Consumer Intention Prediction Using Twitter Data

Kaustubh Pawar, Karan Deshmukh, Rushikesh Sawant, Onkar Sonkule

Department of Computer Engineering, Zeal College of Engineering and Research Pune, Maharashtra, India

ABSTRACT

The ecommerce business, and specifically customers buying things online, has seen a considerable increase recently. There has been a lot of research into determining a user's purchase patterns and, more significantly, the factors that influence such patterns. may impact whether or not the consumer will purchase the product In this research, we'll look at be investigating whether it is possible to predict and buying intent of a user for a product and target that user with a message. A commercial or a bargain Furthermore, we want to create tools that will assist the Businesses estimate the number of prospective clients for their products. From their tweets and user profile data, they may determine their purchasing intent.

KEYWORDS: Consumer reviews, sentiment analysis, social perception score, linear regression analysis.

I. INTRODUCTION

We want to create a machine learning strategy that will identify potential clients for a product by evaluating purchase intent in terms of Twitter tweets. Because text analytics can be performed manually, it is in efficient, we must employ a text analytical machine learning model technique. Finding patterns and trends will be more faster and more efficient text mining and natural language processing technologies. In some ways, the process of detecting purchasing intentions is similar to that of finding wishes in product reviews.

After the introduction of brick-and-mortar businesses, the trend of online shopping has deservedly increased. E-retailers produ ced projected revenue of 1.9 trillion US dollars (7.4 percent of total retail sales) from 1.61 billion clients worldwide in 2016. Amazon, the world's largest online retailer, has over 310 million active user accounts, with about 136 billion dollars in sales in 2016. (Statista, 2017). In the first month after demonetization, digital payments grew by 271 percent in the world's third-largest purchasing power parity country (India), whereas cash on delivery fell by 30-40 percent (Chronicle, 2017). Marketing managers regularly and use purchase intentions as a factor in making decisions regarding new and existing products and services. Until now, many businesses have relied on client feedback. Our technique is complicated by the fact that we must cal culate how to measure the purchase.From a tweet's intentions Investigating and selecting several types of text analysis met ho ds. Choosing the ideal option will be a difficult process. Taking measurements of our machine's output. Learning models and then selecting

on the optimal one involves a number of elements that we shall discuss.

II. LITERATURE SURVEY

Several research studies have been conducted to examine the insights into online customer purchasing behavior. Only a few, however, have addressed client purchasing objectives. Studies on recognising wishes from text, particularly (Ramanand Bhavsar and Pednekar). Consider the challenge of identifying product review buy wishes. Among these desires are product recommendations or a desire to purchase the product to do so, they employed linguistic riles. There aware two types of desires Although rule-based ways to purchasing or recognising desires are available and are effective, coverage is inadequate, and they are difficult to extend. The problem of detecting purchase intentions is similar to that of finding wishes in product reviews. We provide a Machine instead of a rule-based method.

The sentiment 140API, which allows you to detect sentiment of a certain brand, product, or any subject on twitter, is one of the most frequent pre-processing approaches for twitter data. As part of the speech tagger, the Twitter NLP library .

LITERATURE REVIEW PIC BELOW

III. METHODOLOGY

When the corpus was complete, we tested various text analysis models to see which one produced the best results. The models we utilise are as follows:

A. Linear Regression:
When the dependent variable is binary, logistic regression is the proper regression strategy to use. The logistic regression, like other regression studies, is a predictive analysis. It's a term used to describe data.

B. Vectors Machines are supported:
SVM stands for supervised machine learning method that does not do complicated data transformations. Then it tries to split data into the designated classes.

C. Naive Byes:
The Bayes theorem is used with the "naive" assumption of conditional independence between any pair of characteristics given the features of the class variable.

D. Decision Tree:
A non-parametric supervised learning approach for classification and regression, decision trees are utilised. The objective is to learn basic decision rules from data attributes to develop a model that predicts the value of a target variable.

E. Neural Network:
A neural network is a deep learning machine method that is organised in layers of neurons. There are three levels of neurons: an input layer, an output layer, and buried layers. The neurons in these layers learn from the original input and subsequent runs, making the neuron network adaptable.

IV. TECHNOLOGY USED

A. NumPy:
Here, in this our project we use python module called NumPy for manipulating the data of our dataset and do some mathematical operation to calculate desired output from that using python NumPy module. NumPy is a general- purpose array-processing

package. It provides a high- performance multidimensional array object, and tools for working with these arrays. It's the abecedarian package for scientific computing with Python. In NumPy, number of confines of the array is called rank of the array. A tuple of integers giving the size of the array along each dimension is known as shape of the array. An array class in NumPy is called as and array. Rudiments in NumPy arrays are penetrated by using square classes and can be initialized by using nested Python Lists.

B. Panda:

We have used Panda module in this project to manipulate huge dataset with the functionality of pandas module effectively and we get desired dataset in form in which we want by manipulating raw dataset with the help of pandas. Pandas is an open-source library that's erected on top of NumPy library. Pandas is fast and it has high-performance productivity for druggies. Advantage Of Panda are as follow:

1. Fast and effective for manipulating and assaying data.
2. Data from different train objects can be loaded.
3. Easy running of missing data (represented as NaN) in floating point.

C. Matplotlib:

It is a python data visualization library or module. We have used this module in our project to visualize table format data into graph format, so that if there is any kind of error or inaccuracy so that we can easily get it out by data visualization as compared to data in table format and we able correct it immediately, so in directly it will increase our speed. Matplotlib is easy to use and an amazing imaging library in Python. It's erected on NumPy arrays and designed. to work with the broader SciPy mound and consists of several plots like line, bar, smatter, histogram, etc. Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multiplatform data visualization library erected on NumPy arrays and designed to work with the broader SciPy mound. It was introduced by John Hunter in the time 2002. One of the topmost benefits of visualization is that it allows us visual access to huge quantities of data in fluently digestible illustrations. Matplotlib consists of several plots like line, bar, smatter, histogram etc.

D. Flask:

we have used flask python module in our project, the purpose of this module is to connect backend machine learning module to CSS and Bootstrap and flask module will help to connect ML backend module to frontend to calculate estimated price. So here used need to give some basic information related to car purchasing and this information or input will pre-process in our machine and output or estimated car price will shown on website with the help of flask. Flask is an API of Python that allows us to build up web- applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web-Application. A Web-Application Framework or Web Framework is the collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc. Flask is based on WSGI (Web Server Gateway Interface) toolkit and Jinja2 tem MODELLING.

V. MODELLING

A. Analyzing:

In our project, we utilised the agile approach, which includes analysing as the first stage. We evaluated our dataset and the models we intended to use for our project throughout this part of the model to ensure that our results were as accurate as possible.

B. Planning:

In the second step of our model, the planning phase, we planned the tasks that our group members would carry out in order to finish the project within the time frame we set.

C. Design:

During the design process, we create the project's primary core components, such as the frontend and backend.

In order to use Machine Learning algorithms and produce projected outcomes, the two must be combined.

D. Testing:

We eliminated all bugs during this round. Check to determine if our project met all of the user's needs. Validation comes last after the prerequisite check.

E. Deployment:

For the time being, we've deployed our project on localhost 8000, but we can also deploy it on other platforms in the future, in order for all users to have access. The fundamental goal of implementing the Agile style of software elopement is to complete all tasks and processes at the same time. For our project, it also provides an incremental and iterative approach.

VI. SOFTWARE TESTING

Software testing can be stated as the process of verifying and validating that software or application is bug-free, meets the technical requirements as guided by its design and devel- opment, and meets the user require ments effectively and effi- ciently with handling all the exceptional and boundary cases. The process of software testing aims not only at finding faults in the existing software but also at finding measures to improve the software in terms of efficiency, accuracy, and usability. It mainly aims at measuring the specification, functionality, and performance of a software program or application.Software testing can be divided into two steps:

A. **Verification:** it refers to the set of tasks that ensure that software correctly implements a specific function. Verification: "Are we building the product, right?"
B. **Validation:** it refers to a different set of tasks that ensure that the software that has been built.

VII. ACKNOWLEDGEMENT

VIII. FUTURE WORK

This research's research study and literature evaluations revealed significant gaps, indicating that there are opportunities to expand this work. Incorporating geo-spatial data into sentiment analysis

may aid marketers in gaining insights into the reasons for a product's success or failure, as well as designing or altering the product to fit the preferences of the local population.

Apart from the potential for expansion, the work in this project provides abundant opportunity for study in adjacent areas that were not explored in this project or elsewhere and so provide rich ground for research.

There are few research on the impact of attitudes on product sales in the existing literature. Future research might focus on identifying other elements that function as predictors and obstacles.

IX. CONCLUSION

When compared to previous studies in the sector, our study stands out since we created five distinct models and then evaluated them before selecting the best one based on the product data because of the two issues listed below, we were unable to achieve greater than 80% accuracy. It's a success to reach even 80% accuracy with unbalanced class data and such a little dataset. We have two significant issues to deal with:

A. The issue of class imbalance: We have roughly 2000 positive tweets and 1200 negative tweets in our dataset since we manually tagged it. As a result, our True Negative Rate was extremely low, and our model was unable to effectively forecast the negative.

B. Limited annotated data: We were only able to annotate roughly 3200 tweets since we had to manually annotate each tweet in the dataset, which takes a long time.

X. REFERENCES

[1]. Speech and Language Processing (3rd ed. draft), Dan Jurafsky and James H. Martin.

[2]. Inspirations for code and designs:Building a prediction model, https://www.kaggle.com/gpayen/building-a-prediction model.

[3]. Sentiment analysis:https://www.kaggle.com/laowingkin/amazon-fine-food-reviewsentiment-analysis.

[4]. Inspirations for code and designs: TEXT PREPROCESSING USING PYTHON,https://www.kaggle.com/shashanksai/text-preprocessing-using-python.

[5]. Identifying Purchase Intentions by Extracting Information from Tweets, February 8,2017, RADBOUDU NIVERSITY NIJMEGEN, BACHELOR 'S THESIS In ARTIFICIAL INTELLIGENCE.

[6]. Aparna V.Mote, Utakarsha Musmade, Nikita Deshmukh "Modeling And Forecasting Gross Domestic Product (GDP) Using Linear Regression" 2019 JETIR May 2019, Volume 6, Issue 5.

[7]. Tweet analyst: Using Twitter Data to Analyze Consumer Decision Process, The Berkeley Institute of Design.