International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Finding Purchase Intentions using Twitter Data

Kunal Deshmukh

AISSMS IOIT, India

Abstract—

Recently, there has been a significant rise in the ecommerce industry and more specifically in people buying products online. There has been a lot of research being done on figuring out the buying patterns of a user and more importantly the factors which determine whether the user will buy the product or not. In this study, we will be researching on whether it is possible to identify and predict the purchase intention of a user for a product and target that user towards the product with a personalized advertisement or a deal. Further, we wish to develop a software that will help the businesses identify potential customers for their products by estimating their purchase intention in measurable terms from their tweets and user profile data on twitter. After applying various text analytical models to tweets data, we have found that it is indeed possible to predict if a user have shown purchase intention towards a product or not, and after doing some analysis we have found that people who had initially shown purchase intention towards the product have in most cases also bought the product.

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

I. INTRODUCTION

There have been several research studies for analyzing the insights of online consumers buying behavior. However, only a few have addressed the customers buying intention for products. We want to develop a machine learning approach that will identify potential customers for a product by estimating the purchase intention in measurable terms from tweets on twitter. We have used a text analytical machine learning approach because although text analytics can be performed manually, it is inefficient. By using text mining and natural language processing algorithms it will be much faster and efficient to find patterns and trends. In a way we can say that Purchase Intention detection task is close to the task of identifying wishes in product reviews.

II. LITERATURE REVIEW

There have been several research studies for analyzing the insights of online consumer buying behaviour. However ,only few have adressed the customer buying intentions for products. Studies on identification of wishes from text, specially (Ramanand Bhavsar and Pednekar) consider the task of identifying 'buy' wishes from the product review. These wishes include suggestions for a product or a desire to buy the product .They used linguistic riles to detect these two kind of wishes. Although rule based approaches for buying or identifying the wishes are effective, but their coverage is not satisfactory, and they can't be extended easily. Purchase intentions detection task is close to the task of identifying wishes in the product reviews. Here we don't use rule based approach, but we present a Machine Learning approach with generic features extracted from the Tweets.

Past studies have shown that it is possible to appear Natural Language Processing (NLP) and Named Entity Recognition (NER) to tweets. However, applying NER to tweets is very difficult because often people use the abbreviation or misspelled words and grammatical errors in the tweets .Nonetheless, Finn et al to name the annotated entities in tweets using crowd sourcing. Other studies apply the sentiment analysis to apply on the tweets. The first studies used product or movie review because the reviews are either positive or negative. Wang et al Anta et al analyzed the sentiments of the tweets filtered on a certain hashtag (keywords or phrases starting with a symbol that denate the main topic of the tweet). These studies merely analyze the sentiment of a tweet about the product after the author has bought it.

Some pre processing techniques commonly used for twitter data are the sentiment 140API that allows you to discover sentiment of a particular brand, product or any topic on twitter. Twitter NLP library tokenizer as a part of speech tagger. hirarchical word structures or clusters, and a dependency parser for the tweets.

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.

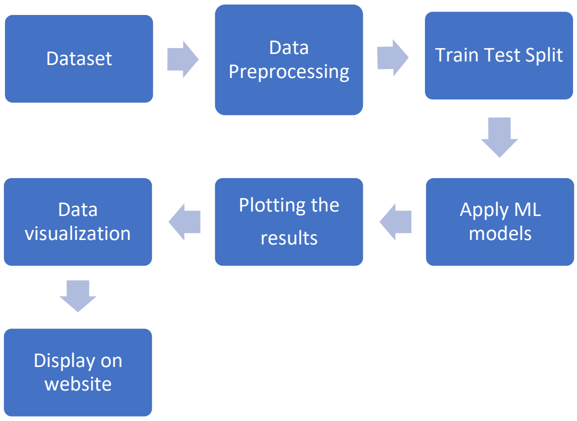
Shital Anil Phand, Jeevan Anil Phand , Twitter sentiment classification using Stanford NLP, 1st International Conference on Intelligent Systems and Information Management (ICISIM), IEEE, 2017, pp. 1-5.

Lokesh Mandloi, Ruchi Patel, Twitter Sentiments Analysis Using Machine Learning Methods, International Conference for Emerging Technology (INCET), IEEE, 2020, pp. 1-5.

J. Ramanand, Krishna Bhavsar, Niranjan Pedanekar, Wishful thinking: finding suggestions and 'buy' wishes from product reviews, Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, pp. 54-61.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, Mark Dredze, Annotating Named Entities in Twitter Data with Crowdsourcing, Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, ACM, 2010, pp. 80-88

III. METHODOLOGY



A. Steps of Machine Learning Modelling

1. Dataset- Our dataset is mainly based on twitter. People who use Twitter application to tweet their reviews regarding a particular product is taken into consideration irrespective of weather the reviews are positive or negative.
2. Data Preprocessing- Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.
3. Train Test Split- Train the data, testing the data and finally applying split conditions which are further given to apply the different ML models.
4. Application of ML models - 1.Logistic regression 2.Decision Tree 3.Naive Bayes 4.Support Vector Machine
5. Plotting the results - Results are calculated taken into consideration the best possible algorithm into consideration which gives the maximum highest accuracy.
6. Data Visualization - Visualization of the data in form of 2 dimensional rows and columns, pie diagrams, Histograms for efficient analysis and prediction of results.
7. Display on website - Deploying the website so that people can visit the website for getting relevant information about the prediction of intentions of consumer regarding the particular product.

B. Data Preprocessing (Text Preprocessing)

1) LOWERCASE: So we started our groundwork by converting text into lower case, to get case uniformity.

2) REMOVE PUNC: Then we passed that lowercase text to punctuations and special characters removal functions. Text may contain unwanted character spaces tabs and etcetera which has no significance use in text classification.

3) STOPWORDSREMOVAL: Text may contain words that are not useful but still a part of sentence as a routine and do not contribute to meaning of sentence. Like 'a", "an", "in" are the words mentioned.

4) COMMON WORD REMOVAL: Then there also lots of repetitive words which form their recurrence do not contribute to the meaning of the sentence.

5) RARE WORDS REMOVAL: We also removed some rare words like name, brand words left out html tags etc. These are the unique words that do not contribute much to the interpretation model.

6) SPELLING CORRECTION: Social media data is full of spelling mistakes. And it's our job to get rid of the mistakes give our model correct word as input.

7) STEMMING: Then we stemmed the words to their roots. Stemming words are the words by cutting the end or beginning of the word.

8) LEMMATIZATION: Then we also performed lemmatization on our text. This analysis is performed in morphological order. A word is traced back to its lemma.

After preprocessing of the tweets we are left with the 1300 tweets to train and test our model.

C. Text Visualization

Text visualization is the process of extracting information from the raw text and perform different analysis to derive important structure and pattern out of the information. This is where natural language processing (NLP) tools come in. They can capture prevailing moods about a particular topic or product (sentiment analysis).

For this we can use seaborn library to visualize the sentiment polarity as well as subjectivity of tweets in dataset

Positive Tweets:

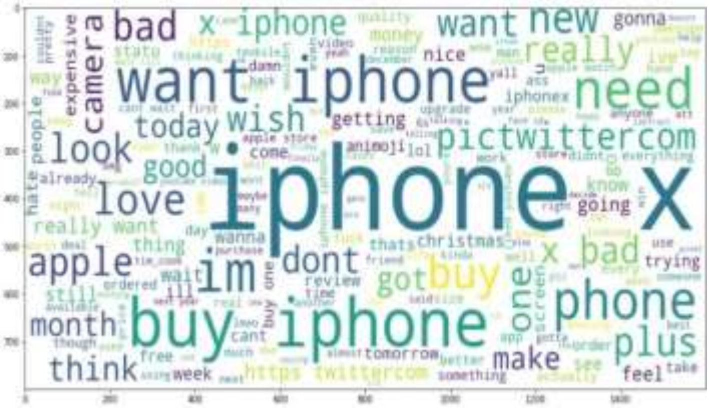
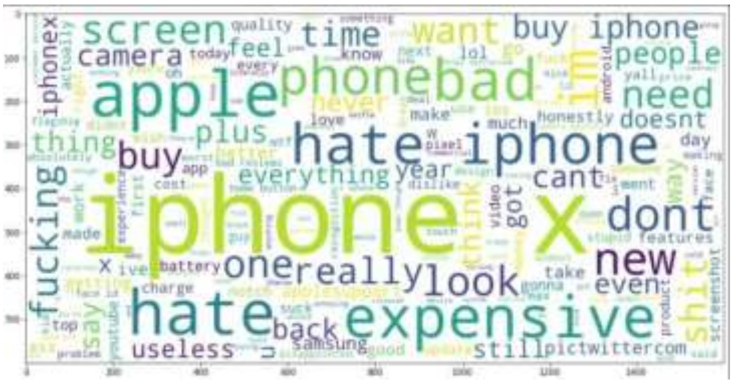


Fig. 4.1

Negative Tweets:

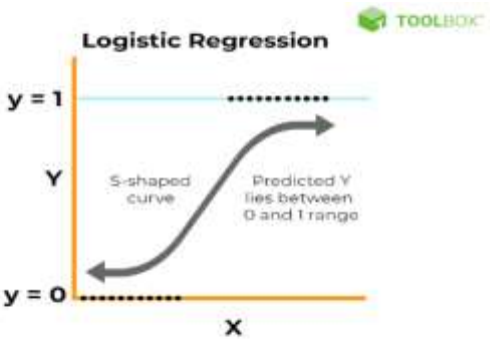


D. ML Models

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

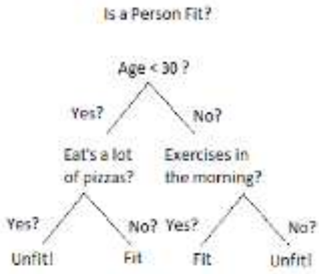
Logistic Regression-

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.



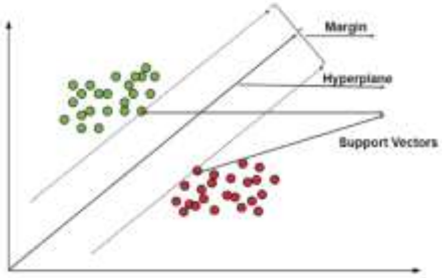
Decision Tree-

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems but mostly it is preferred for solving Classification problems. It is a true-structured classifier, where internal nodes represents the features of the dataset, branches represent the decision rules and each leaf node represents the outcome.



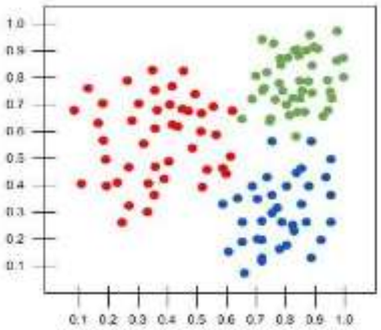
SVM-

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.



Naive Bayes-

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.



IV. Evaluation Features

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses.

To evaluate our models, we use the following techniques:

1. Confusion Matrix: A performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

2. Accuracy: Accuracy = Number of correct predictions Total number of predictions. For binary classification, accuracy can also be calculated in terms of positives and negatives as follows: Accuracy = T P + T N T P + T N + F P + F N. Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives

3. Precision and Recall: Precision and recall are performance metrics used for pattern recognition and classification in machine learning**.** *These concepts a*re essential to build a perfect machine learning model which gives more precise and accurate results.

4. F-measure: The F-measures do not take true negatives into account, hence measures such as the Matthews correlation coefficient, Informedness or Cohen's kappa may be preferred to assess the performance of a binary classifier.

5. True-Negative Rate: A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class. A false positive is an outcome where the model incorrectly predicts the positive class.

V. CONCLUSION AND FUTURE SCOPE

Looking at the other researches that are done in the similar field, our project also stands apart since we have implemented 4 different models and after evaluating them, we choose the best one customized to the product data.

We were not able to get more than 80% accuracy because of the two problems highlighted below. To achieve even 80% accuracy with an imbalance class data and such a small dataset is a victory.

The 2 major problems that we faced were:

1) The imbalance class problem: Since our dataset was manually annotated by us, we had about 2000 positive tweets and 1200 negative tweets. Due to this we were getting a very low True Negative Rate and our model was not accurately predicting the negative class.

2) Limited annotated data: Since we had to manual annotate each tweet in the dataset and this process takes a lot of time, we were only able to annotate about 3200 tweets.

VI. 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 tweet sentiment derived from tweet segmentation using Hadoop, 1st International Conference on Intelligent Systems and Information Management (ICISIM), IEEE, 2017,pp. 194 - 198.

[4] Lokesh Mandloi, Ruchi Patel, Twitter Sentiments Analysis Using Machine Learning Methods, International Conference for Emerging Technology (INCET), IEEE, 2020, pp. 1-5.

[5] J. Ramanand, Krishna Bhavsar, Niranjan Pedanekar, Wishful thinking: finding suggestions and 'buy' wishes from product reviews, Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, pp. 54-61.

[6] J. Ramanand, Krishna Bhavsar, Niranjan Pedanekar, Wishful thinking: finding suggestions and 'buy' wishes from product reviews, Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, pp. 54-61.

[7] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, Mark Dredze, Annotating Named Entities in Twitter Data with Crowdsourcing, Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, ACM, 2010, pp. 80-88