Using Twitter Data Implementation to Determine Purchase Intention

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

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

***Abstract*—** The e-commerce sector has experienced tremendous growth recently, particularly in terms of the number of people making online purchases. Many studies have been carried out to ascertain the buying habits of users and, more crucially, the variables that influence the users' decision to buy the product. In order to target a user with a customized advertisement or offer, we will look into whether it is feasible to recognize and anticipate a user's intention to buy a product. In addition, our goal is to develop software that will help companies find prospective buyers of their goods by quantifying their intent to buy based on the users' Twitter profiles and tweets. After applying different text analytical models to tweet data, we have found that it is possible to predict whether or not a user has expressed a desire to purchase a product. Furthermore, the bulk of users who had initially indicated that they would like to buy the product have done so, according to our analysis.

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

INTRODUCTION

Several studies have been conducted to examine the purchasing habits of internet users. Few, however, have addressed customers' intention to buy products. Our goal is to develop a machine learning method for identifying potential product buyers by quantifying their intent to buy using tweets. Text analytics can be performed manually, but it is inefficient, so we used a text-based machine learning approach. Text mining and natural language processing algorithms will make it much easier and faster to find patterns and trends. We can compare the task of detecting purchase intentions to that of determining desires in product reviews. There are numerous recommendation systems available today that provide the user with various product recommendations; however, the majority of them are ineffective. There is no effective model for businesses to find potential clients. Furthermore, a number of research studies have been conducted to examine internet users' purchasing patterns. Few, however, have addressed customers' intention to buy products.

RELATED WORK

The topic of predicting purchase intentions has been the subject of numerous research investigations, each employing a different strategy. We think Ramanand et al.'s work [3], in which a corpus from well-known consumer review sites was produced, is the first known attempt in the literature. They employed a lexical rule-based methodology to forecast consumers' propensity to buy based on product reviews. For customer intention, Hamroun et al. used a semantic pattern-based method [4]. For Part-of-Speech (PoS) tagging in tweets and creating ontological representations of words, tools like OpenNLP and WordNet are useful. Using PoS tags in conjunction with these ontological representations, patterns were matched and intent predictions were made. Oele tried to use knowledge-rich, knowledge-poor, and their combinations to train several machine learning models in order to determine buy intents. Ten-fold cross-validation was used to test the models. Ninety percent of the potential customers could be predicted by the best model [5]. Deep learning techniques for forecasting purchase intentions from Twitter data were examined by Korpusik et al. [6]. However, because their data was restricted to Twitter users who finally tweeted after making their initial purchase, their problem set differed from this study's. They also increased the restrictions on data collection. Rather than predicting a tendency to buy, their study predicts whether a consumer will "buy" or "will not buy." Since our problem can be viewed as a derivative of the sentiment analysis task, as was previously discussed, we have also reviewed studies related to sentiment analysis on Twitter. Go et al. addressed the problem of sentiment analysis using data from Twitter by converting it to a self-supervised task with noisy labels. They used SVM, Naïve Bayes, and Max Entropy algorithms in addition to n-gram features [7]. Gamallo et al. employed a Naïve Bayes classifier but additionally carried out feature engineering steps like PoS tagging and Negation handling. They also eliminated neutral tweets by employing a polarity lexicon [8]. Eshak et al. defined s-commerce as commerce platforms that use online social media platforms. They demonstrated the advantages of ML techniques by contrasting lexicon- and ML-based approaches [9]. Recently, binary logistic regression was used to assess Twitter data for negative purchase intention. When compared to other cutting-edge machine learning techniques, the model created in this study had a higher F1 score [10]. A model for assessing user intention based on artificial intelligence was created by Sharma and Shafiq [11]. Based on reviews, the model demonstrated great precision, accuracy, and F1 score in identifying various online users' intents. This study suggests a deep learning model to forecast an internet user's intention to make a purchase. When the suggested model is used with data from Twitter, it exhibits good accuracy.

I. FLOWCHART

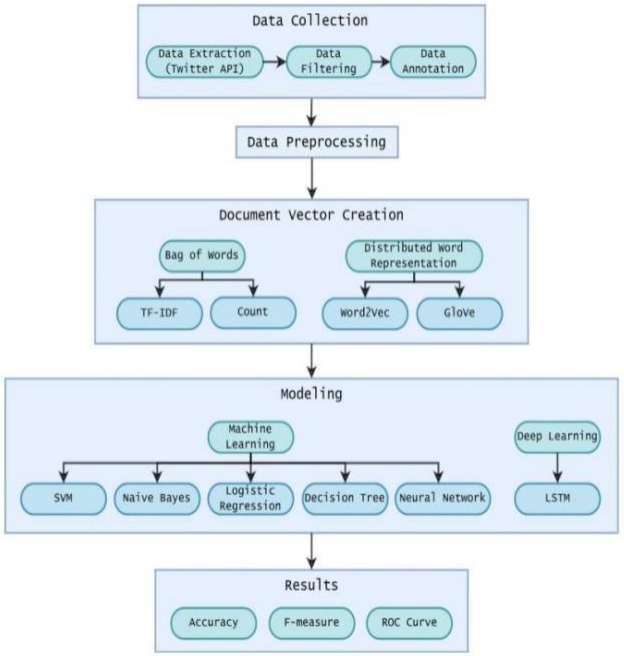

Figure 1 Flowchart

II. METHODOLOGY

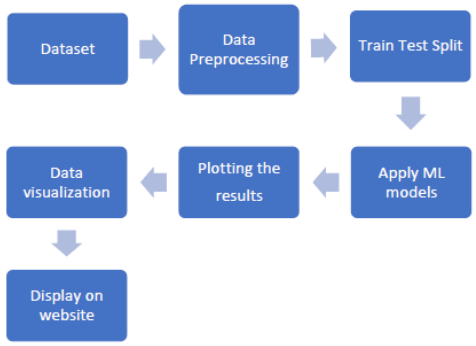

Figure 2 System Architecture

III. ALGORITHM

1.Dataset: We obtained the majority of our data from Twitter. Whether a review is positive or negative, people who tweet about a product via the Twitter app are considered.

2.Data Preprocessing: This step prepares the unprocessed data so that a machine learning model can use it. It is the first and most important step in the creation of a machine learning model. Cleaning and preparing data for a machine learning model are critical steps in the data preprocessing process, which improves the accuracy and efficiency of the machine learning model.

3.Train Test Split: After training and testing the data, split conditions are applied to various machine learning models.

4.Applying machine learning models: 1. Logistic regression. 2. The Tree of Decision 3. Ignorant Bayes 4. Vector Machine Support.

5.Plotting the results: The best algorithm with the highest level of accuracy is used to calculate the results.

6.Data Visualization: Data can be represented as two-dimensional rows and columns, pie diagrams, or histograms to aid in analysis and prediction.

7.Display on website: Configuring the website so that visitors can access relevant data about the consumer's anticipated intentions with regard to that specific product.

A. Data Preprocessing (Text Preprocessing)

1) LOWERCASE: To achieve case consistency, we began our foundational work by converting the text into lower case.

2) REMOVE PUNC: Next, we passed the lowercase text to functions that remove punctuation and special characters. Unwanted characters, spaces, tabs, and other elements that serve no purpose in text classification may appear in the text.

3) STOP WORDS REMOVAL: Text may contain words that are commonly used in sentences but serve no purpose or add to the meaning of the sentence. The terms "a," "an," and "in" are mentioned.

4) REMOVAL OF COMMON WORDS: Furthermore, the sentence contains a large number of words that are repeated but do not contribute to its meaning.

5) RARE WORDS REMOVAL: We also removed a few uncommon words, such as names and brand names that were not enclosed in HTML tags. These are the special terms that don't add much to the interpretation model.

6) SPELLING CORRECTIONS: Spelling mistakes abound in social media data. It is our responsibility to correct any errors and provide accurate words for our model.

7) STEMMING: Following that, we returned to the words' origins. Stemming words are those with missing ends or beginnings.

8) LEMMATIZATION: Next, we lemmatized our text. The order of this analysis is morphological. One can follow a word back to its lemma. After preprocessing, we are left with 1300 tweets to test and train our model.

B. Text Visualization

Text visualization refers to the process of extracting information from raw text and applying various analyses to identify significant structure and pattern. NLP tools can help with this. They can identify popular attitudes toward a particular subject or item (sentiment analysis). To visualize the subjectivity and sentiment polarity of the tweets in the dataset, we can use the Seaborn library.

Positive Tweets:

Figure 3 Positive Tweets

Negative Tweets:

Figure 4 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-

Logistic regression is one of the most widely used supervised learning algorithms. It predicts the categorical dependent variable based on a given set of independent variables.

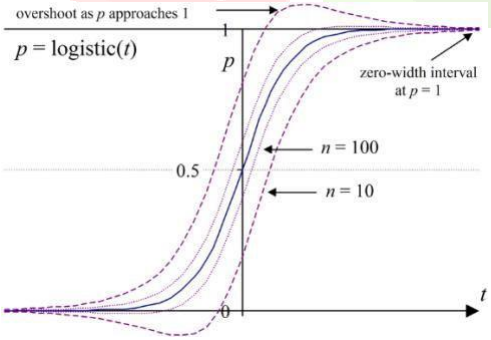

Figure 5 Logistic Regression

Decision Tree-

Although decision trees are a type of supervised learning technique, their primary application is to solve classification problems. However, they can also be used to resolve regression issues. It is a true-structured classifier, with internal nodes representing dataset features, branches for decision rules, and leaf nodes representing individual results.

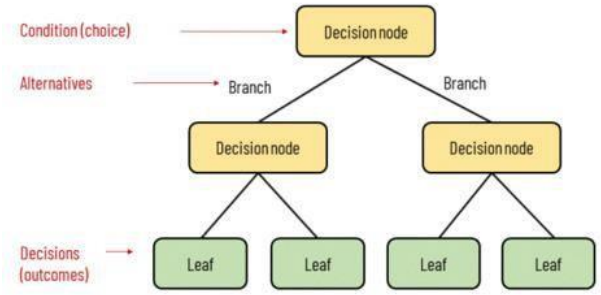

Figure 6 Decision Tree

SVM-

To make it easier to classify new data points in the future, the SVM algorithm seeks to identify the optimal line or decision boundary that divides an n-dimensional space into categories. We call this optimal decision boundary a hyperplane.

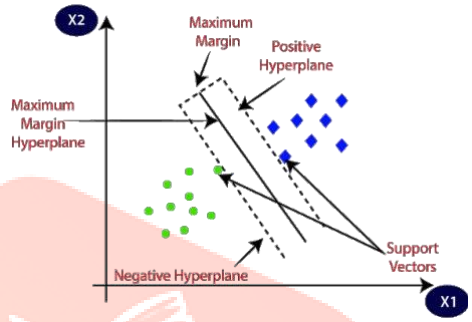

Figure 7 SVM

Naive Bayes-

The Naïve Bayes classifier is a simple and efficient classification algorithm that helps create machine learning models with fast prediction capabilities. As a probabilistic classifier, it predicts the likelihood that an object will occur.

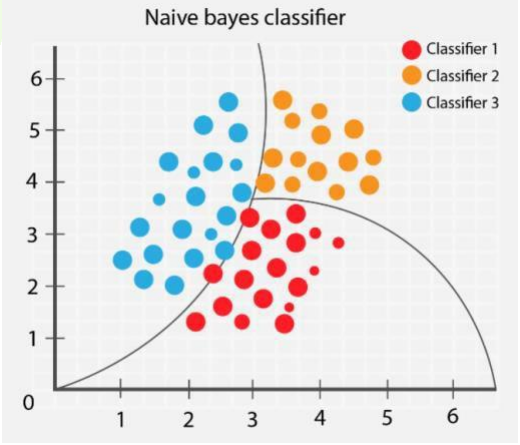

Figure 8 Naïve Bayes

Neural Network-

A method in artificial intelligence that teaches machines to manage information in a way that is similar to how the human brain functions. Propagation functions,

weights, biases, connections, propagation functions, and a learning rule make up neural networks.

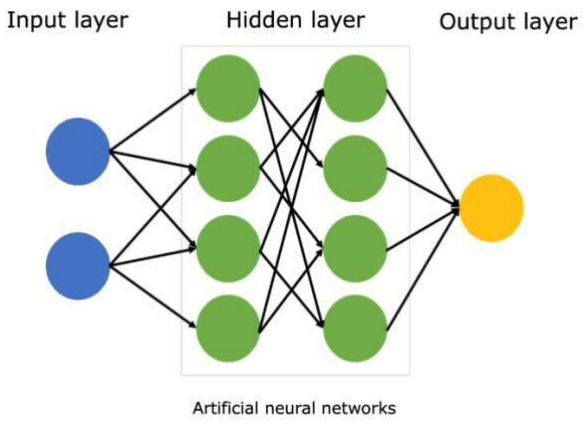

Figure 9 Neural Network

IV. EVALUATION FEATURES

Model evaluation is the process of applying different assessment metrics to examine the strengths and weaknesses and performance of a machine learning model. In order to evaluate our models, we use the following techniques:

1. Confusion Matrix: An evaluation tool for machine learning classification problems that can produce multiple classes as output. The table shows five distinct combinations of expected and actual values.

2. Precision: The number of accurate forecasts is the total count of forecasts. Accuracy for binary classification can also be computed in terms of positives and negatives: T P + T N + F P + F N = Accuracy. False Negatives (FN), False Positives (FP), True Negatives (TN), and True Positives (TP).

3. Precision and Recall: In machine learning, precision and recall are performance metrics used to classify and recognize patterns. These concepts are required for developing the ideal machine learning model that produces more precise and accurate results.

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

5. True-Negative Rate: A true positive is an outcome where the model correctly predicts the positive class. A true negative, on the other hand, is an outcome in which the model correctly predicts the negative class. False positives occur when the model incorrectly predicts the positive class.

V. ADVANTAGES AND DISADVANTAGES

A. Advantages :

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

2) Real-time insights: Our project can provide businesses with timely insights into their target audience's purchasing intentions, allowing them to make necessary changes to their marketing strategies.

3) Objective Analysis: By eliminating potential bias from manual analysis, machine learning algorithms provide a more objective assessment of purchase intention.

4) Predictive Power: By identifying patterns and trends in Twitter data, your model may be able to predict future purchase intentions, allowing businesses to proactively target potential customers.

5) Cost-effective: Once developed, the model can be applied to a large volume of data at a lower cost than hiring.

VI. FUTURE SCOPE

Take advantage of Twitter's real-time data streams to assist companies in responding swiftly to changing consumer attitudes and behaviours. One approach to achieve this could be to make use of state-of-the-art analytics and machine learning models that can adapt in real-time to new conversations and trends. Extend the analysis's domain beyond text-based tweets to include multimedia content like photos and videos. Textual and visual data combined provide a deeper understanding of customer behaviour and preferences. Be mindful of the tweets' context. A user's intent to buy can be strongly influenced by factors like location, weather, events, and user demographics [18].
Further research could focus on incorporating these contextual factors into prediction models. Analyse the ways in which information from different social media platforms can be merged to produce a comprehensive picture of consumer behaviour. Understanding how customers communicate and express themselves on different platforms can help to improve predictive accuracy. Investigate Twitter user actions and behaviours in addition to sentiment analysis. This could entail keeping an eye on how customers respond to specific product mentions, interact with brands, and determine whether or not these actions affect their inclination to purchase.

CONCLUSION

Since we tested four different models and chose the best model based on the product data, our project stands out when compared to other studies in the same field. The two problems that are mentioned below kept us from surpassing 80% accuracy. Achieving even 80% accuracy with a limited dataset and unbalanced class data is a noteworthy accomplishment.

The two primary problems we ran into were:
1) The unequal class problem: Because we annotated our dataset by hand, we had about 2000 positive tweets and 1200 negative tweets. Because of this, we were getting extremely low True Negative Rates and our model was failing to accurately predict the negative class.

2) Limited annotated data: Due to the time-consuming nature of manually annotating each tweet in the dataset, we were only able to annotate about 3200 tweets.

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."