Predicting Consumer Purchase Intentions Using Twitter Data

Raghav A[1], Jatin M Bhangaonkar[2], Aruna Kamble[3]

[1, 2, 3] Dept of Computer Engineering

[1, 2, 3] Bharati Vidyapeeth College of Engineering, Sector 07, C.B.D.Belapur , Navi Mumbai, 400614, India.

Abstract- *Significant rise in the ecommerce industry and more specifically in people buying products online has been observed during the recent COVID times. Lot of research has been done on figuring out the patterns of a user and the probablility of them buying a product. In this study, we will be researching on whether it is possible to identify and predict the purchase intention of a user for a desired product and this be used to target product specified ads towards the user. 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. By applying various text 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- Data Mining, Information Retrieval, Machine Learning, Artificial Intelligence, Python.

I. INTRODUCTION

There have been several research studies for analyzing the consumer purchase behavior of online buying . However, only a few have addressed the customers buying intention for specific products. We propose to develop a machine learning web based app that will identify potential customers for a product by estimating the purchase intention inanalytical terms from tweets on twitter. We have used a text mining machine learning approach because although text analytics can be performed manually, it is inefficient. Text mining and natural language processing algorithms will make it 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.Different algorithms have been tested to observe efficiencies.

II. LITERATURE 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(intent) 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 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 of abbrevistions, misspelled words and grammatical mistakes. Nonetheless, Finin et al. (2010) tried to explain named entities in tweets using crowdsourcing. 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 merely analyze the sentiment of a tweet about a product after the author has bought it or general sentiments reguarding the specific product. 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. Such sentiments can be used for political analysis or most loved movie genre and much more. 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. Dictionary based TextBlob library used for processing text also exist. It provides a consistent API for diving into common NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis etc.

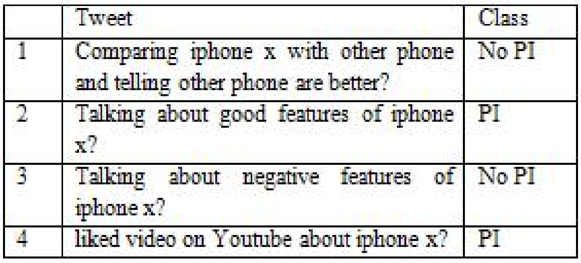
Some common machine learning algorithms that are used for text analysis are Linear Regression, Random Forest, Naive Bayes and Support Vector Machine.

III. 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 using the data to train text analytical models.

A. 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 10,000 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:(table 1)



	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 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.

B. 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. The unwanted special characters, spaces, tabs and etcetera which has no significant use in text classification were removed.
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.
4. COMMON WORD REMOVAL: Repeatative words might also occur in the sentences.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: Spelling errors are abundant on Social Media. 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. It reduced words to its word stem and is used vastly in Natural Language Understanding(NLU). For our purpose, we used Porters Stemmer, which is available with NLTK.
8. LEMMATIZATION: The next step was to perfomlemmatization.This is analyzed inmorphological order and inflected form of words are grouped together. A word is traced back to its lemma, and lemma is returned as the output.

IV. EVALUATION

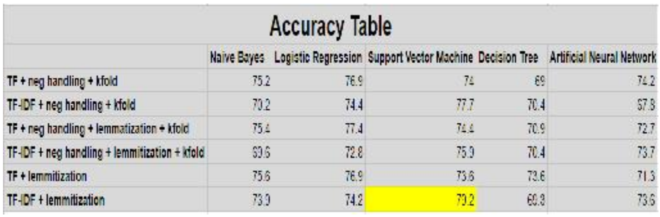
To evaluate our models, we used the following techniques:

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

The True Positive Rate and the shape of the ROC curve were considered for more insights.

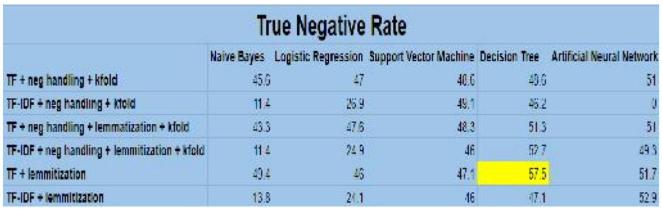
After evaluating our model,we have obtained following results:

1st test:



Accuracy Table

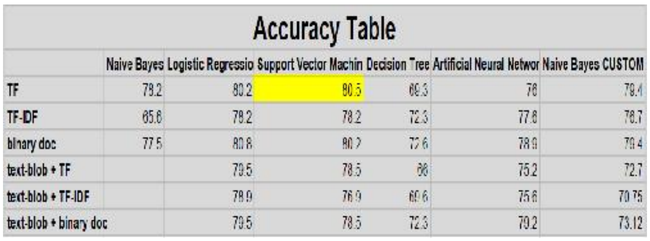
	Naive Bayes	Logistic Regression	Support Vector Machine	Decision Tree	Artificial Neural Network
TF + neg handling + kfold	75.2	76.9	74	69	74.2
TF-IDF + neg handling + kfold	70.2	74.4	77.7	70.4	67.8
TF + neg handling + lemmatization + kfold	75.4	77.4	74.4	70.9	72.7
TF-IDF + neg handling + lemmitization + kfold	69.6	72.8	75.9	70.4	73.7
TF + lemmitization	75.6	76.9	73.6	73.6	71.3
TF-IDF + lemmitization	73.9	74.2	79.2	69.3	73.6



True Negative Rate

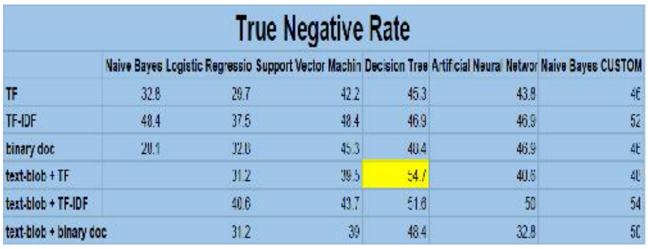
	Naive Bayes	Logistic Regression	Support Vector Machine	Decision Tree	Artificial Neural Network
TF + neg handling + kfold	45.6	47	40.6	40.6	51
TF-IDF + neg handling + kfold	11.4	26.9	49.1	46.2	0
TF + neg handling + lemmatization + kfold	43.3	47.6	48.3	51.3	51
TF-IDF + neg handling + lemmitization + kfold	11.4	24.9	46	52.7	49.3
TF + lemmitization	40.4	46	47.1	57.5	51.7
TF-IDF + lemmitization	13.8	24.1	46	47.1	52.9

For our second attempt after reorganizing the data preprocessing steps and adding code for negation



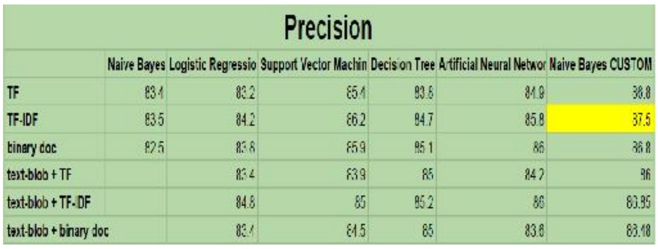
Accuracy Table

	Naive Bayes	Logistic Regressio	Support Vector Machin	Decision Tree	Artificial Neural Networ	Naive Bayes CUSTOM
TF	78.2	80.2	80.5	69.3	78	79.4
TF-IDF	65.6	78.2	78.2	72.3	77.8	78.7
binary doc	77.5	80.8	80.2	73.6	78.9	79.4
text-blob + TF		79.5	78.5	68	75.2	72.7
text-blob + TF-IDF		78.9	76.9	69.6	75.6	70.75
text-blob + binary doc		79.5	78.5	72.3	79.2	73.12



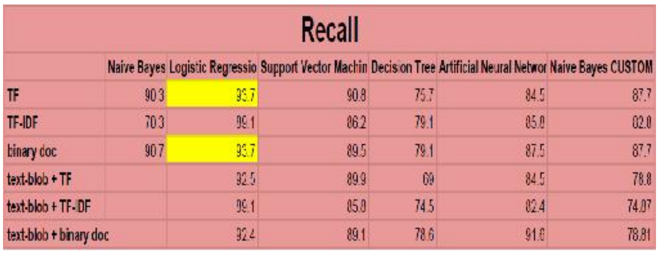
True Negative Rate

	Naive Bayes	Logistic Regressio	Support Vector Machin	Decision Tree	Artificial Neural Networ	Naive Bayes CUSTOM
TF	32.8	29.7	42.2	45.3	43.8	46
TF-IDF	48.4	37.5	48.4	46.9	46.9	52
binary doc	20.1	32.8	45.3	40.4	46.9	46
text-blob + TF		31.2	39.5	54.7	40.6	40
text-blob + TF-IDF		40.6	43.7	51.8	50	54
text-blob + binary doc		31.2	39	48.4	32.8	50



Precision

	Naive Bayes	Logistic Regressio	Support Vector Machin	Decision Tree	Artificial Neural Networ	Naive Bayes CUSTOM
TF	83.4	83.2	85.4	83.6	84.9	36.8
TF-IDF	83.5	84.2	86.2	84.7	85.8	37.5
binary doc	82.5	83.8	85.9	85.1	86	36.8
text-blob + TF		83.4	83.9	85	84.7	86
text-blob + TF-IDF		84.8	85	85.2	86	83.85
text-blob + binary doc		83.4	84.5	85	83.8	83.18



Recall

	Naive Bayes	Logistic Regressio	Support Vector Machin	Decision Tree	Artificial Neural Networ	Naive Bayes CUSTOM
TF	90.3	93.7	90.8	75.7	84.5	87.7
TF-IDF	70.3	99.1	86.2	79.1	85.0	82.0
binary doc	90.7	93.7	89.5	79.1	87.5	87.7
text-blob + TF		92.5	89.9	69	84.5	78.8
text-blob + TF-IDF		99.1	85.0	74.5	82.4	74.07
text-blob + binary doc		92.4	89.1	78.6	91.6	78.81

All these are form the confusion Matrix which will help us determine the probability of the consumer buying the product.

VI. CONCLUSIONS

The results were observed to be falling a little short of perfectly accurate but were promising enough to pursue this project. 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 a very low True Negative Rate were observed and the negative classification was affected.
2. Limited annotated data: Since we had to manually annotate each tweet in the dataset and this was time inefficient.

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.

The accuracy was affected due to the problems mentioned above.But,to achieve even 80% accuracy with an imbalance class data and such a small dataset is a win.

REFERENCES

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

[2] Building a prediction model, https://www.kaggle.com/gpayen/building-a-prediction-model

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

[4] 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, RADBOUD U NIVERSITY NIJMEGEN, BACHELOR 'S THESIS IN ARTIFICIAL INTELLIGENCE.

[6] Tweetalyst: Using Twitter Data to Analyze Consumer Decision Process, The Berkeley Institute of Design.

[7] 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.

[8] 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.

[9] https://www.kaggle.com/snap/amazon-fine-food-reviews

[10] https://scikit-learn.org/stable/