Finding Purchase Intention Using Twitter Data

Supervisor: Mr Jhalak Dutta

Author: Rishav Giri

Co-Authors: Sarasij Jana, Shouvik Chatterjee

Table of Contents

[Abstract 3](#_Toc11663125)

[Introduction 3](#_Toc11663126)

[Literary Review 3](#_Toc11663127)

[Proposed Approach 4](#_Toc11663128)

[Data collection and annotation 4](#_Toc11663129)

[Data preprocessing 5](#_Toc11663130)

[Data preprocessing techniques: 5](#_Toc11663131)

[Document Vector 5](#_Toc11663132)

[Modelling 6](#_Toc11663133)

[Evaluation: 6](#_Toc11663134)

[Conclusion: 7](#_Toc11663135)

[References: 8](#_Toc11663136)

# 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.

# 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.

# Literary Review

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. Studies on identification of wishes from texts, specifically Ramanand et al. (Ramanand, Bhavsar, and Pedanekar 2010) consider the task of identifying ‘buy’ wishes from product reviews. These wishes include suggestions for a product or a desire to buy a product. They used linguistic rules to detect these two kinds of wishes. Although rule-based approaches for identifying the wishes are effective, but their coverage is not satisfactory, and they can’t be extended easily. Purchase Intention detection task is close to the task of identifying wishes in product reviews. Here we don’t use the 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 apply Natural Language Processing (NLP) and Named Entity Recognition (NER) to tweets (Li et al., 2012) (Liu et al., 2011). However, applying NER to tweets is very difficult because people often use abbreviations or (deliberate) misspelled words and grammatical errors in tweets. Nonetheless, Finin et al. (2010) tried to annotate named entities in tweets using crowdsourcing. Other studies used these techniques to apply sentiment analysis to tweets. The first studies used product or movie reviews because these reviews are either positive or negative. Wang et al. (2011) and Anta et al. (2013) analyzed the sentiment of tweets filtered on a certain hashtag (keywords or phrases starting with the symbol that denote the main topic of a tweet). These studies merely analyze the sentiment of a tweet about a product after the author has bought it. We will however be extracting features from tweets to find whether the user has shown purchase intention towards the product or not.

More recently, research articles like *Identifying Purchase Intentions by Extracting Information from Tweets* ( February 8, 2017, RADBOUD U NIVERSITY NIJMEGEN) and *Tweetalyst: Using Twitter Data to Analyze Consumer Decision Process* (The Berkeley Institute of Design) investigate if an artificial intelligence approach can predict (from existing user created content on twitter) if someone is a potential customer for a specific company or product and identify users at different stages of the decision process of buying a given product. Further looking at research reports like *The Impact of Social Network Marketing on Consumer Purchase Intention in Pakistan: Consumer Engagement as a Mediator* (Asian Journal of Business and Accounting 10(1), 2017) give us an insight of the impact of social network marketing on consumer purchase intention and how it is affected by the mediating role of consumer engagement. Based on UGT theory (Uses and Gratification Theory).

Some preprocessing techniques commanly used for twitter data are the sentiment140 API (Sentiment140 allows you to discover the sentiment of a brand, product, or topic on Twitter), the TweetNLP library (a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets), unigrams, bigrams and stemming. There are also some dictionary-based approaches such as using the textBlob library (TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more).

The common machine learning algorithms that are used for text analysis are Linear Regression, Random Forest, Naive Bayes and Support Vector Machine. We will be looking at these models later in detail.

# Proposed Approach

In this section, we describe the details of our approach to tackle the problem of purchase intention detection. We will begin by describing our data collection and annotation process. Then we will describe our approach for data preprocessing and transforming the data to train text analytical models.

## Data collection and annotation

As there are no annotated Twitter tweets corpora available publicly for detection of purchase intent, we had to create our own. This was done using a web crawler developed by JohnBakerFish which crawled the website to collect the data. We had collected over 100,000 tweets but since they were not annotated, we had to cut down to just 3200 tweets which were randomly selected out of the dataset and we manually annotated them using a basic criterion we had defined:

Criteria for Labelling of tweets

|  |  |  |
| --- | --- | --- |
|  | Tweet | Class |
| 1 | Comparing iphone x with other phone and telling other phone are better? | No PI |
| 2 | Talking about good features of iphone x? | PI |
| 3 | Talking about negative features of iphone x? | No PI |
| 4 | liked video on Youtube about iphone x? | PI |

We used just 3200 tweets out of such a large dataset as we were limited by time. We defined definition of Purchase Intention as object that is having action word like (buy, want, desire) associated with it. Each tweet was read by 3 people and final class was decided by maximum voting.

## Data preprocessing

### Data preprocessing techniques:

Next, we preprocessed the tweets using these techniques:

1. LOWERCASE: So, we started our groundwork by converting our text into lower case, to get case uniformity.
2. REMOVE PUNC: Then we passed that lower case text to punctuations and special characters removal function. Text may contain unwanted special characters, spaces, tabs and etcetera which has no significant use in text classification.
3. STOPWORDS REMOVAL: Text also contains useless words which are routine part of the sentence and grammar but do not contribute to the meaning of the sentence. Likes of “the”, “a”, “an”, “in” and etcetera are the words mentioned above. So, we do not need these words, and it is better to remove these.
4. COMMON WORD REMOVAL: Then there also lots of repetitive words which from their recurrence do not contribute to the meaning in the sentence. This can also be the result of mistake as the data we are analyzing is an informal data where formal sentence norms are not taken into consideration.
5. RARE WORDS REMOVAL: We also removed some rare words like names, brand words (not iphone x), left out html tags etc. These are unique words which do not contribute much to interpretation in the model.
6. SPELLING CORRECTION: Social media data is full of spelling mistakes. And it is our job to get rid of these mistakes and give our model the correct word as an input.
7. STEMMING: Then we stemmed the words to their root. Stemming works like by cutting the end or beginning of the word, considering the common prefixes or suffixes that can be found in that word. For our purpose, we used Porters Stemmer, which is available with NLTK.
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, and lemma is returned as the output.

After preprocessing the tweets, we are left with about 1300 tweets for training data and remaining for testing.

### Document Vector

Next, we made 3 types of document vectors:

1. TF: First is the term frequency document vector. We have stored text and its labeled class in data frame. And we have constructed a new data frame with columns as the words and document count as the rows. So, individual frequency of words in a document count is recorded.
2. IDF: It is a weighting method to retrieve information from the document. Term frequency and inverse document frequency scores calculated and then product of TF\*IDF is called TF-IDF. IDF is important in finding how relevant a word is. Normally words like ‘is’, ‘the’, ‘and’ etc. have greater TF. So IDF calculated a weight to tell how important least occurring words are.
3. TF-IDF with textblob library: With the help of textblob library we calculated sentiments of individual word and then multiplied the sentiment score with TF and TF-IDF of that word.

## Modelling

Once the corpus was ready, we then used different text analytical models to test which one gave the best results. We used the following models:

1. Support Vector Machine (SVM): Simply put, SVM is a supervised machine learning algorithm which does complex transformation on the data. And then it tries to separate data on classes we have defined on our data.
2. Naive Bayes: Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable.
3. Logistic Regression: Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
4. Decision Tree: Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
5. Neural Network: It is deep learning machine algorithm, which is arranged in a layer of neuron. There is an input layer, output layer and hidden layers of neurons. Neuron network is adaptive as neurons in these layers learn from their initial input and subsequent runs.

# Evaluation:

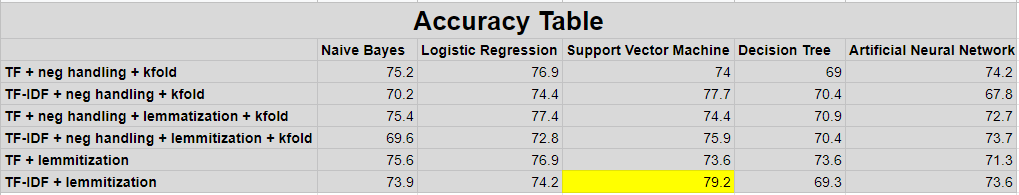
To evaluate our models, we used the following techniques:

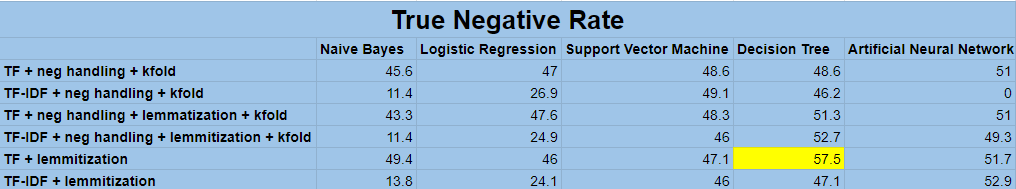
1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall
5. F-Measure
6. True Negative Rate

Further we have also considered The True Positive Rate and the shape of the ROC curve for more insights.

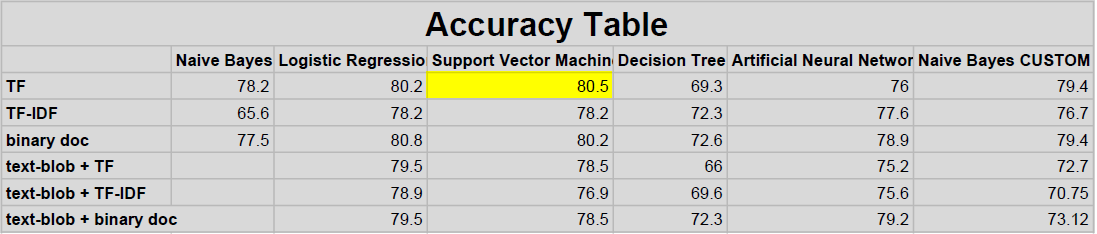
After evaluating our model here are the following results that we have gotten:

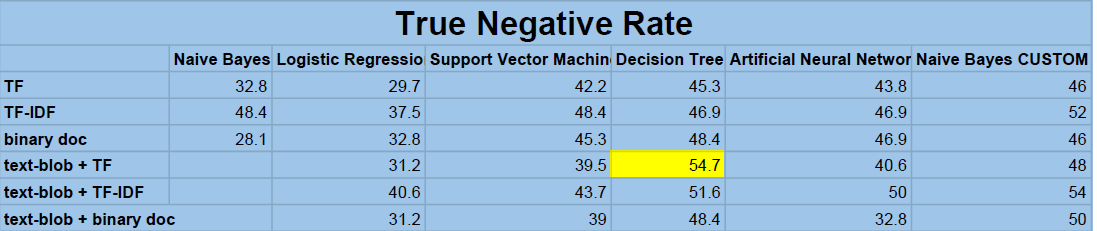
For our first attempt this is the results that we got:

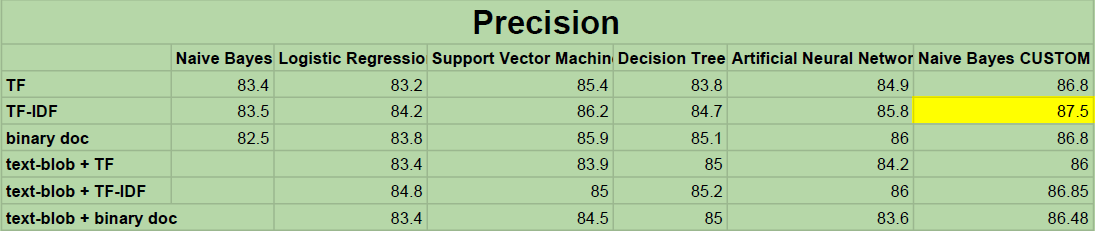


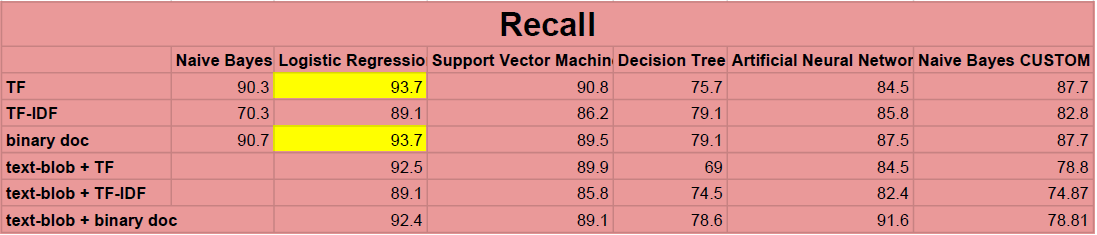


For our second attempt after reorganizing the data preprocessing steps and adding code for negation handling, we got these results:









# Conclusion:

Our results were quite promising since we had created our own dataset and were building the model from scratch. We had to create our own dataset because there does not exist a publicly available dataset for purchase intention based on twitter tweets.

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.

Looking at the other researches that are done in the similar field, our project also stands apart since we have implemented 5 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 above. To achieve even 80% accuracy with an imbalance class data and such a small dataset is a victory.

# References:

1. Books:
   1. Speech and Language Processing (3rd ed. draft), Dan Jurafsky and James H. Martin.
2. Inspirations for code and designs:
   1. Building a prediction model, https://www.kaggle.com/gpayen/building-a-prediction-model
   2. Sentiment analysis, https://www.kaggle.com/laowingkin/amazon-fine-food-review-sentiment-analysis.
   3. TEXT PREPROCESSING USING PYTHON, https://www.kaggle.com/shashanksai/text-preprocessing-using-python.
3. Relevant Papers:
   1. Identifying Purchase Intentions by Extracting Information from Tweets, February 8, 2017, RADBOUD U NIVERSITY NIJMEGEN, BACHELOR ’S THESIS IN ARTIFICIAL INTELLIGENCE.
   2. Tweetalyst: Using Twitter Data to Analyze Consumer Decision Process, The Berkeley Institute of Design.
   3. The Impact of Social Network Marketing on Consumer Purchase Intention in Pakistan: Consumer Engagement as a Mediator, Asian Journal of Business and Accounting 10(1), 2017.
   4. Using Twitter Data to Infer Personal Values of Japanese Consumers, 29th Pacific Asia Conference on Language, Information and Computation pages 480 – 487 Shanghai, China, October 30 - November 1, 2015, Copyright 2015 by Yinjun Hu and Yasuo Tanida.
4. Websites:
   1. https://www.kaggle.com/snap/amazon-fine-food-reviews
   2. https://scikit-learn.org/stable/