IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.7, Special Issue.3, Feb. 2019 E-ISSN: 2347-2693

Feature Selection and Classification for Sentiment Analysisof Amazon Product Reviews

Smita Suresh Daniel [1*], Ani Thomas [2],Neelam Sahu[3]

[1,3]Dept of ComputerScience ,Dr. C.V. Raman University Kota ,C.G ,India
[2]Dept of Information Technology , BITDurg , CSVT University , Bhilai, C.G, India

[*]*Corresponding Author: smitasuresh01@redifmail.com, Tel.: +00-9826108928*

Available online at: www.ijcseonline.org

Abstract—Online reviews provide accessible and plentiful data for relatively easy analysis for a given product.This paper seeks to apply and extend the current work in the field of Natural Language processing and sentiment analysis to retrieve information from Amazon Product reviews classify them using Naïve bayes classifier . This work presents a methodology that shows how text data can provide insight into various features of a product found in the customer reviews and feature selection method.

Keywords—Feature selection , Sentiment classification, Categorization

I. INTRODUCTION

Sentiment Analysis is a current research area in text mining. It is an important source of decision making which is used to extract, and identify product reviews. However, the growing scale of data demands automatic data analysis techniques .Information Extraction aims to obtain writer's feelings expressed in positive or negative comments. Byanalyzing a large numbers of documents, itattempts to identify the opinion sentiment that hold towards an object. It makes use of natural language processing (NLP) and computational technique to automate the extraction or classification of sentiment from typically unstructured text.

Amazon is one of the largest online vendor in the World. We focused on understanding issues and problems faced by the customers on buying a product. Our main goal is to collect the online Amazon reviews for a product , extract the most important features of the product talked about by the reviewers in a given datatset and then quantify the polarity of these reviews as positive or negative for each of the features of the product.

The proposed system is implemented using Python 2.7 .

The summary of this paper is as given.

- The extracted raw data are preprocessed using Natural Language Toolkit techniques .
- Extract the important features of the product .
- Naive Bayes classifier is used for training and testing . Feature selection and also evaluatingthe sentimental polarity using 1000 review for the product Kindle Store.
- In order to select the best features, frequency distribution is used .Based on these results, we implemented a multi-label classification algorithm to categorize them.
- The sentiments of each category was evaluated using polarity algorithm.

II. RELATED WORK

Sentiment Analysis is the most important research area in various various fields like political, educational, business etc. Pang et al. has first carried out the sentiment classification in different areas of product review using star ratings as polarity targets .A model for classification of reviews along with sentimental analysis is dicussed in paper [6]. Some of the Machine learning Techniques like Naïve Bayes ,Maximum Entropy and Support Vector Machines has been discussed in paper [1].

III. METHODOLOGY

This section explains the methodology of the proposed study as presented in Figure 1. First, the review documents were collected and pre-processed with basic natural language processing techniques like word tokenization, stop word removal, POS taggingand stemming. The residual tokens were arranged as per their frequencies or occurrences in whole documents set. Then feature selection method is utilized to pick out top n-ranked attributes of the product . Then these features are used for categorizing the reviews based on the the attributes, Then we trained the Naïve Bayes Classifier and sentiment based classification was done.Sentiment analysis is to classify these reviews

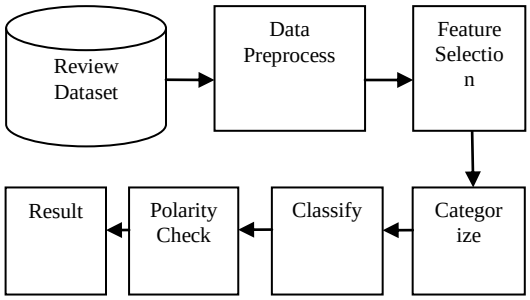

Figure 1. Flow Diagram Of The System

IV. DATA PREPROCESSING

A. Data Acquisition

For our study purpose we obtained our dataset fromthe Amazon site for reviews of the product Kindle which was meant for our study purpose. It is a textfile.It contained columns for information about the review such as the product ID, rating, review date, and review text etc. but to keep it less complicated we collected the review text part for data analysis . There was a lot of irregularity and diversity of the language used with much noise and its a challenging task to use it without a content analysis.

A total of 1000 reviews were collected as data set and where 400 positive and 400 negative reviews were used as training set 200 for testing.

B. Data Pre-Processing

Each document contained a large amount of informal language, sarcasm, acronyms, and mis-spellings, and meaning is often ambiguous and subject to human interpretation. Faulty assumptions are likely to arise if automatic algorithms are used without taking a qualitative look at the data.The text pre-processing is the process of refining the review which is in the form of text where several unwanted words in the review or a part of text such that provide no meaning to the full sentences[3].

From each of the review texts all punctuation except periods, apostrophes, and hyphens were removed .Users sometimes repeat letters in words so that to emphasize the words, for example, 'greeeeat " ,"tooogooood". Besides, common stopwords such as "a, an, and, of, he, she, it", non-letter symbols, and punctuation also bring noise to the text. So we pre-processed the texts before training the classifier:

1. We removed all the unwanted text from the reviews like repeated letters, stopwords , non-letter symbols from the text texts.
2. Negative words are useful for detecting negative emotion and issues. So we substituted words ending with "n't" and other common negative words (e.g.not ,no,nothing, never, none, cannot) with "negkok".
3. We removed all words that contain non-letter symbolsand punctuation. This included the removal of @ and http links.
4. For repeating letters in words, our strategy was that when we detected two identical letters repeating, we kept both of them. If we detected more than two identical letters repeating, we replaced them with one letter. Therefore, "sooo" were corrected to "so" ."muuchh" was kept as "much" . Originally correct words such as "too" and "sleep" were kept as they were.
5. We used astopword list which we created to remove the common stopwords. We kept words like "much, more, all, always, still, only" , because the reviewers frequently use these words to express extent.
6 The stemmer in the NLTK toolkit was used to perform stemming in orderto unify different forms of a word, plurals and different forms of a verb.
7. Lemmatize these words for smoothing them after stemming.

C. Tokenization

This process splits the text of a document into sequence of tokens. The splitting points are defined using all non letter characters. This results in tokens consisting of one single word (unigrams).

D. Pruning

The review data set was pruned to ignore the too frequent and too infrequent words. Absolute pruning scheme was used for the task. Two parameters were used for the pruning task namely, *prune below* and *prune above*. The value of these parameters was set as: *pruned below=5 and pruned above =200* i.e. ignoring the words that appear in less than 5 documents and inmore than 200 documents.

E. Filtering tokens

Length based filtration scheme was applied for reducing the generated token set. The parameters used to filter out the tokens are the minimum length and maximum length. The parameters define the range for selecting the tokens. In the proposed model the minimum length was set to 4 characters and maximum length to 25 characters i.e. tokens with less than 4 characters and more than 25 characters were discarded.

F. Stemming

Stemming defines a technique that is used to find the root or stem of a word. The filtered token set undergoes stemming to reduce the length of words until a minimum length is reached. This resulted in reducing the different grammatical forms of a word to a single term. We used potter stemmer for the data set .

The general rules for dropping the endings from words include:

If a word ends in *es* drop the *s*.

i. If a word ends in *ing*, delete the *ing*unless the remaining word consists of a single letter or *th*.
ii. If a word ends in a consonant, other than *s*, followed by *s* then delete *s*.

Table 1.An Example of Preprocessed Review

Words and their stem.			
Review	**After Stopping**	**After Stem**	**After Lemmatize**
super easy to read screen in all lighted conditions .	supeeasy read screen lighted conditions	super eas read screen light condition	super easy read screen light condition

Table 1 shows different Grammatical forms of a Word and the corresponding stopped and Stemmed words

The stemming technique increases the efficiency and effectiveness of the information retrieval and text mining processes. Matching the similar words results in improved recall rate and also reduces the indexing size as much as 40-50%.

G. Polarity Check

Our classifier accuracy was 76% .ThePreprocessed dataset is connected to the polarizer method where each review is collected and split into sentences and each sentence is tokenised and stored in a list And each word is classified as positive or negative by Naïve Bayes classifier ., If positive then + 1 is incremented to the variable pos if negative then +1 is incremented to a variable neg .The sentence is tested for the word "negtok" if this word exists in list then everything after not is tested for whether it is a positive word or a negative word and consecutive words polarity is finally the sentiment orientation is changed , for example "screen is not good" this is classified as negative as the word "good" is negated by the word "not".Thuspolarity of each word is found by the classifier and collective polarity isconsidered for each sentence. An Average Score is calculated if there is more than one sentence in the review by taking the total score and dividing it by the number of sentences in the review .

V. FEATURE EXTRACTION

Features or attributes of a product are generally represented as nouns and extracting nouns and their frequencies in the dataset provide an insight about the various attributes of the product . Then we reduce the original feature set by removing irrelevant features with less frequencies . Based on these review, categorization is done. There are five commonly used feature selection methods in data mining research, i.e., DF, IG, CHI, GR and Relief-F. All these feature selection methods compute a score for each individual feature and then select top ranked features as per that score. Reviews are categorized based on it

A. Document Frequency (DF)

Document Frequency measures the number of documents in which the feature appears in a dataset. This method removes those features whose document frequency is less than or greater than a predefined threshold frequency. Selecting frequent features will improve the likelihood that the features will also be comprised by prospective future test cases. The basic assumption is that both rare and common features are either non-informative for sentiment category prediction, or not impactful to improve classificationaccuracy Research literature shows that this method is simplest, scalable and effective for text classification .Also most of the adjectives, adverb, and a small set of nouns and verbs can acquire semantic orientation

The schemes used for word vector creation includes: Term Occurrence, Term frequency and TF-IDF (term frequency-inverse document frequency).These are based on the following values:
fij: total occurrences of the termi in the document j.
fdj: total number of terms occurring in document j.
fti: total number of documents in which the term ioccurs.

Term occurrence**:**defines the absolute number of occurrences of a term.

Term occurence = fij

Binary term occurrence**:**term occurrence is defined as the binary value.

Binary Occurrence = 1 for fij>0 and = 0 otherwise

B. TF-IDF:

It describes how important a word is for a document. It consists of two parts: termfrequency (TF) and invert document frequency (IDF).

TF-IDF= (fij /fdj)log(1/fti).

C.Algorithm for Feature Extraction

1. Input Data set.
2. Generate New Data set by using Preprocessing
3. Using POS tagger, Extract only the noun words in a word vector .
4. Construct a tokenset using noun words from the data set .
5. Find ten most frequently used nouns, using frequency distribution from the token set using TF-IDF .
6 Filter the required features from this List .
7. Categorize reviews based on it .

VI. CLASSIFICATION

Machine learning approaches simulate the way humans learn from their past experiences to acquire knowledge and apply it in making future decisions. These learning techniques are widely used in artificial intelligence and document classification. The classification using machine learning can be summed up in two steps:

1. Learning the model using the training dataset

2. Applying the trained model to the test dataset.

Sentiment analysis is a text classification problem and thus any existing supervised classification method can be applied. Our work uses the Naive Bayes classifier for classifying the Amazon Product reviews .

Naïve Bayes classifier is a simple probabilistic classifier that is based on the Bayes theorem. Thisclassification technique assumes that the presence or absence of any feature in the document is independent of the presence or absence of any other feature. Naïve Bayes classifier considers a document as a bag of words and assumes that the probability of a word in the document is independent of its position in the document and the presence of other word. For a document d and class c:
$p(c|d)=p(d|c)p(c)/p(d)$
Our classifier has an accuracy of 76 %.

VII. RESULTS AND DISCUSSION

After preprocessing the test data file of Amazon review files and passing it to the features Extraction method resulted in a list of 10 most used nouns and their frequencies as shown below .

[('kindle', 211), ('screen', 156), ('library', 136), ('price', 118), ('device', 79), ('battery', 78), ('time', 40), ('life', 38), ('replacement', 37), ('months', 32)]

Out of these we selected 5 important features for categorization and classification as shown.
("screen', battery, 'price', 'weight ', 'library") .

The training dataset was categorizedusing Multinomial naive bayes classifier into five different categories based on it and the following words were extracted as tokens for each category and a bag of words was used to categorize each review based on it .

Table 2

Sno	Appearance of Words in each Category	
	Category	**Words**
1	Screen	Touch , eye , glare, backlit, uneven , freeze ,lighting navigation, greyscale , navigation ,strain
2	Battery	Charge ,lifereplacement,time, port, time, connecting , hours, weeks, recharge
3	Price	Cost, afford, sale, purchase,expensive, buy, money Cheap,
4	Weight	Carry, easy, hand, handbag,light , weight, convenient,portability, heavy , travel
5	Library	Books, storage, download, read , carry, store , cheap, font, bookstore, travel

and their total polarity of each category was quantified which is as shown in the graph . we only focused on the above five attributes of the product kindle. The total polarity of each attributes was stored in variables. Then we add all positive score and all negative score and graphical representation of the result is shown below

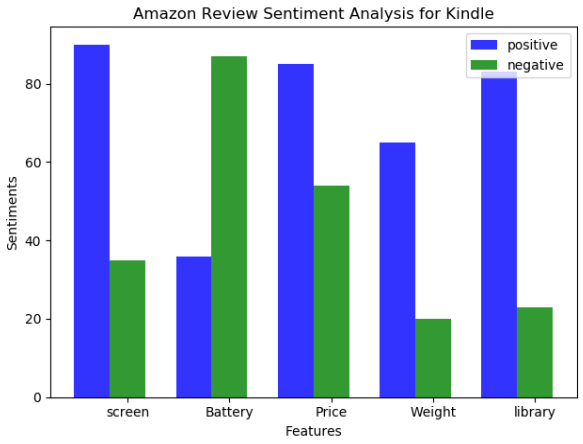

Figure 2.Result of Graphical Analysis of polarity of each category

From the above we can interpret the positivity and negativity about each of the features of the product kindle where we find that the battery is having higher negative value than positive , and the rest of the features have higer positive values which shows that the users are happy with these features.

VIII CONCLUSION AND FUTURE SCOPE

Sentimental Analysis is an easier andcost effective way to understand how the people are feeling about a particular subject of matter [5].. The Naive Bayes classifier used in the algorithm and polarity algorithm results depicted clearly the sentiment of the buyers and thus interpret the data .

It has been observed that the pre-processing of the data greatly affects quality of detected sentiments. We find the sentiments for each category of featuresseparately. The polarity algorithm finds score of eachwords. Then sentiments are classified as positive, negative. The analysis for each attribute of the product results in finding what is peoples opinion about the various product features which can be represented graphically in experimental results section.These results can guide the owners of the product to detect customer attitude and improve on the aspects that seem negative or is disliked by the targeted audienceand can improve their online reputation .

ACKNOWLEDGMENT

I express my sincere regards to my guide Dr. Ani Thomas and Dr. Neelam Sahu for the valuable guidance , inputs and whole-hearted cooperation for developing this work.

REFERENCES

[1] Y. Yang and J.O Pedersen, "A Comparative study on Feature Selection in Text Categorization ",In *International Conference on Machine Learning (ICML)*, 1997.

[2] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment analysis of Review Datasets using Naïve Bayes' and k-NN Classifiers." ,Int. J. of Information Engineering and Electronic Business,vol**. 4**, pp**. 54-62**, 2016.

[3] I. Hemalatha , G. P Saradhi V., &A. Govardhan, "Preprocessing the Informal Text for efficient Sentiment Analysis" , International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),Volume **1**, Issue **2**,2012.

[4] Wei Gao , Fabrizio Sebastiani ," TweetSentiment: From Classification to Quantification',Springer-Verlag Wien **2016**

[5] M. Bouazizi and T. Ohtsuki, ''Sentiment analysis in Twitter: Fromclassification to quantification of sentiments within tweets,'' inProc. IEEEGLOBECOM, Dec. 2016**, pp. 1–6**.

[6] Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande," Amazon Review Classification and SentimentAnalysis", International Journal of Computer Science and Information Technologies, **Vol. 6 (6)** , 2015, **5107-5110**

Authors Profile

Smita Suresh is working as Assistant Professor in Department of Computer Science at St. Thomas College, Bhilai, DurgUniversity,Chattisgarh , India and is currently pursuing Ph.D from Dr. C.V. Raman University , Kota, Bilaspur C.G.

Dr. Ani Thomas is Professor in Department of Information Technology at BIT Durg, CSVT University Chattisgarh , India since 2000 She did her Ph,D from CSVT University in 2012. She is a life member of ISE & Computer Society of India . She has published more than 20 research papers in reputed international journals index by Scopus and conferences including IEEE and it's also available online. Her main research work focuses on Text Mining and Machine learning. She has 20 years of teaching experience and 10 years of Research Experience.

*Dr.Neelam Sahu*is Proffessor in Department Computer Science at Dr C.V. Raman University ,Kota Bilaspur , C.G since 2016, She is a member of International Association of Engineers and has published more than 25 research papers in reputed international journals and conferences and it's also available online. His main research work focuses on SoftComputing based data in Fuzzy,Cyber Crime and Cloud Computing . She has 5 years of Research Experience.

