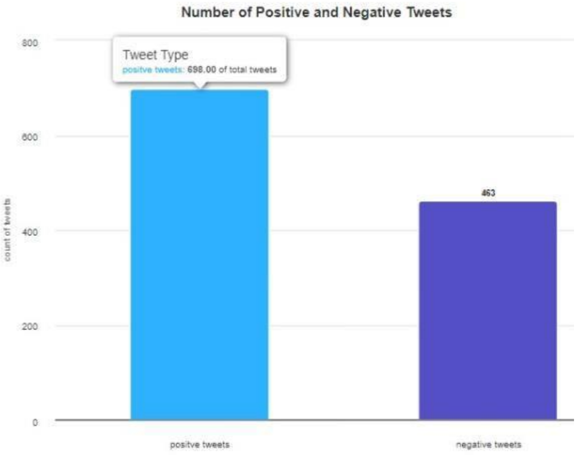
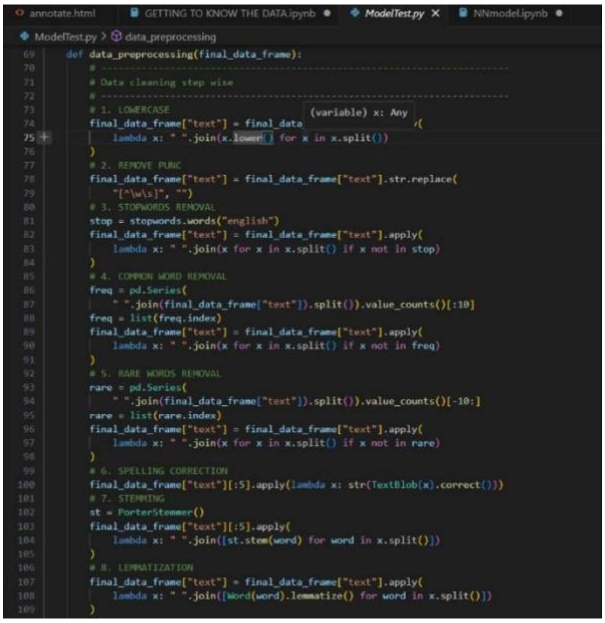
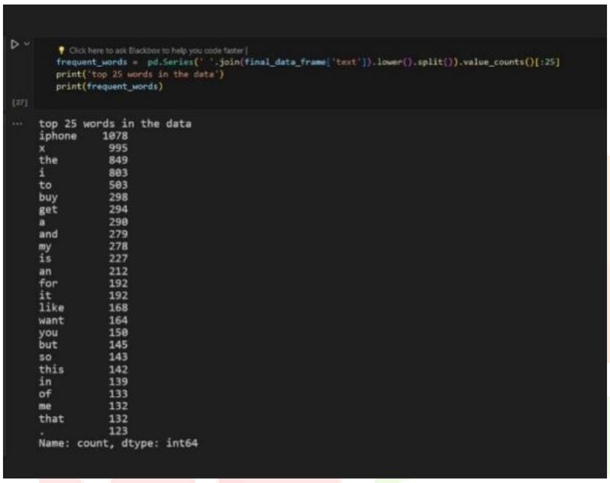


Figure 10


annotate.html
GETTING TO KNOW THE DATA.ipynb
ModelTest.py
NNmodel.ipynb
ModelTest.py > data_preprocessing

```
def data_preprocessing(final_data_frame):
    # ----------------------------------------------------------------------
    # Data cleaning step wise
    # ----------------------------------------------------------------------
    # 1. LOWERCASE
    final_data_frame["text"] = final_data (variable) x: Any (
        lambda x: " ".join(x.lower() for x in x.split())
    )
    # 2. REMOVE PUNC
    final_data_frame["text"] = final_data_frame["text"].str.replace(
        "[^\w\s]", "")
    # 3. STOPWORDS REMOVAL
    stop = stopwords.words("english")
    final_data_frame["text"] = final_data_frame["text"].apply(
        lambda x: " ".join(x for x in x.split() if x not in stop)
    )
    # 4. COMMON WORD REMOVAL
    freq = pd.Series(
        " ".join(final_data_frame["text"]).split()).value_counts()[:10]
    freq = list(freq.index)
    final_data_frame["text"] = final_data_frame["text"].apply(
        lambda x: " ".join(x for x in x.split() if x not in freq)
    )
    # 5. RARE WORDS REMOVAL
    rare = pd.Series(
        " ".join(final_data_frame["text"]).split()).value_counts()[-10:]
    rare = list(rare.index)
    final_data_frame["text"] = final_data_frame["text"].apply(
        lambda x: " ".join(x for x in x.split() if x not in rare)
    )
    # 6. SPELLING CORRECTION
    final_data_frame["text"][:5].apply(lambda x: str(TextBlob(x).correct()))
    # 7. STEMMING
    st = PorterStemmer()
    final_data_frame["text"][:5].apply(
        lambda x: " ".join([st.stem(word) for word in x.split()])
    )
    # 8. LEMMATIZATION
    final_data_frame["text"] = final_data_frame["text"].apply(
        lambda x: " ".join([Word(word).lemmatize() for word in x.split()])
    )
```

Figure 11



```
frequent_words = pd.Series(' '.join(final_data_frame['text']).lower().split()).value_counts()[:25]
print('top 25 words in the data')
print(frequent_words)
```

```
top 25 words in the data
iphone    1078
x          995
the        849
i          803
to         503
buy        298
get        294
a          290
and        279
my         278
is         227
an         212
for        192
it         192
like       168
want       164
you        150
but        145
so         143
this       142
in         139
of         133
me         132
that       132
.          123
Name: count, dtype: int64
```

Figure 12

VII. ADVANTAGES AND DISADVANTAGES

A. *Advantages :*

1)Effective and Scalable: Machine learning techniques are suited for real-time or big data applications because they process massive amounts of Twitter data far more quickly and efficiently than manual analysis.

2)Real-time insights: Our project can offer businesses timely insights into the purchase intentions of their target audience, allowing them to make necessary adjustments to their marketing plans.

3)Objective Analysis: By eliminating potential bias that could result from manual analysis, machine learning algorithms offer a more objective analysis of purchase intention.

4)Predictive Power: Your model may be able to forecast future purchase intentions by identifying patterns and trends from Twitter data, giving businesses the ability to proactively target potential clients.

5)Economical: Compared to hiring, the model's application to a large volume of data can be done at a comparatively low cost after it is developed.

VII. FUTURE SCOPE

Use Twitter's real-time data streams to help businesses react quickly to shifting customer attitudes and behaviour. Utilizing cutting-edge analytics and machine learning models that can adjust in real-time to new conversations and trends could be one way to do this. Expand the scope of the analysis to encompass multimedia content such as images and videos, in addition to text-based tweets. A more thorough knowledge of customer behaviour and preferences can be gotten by combining textual and visual data. Take into account the context of tweets. Purchase intentions can be substantially influenced by factors such as user demographics, location, weather, and events.

Subsequent investigations could concentrate on integrating these contextual factors into models of prediction. Examine how data from various social media platforms can be combined to obtain a comprehensive understanding of customer behaviour. Predictive accuracy can be improved by knowing how customers communicate and express themselves on various platforms. Go beyond sentiment analysis and examine Twitter user actions and behaviours. This might involve monitoring how consumers interact with particular product mentions, how they engage with brands, and whether or not these behaviours influence their propensity to buy.

VIII. CONCLUSION

When we compare our project to other studies in the same field, we can see that it is unique since we have tested four different models and have selected the best model based on the product data. The two issues listed below prevented us from achieving an accuracy rate of greater than 80%. It is a success to reach even 80% accuracy with such a small dataset and imbalanced class data.

The two main issues we encountered were:

1) The imbalance class issue: We had roughly 2000 positive tweets and 1200 negative tweets because we manually annotated our dataset. As a result, our model was not correctly predicting the negative class, and we were receiving very low True Negative Rates.

2) Limited annotated data: We were only able to annotate roughly 3200 tweets because we had to manually annotate every tweet in the dataset, which takes a lot of time.

ACKNOWLEDGMENTS

We want to especially thank our respected internal guide Prof. Sonu Khapekar for her guidance and encouragement which has helped us to achieve our goal. Her valuable advice helped us to complete our project successfully. Our Head of

Department Dr. Saurabh Saoji has also been very helpful, and we appreciate the support he provided us. He also gave us valuable input during our work. We would like to convey our gratitude to Dr. Vilas Deotare (Principal) all the teaching and non-teaching staff members of the Computer Engineering Department who gave us the freedom to explore and guided us the right way, also our friends and families for their valuable suggestions and support.

REFERENCES

[1] Rehab S. Ghaly, Emad Elabd, Mostafa Abdelazim Mostafa, Tweet's classification, hashtags suggestion and tweets linking in social semantic web, IEEE, SAI Computing Conference (SAI), 2016, pp. 1140-1146.

[2] Shital Anil Phand, Jeevan Anil Phand , Twitter sentiment classification using Stanford NLP, 1st International Conference on Intelligent Systems and Information Management (ICISIM), IEEE, 2017, pp. 1-5.

[3] Swati Powar, Subhash Shinde, Named entity recognition and tweetsentiment derived from tweet segmentation using Hadoop, 1st International Conference on Intelligent Systems and Information Management (ICISIM), IEEE, 2017,pp. 194 - 198.

[4] J. Kim, H. Lee, and H. Kim, "Factors affecting online search intention and online purchase intention," Seoul J. Bus., vol. 10, 2004.

[5] J. Ramanand, K. Bhavsar, and N. Pedanekar, "Wishful thinking-finding suggestions and'buy'wishes from product reviews," in Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, 2010, pp. 54–61.

[6] M. Hamroun, M. S. Gouider, and L. B. Said, "Customer intentions analysis of twitter based on semantic patterns," in The 11th international conference on semantics, knowledge and grids, 2015, pp. 2–6.

[7] M. J. A. Oele, "Identifying Purchase Intentions by Extracting Information from Tweets," 2017.

[8] M. Korpusik, S. Sakaki, F. Chen, and Y.-Y. Chen, "Recurrent Neural Networks for Customer Purchase Prediction on Twitter.," CBREcsys Recsys, vol. 1673, pp. 47–50, 2016.

[9] A. Go, L. Huang, and R. Bhayani, "Twitter sentiment analysis," Entropy, vol. 17, p. 252, 2009.

[10] P. Gamallo and M. Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets.," in Semeval@ coling, 2014, pp. 171–175.

[11] M. I. Eshak, R. Ahmad, and A. Sarlan, "A preliminary study on hybrid sentiment model for customer purchase intention analysis in social commerce," in 2017 IEEE conference on big data and analytics (ICBDA), 2017, pp. 61–66.

[12] S. Atouati, X. Lu, and M. Sozio, "Negative purchase intent identification in Twitter," in Proceedings of The Web Conference 2020, 2020, pp. 2796–2802.

[13] A. Sharma and M. O. Shafiq, "A Comprehensive Artificial Intelligence Based User Intention Assessment Model from Online Reviews and Social Media," Appl. Artif. Intell., pp. 1–26, 2022.

[14] D. Kumar, H. D. Mathur, S. Bhanot, and R. C. Bansal, "Forecasting of solar and wind power using LSTM RNN for load frequency control in isolated microgrid," Int. J. Model. Simul., vol. 41, no. 4, pp. 311–323, 2021.

[15] C. Olah, "Understanding LSTM Networks–colah's blog," Colah Github Io, 2015.

[17] Jon, "TweetScraper." Nov. 18, 2022. Accessed: Sep. 17, 2019. [Online]. Available: https://github.com/jonbakerfish/TweetScraper

[18] K. Crystal, "Scraping Twitter with Tweet Scraper and Python," Jun. 11, 2019. https://medium.com/@kevin.a.crystal/scraping-twitter-with-tweetscraper-and-python-ea783b40443b (accessed Jun. 08, 2021).

[19] P. A. Pavlou, "Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model," Int. J. Electron. Commer., vol. 7, no. 3, pp. 101–134, 2003.

[20] K. V. Ghag and K. Shah, "Negation handling for sentiment classification," in 2016 International Conference on Computing Communication Control and automation (ICCUBEA), 2016, pp. 1–6.

[21] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition."