Sentiment Mining Based on Products Reviews Using Machine Learning

Gurjeet Kaur¹* Richa Dutta²

¹ Research Scholar, M.Tech (CSE) Yamuna Group of Institutes, Yamuna Nagar

² Assistant Professor, Yamuna Group of Institutes, Yamuna Nagar

Abstract – The increasing use of Internet and online activities (such as chatting, conferencing, surveillances, hotel and ticket reservation, B2C and B2B e-commerce, various social media platforms, blogging and micro-blogging, gets for us a very huge database of structured and unstructured data, referred to as Big Data, leading us to extract, transform, load, and analyse this data for interpretations. Such data can be examined using a mixture of Data Mining, Web Mining and Text Mining techniques in various real life applications. The base research by Rushleen et al [3] performed mining on tweets obtained on the Samsung Electronics twitter handle. The algorithm accurately analysis the positive, negative and moderate tweets. The algorithm accuracy is measured in terms of accuracy percentage and time complexity. These values are found to be 80% and O(m*n) respectively. This paper focuses on extracting the features from product reviews taken from Cornell University Movies Review Database given by reviewers to state their opinions. This is done at aspect level of analysis using ontology. Then it determines whether they are positive or negative thereby giving a scaling system to identify the effectiveness of a product. The scaling system is in the form of a -5 to 0 to +5 marking system where negative and positive values indicate the customer sentiment. Output of such analysis is then summarized using machine learning techniques using Weka Tool.

Keywords: Natural Language Processing, Ontology, Text Mining, Sentiment Analysis, Classification Algorithm

1. INTRODUCTION

Sentiment analysis (SA) is a process that involves the computation of sentiment lying in a sentence (Povoda, et. al., 2016). It is the task of classifying opinions, sentiments and attitudes pertaining to different fields, as expressed in input text. This study has been of great help in fields like observing public sentiment regarding political party, market behaviour or intelligence, the analysis of customer satisfaction, a concept or a movie sales prediction etc.

Natural language processing (NLP): NLP is aimed at making the machine behave like human and improving the interaction between them (Llombart, 2017). This is possible from a system that can understand and produce language as good as human can produce. The advent of internet has enormously increased the problem of information overload.

E-commerce is increasingly becoming a part of our life with every passing day and the reviews of existing consumers of any product are the best and most impartial way to validate the authenticity of quality of a product. It is almost impossible to go through all the reviews for any product because it can be very huge at times. So we need a way to do this analysis and provide a numerical analysis of the customer sentiment after having used the product. The review rating system under study here can be used to judge the customers average sentiment about the product.

1.1 Feature Extraction

While trying to extract the particular feature from a review or comment, the first thing is how to represent the review (Zou, et. al., 2015). The conventional VSM (Vector space model) is one of common ways for this. The VSM considers the documents as vectors and works in a quantitative way. The review sentences are added to individual vectors by creating tokens from them. The filtering of stop-words is performed on the vector. Then, stemming is done on each word in the vector. The effectiveness of these words is evaluated by using quantitative methods and their statistical information including the frequency of each word stem, or another similar measure, in the comment as the
corresponding element in the vector used to represent that text. The following steps are performed for text classification.

**Parsing the documents**

This step removes all abbreviations and non-alpha characters from the review sentence.

**Case folding**

Case-folding means converting all the characters in a text into the same case.

**Removing stop words**

Stopwords in English language are like conjunctions, articles, pronouns and prepositions which provide structure to the sentence. There are certain words that occur frequently and don’t have any useful information contained. This step removes such words from sentences in consideration.

**Stemming**

Stemming means reducing the derived words to their original word like buffering to buffer. Porter stemmer performs suffix stripping. The steps in Porter’s stemming algorithm are:

1. Remove the plurals and suffixes like –ed or–ing from each token.
2. Convert each existence of y to i when another vowel in stem.
3. The double suffixes are mapped to single ones: -ization, -ational etc.
4. The suffixes like -full, -ness etc. are either removed or properly dealt with.
5. Remove the suffixes like –ant, -ence, etc.
6. Gets rid of a final –e.

**Term Weighting**

In this step, weight is assigned to a word based on number of times it occurs in the comment. This method is called term frequency and inverse term frequency which is a traditional method to assign a weight to the words.

The explicit feature can be extracted as:

- Depending on frequency of nouns and noun phrases.
- Based on the relations between Opinion and Target.

1.2 **Machine Learning Methods**

In the process of this research, we have gone through several machine learning techniques. We discuss here the most prominent machine learning methodologies and detail about each one of them.

1. **Naïve Bayes method**- In this algorithm, the occurrences of values and combinations of values are counted from historical data related to conditional probabilities. This technique is also termed as conditional probability. This technique is based and dependent on hypothetical data collection and manipulation. It is more suitable to large data sets, for e.g weather forecast mining.

2. **Decision Tree**- The classifier technique decision tree, performs recursive partition of the instance space. The positive and negative results are taken as leaf nodes in this technique. The root of a directed tree is created from the most significant review and it never has any incoming edges. Decision tree is used as trade-off for highly predictive techniques.

3. **K- Nearest neighbour**- K-NN is a non-parametric classification technique which is much less preferred than SVM and Naïve Bayes techniques. It does not perform any generalization and so, the training data points are not included. K-NN performs lazy learning which means that the function is only estimated locally the computation gets delayed until classification. This is comparatively simpler technique than all other machine learning algorithms.

4. **Support vector machine**- SVM stays the most prominently methodology for sentiment analysis in social networks. Due to its dependency on size of dataset, SVM is unfavourable for large scale data mining due to training complexity (Kasthuri and Kumar, 2014). The support vectors which have data located closest to the boundary, describe the boundary function. The data is transformed into a higher dimension, using a nonlinear mapping, in SVM (Kasthuri and Kumar, 2014). SVM searches for optimal linear hyperplane within the limits of this new dimension.

Let’s take an example of 12 samples of peoples where we have two classes of people – first who buy a computer and the other who doesn’t.
In the next step, the hyperplane that separates the nearest training tuples with maximum distance is identified using SVM.

5. **K-Means Clustering** - This method classifies the dataset through a certain number of K clusters. Then it defines K centres for K clusters which are placed as far as possible from each other. Then each point in dataset is associated to the nearest data centre. The following formula finds the new cluster centre (Maryam, 2014).

\[ V_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i \]

Where \( c_i \) represents the number of data points in \( i \)th cluster and \( x_i \) is the data point. The above process is repeated till no new centre is reassigned.

Lets take an example for K-Means clustering (Maryam, 2014) We have 20 positive and negative samples with us. We have reduced them to 6 cluster centers C1,C2,C3….C6 by K-Means algorithm.
in [2] dealt with sentiment analysis in text documents, especially text valence detection. The proposed solution was based on Support Vector Machines classifier. This classifier was trained with huge amount of data and complex word combinations were analysed. Datasets used for training and testing were automatically obtained from real user feedback on products from different web pages (and different product segments). The proposed solution was evaluated with different languages – English, German, Czech and Spanish. This paper improved accuracy achieved with the Big Data approach about 11%. The best accuracy achieved in this work was 95.31% for recognition of positive and negative text valence. The described learning was fully automatic, and was flexible to be applied to any language and no complicated preprocessing was needed.

According to Raghuvanshi et al [3], it was mandatory to mine the opinion on the web, to perform a well-defined task, so that the researcher could retrieve the information from the available product and services data from internet. The author had started the discussion with the introduction on sentiment analysis, which gave an insight into sentiment analysis. Different types of sentiment analysis model based on word, sentence, feature and document were also discussed and found that document level model was more promising for better results. The detail discussion on various methods proposed by different researchers was also presented. Different types of sentiment analysis techniques gave a research direction in different directions. Finally a method was proposed based on the naïve bayes classifier. Huang Zou et. al in [4] performed sentiment classification using machine learning techniques to improve precision and accuracy. The authors generated syntax trees of the sentences, with the analysis of syntactic features of the sentences. The features were trained on data of movie reviews, using Naïve Bayes and support vector machines, the most preferred machine learning techniques. The research proposed and generated a more accurate solution for sentiment classification by examining the syntax tree for features and factors. The authors used the dataset published by Cornell. The dataset contained 2000 movie reviews, half of which were positive while the other half were negative. Preprocess were conducted firstly to remove unrecognized words in English, extend short form words, and correct wrongly spelled words. Syntax trees were constructed firstly and words dependencies were also generated to reveal the grammatical and logistic relationship between words in sentences. Then features about clause information as well as POS tags were generated, and words-bag collection was optimized by using bigrams with words dependencies. The researcher then used SVM and Naïve Bayes to analyse. From the result the authors observed that words dependencies and POS tags improved the accuracy of bigram method.

Deepshikha et. al [7], extracted the features from bank reviews extracted from mouths hut.com and myBankTracker.com sites. This was done at aspect level of analysis using ontology. Then it determined whether they are positive or negative. Output of such analysis was then summarized. The research infers that the average human online customer didn’t have knowledge on identifying relevant sites and getting his inference by extracting and summarizing the opinions stated in those websites. Automated sentiment analysis systems were thus needed. The research used a combination approach of domain ontology and Stanford dependency relation which intended to enhance the sentiment classification. By using this approach one could view the strength or the weakness of the features of a particular bank in more detail.

3. PROBLEM DEFINITION

E-commerce is increasingly becoming a part of our life with every passing day and the reviews of existing consumers of any product are the best and most impartial way to validate the authenticity of quality of a product. It is almost impossible to go through all the reviews for any product because it can be very huge at times. So we need a way to do this analysis and provide a numerical analysis of the customer sentiment after having used the product. The review rating system under study here can be used to judge the customers average sentiment about the product. The base research by Rushleen et al achieved 80% accurate results when compared to actual results. This level of accuracy is not acceptable in current industry standards where every service provided has a cut throat competition.

4. BASE RESEARCH

In the base research, Rushleen Kaur et al in [3] aimed to undertake a stepwise methodology to determine the effects of an average person’s tweets over fluctuation of stock prices of Samsung electronics ltd. It involved extracting tweets from tweeter, data cleaning and application of suitable algorithm to extract the correct sentiment from these. The authors studied the vast impact by twitter feeds. The algorithm accurately analysis the positive, negative and moderate tweets. The algorithm accuracy is measured in terms of accuracy percentage and time complexity. These values are found to be 80% and O(m*n) respectively.

5. PROPOSED WORK

SAMPLE DATA – REVIEWS
The customer reviews of different products for any particular enterprise are considered to extract entity level sentiments. The two most common subtasks in text mining and SA are - Data acquisition and data preprocessing. The analysis of product reviews is performed in two steps:

(a) extracting most important features of product
(b) assigning an overall score for each of them.

This allows us to structure information from reviews by summarizing them in a comprehensive and concise form. We can understand this problem statement like: We have a review as form of sentence $S_n$, and we need to identify the sentiment scores $S_{cS_{Sn}}$ of relevant features.

**ALGORITHM**

Step1: Input all reviews in form of Text

Step2: Organize the reviews into array of sentences and the array elements represent individual review.

Step3: Repeat steps for $i=1$ to length (array)

Step4: We denote each review by $s[i]

Step5: Split $s[i]$ into different factors and the sentiment of each factor is evaluated by sentiment = evaluate (factor)

Step6: Final total Sentiment of all elements

Step7: Evaluate the sentiment to order of 5

Step 9: Find comparison of evaluated result from proposed technique and from the base technique. The comparison will be done using Naïve Bayes and Decision tree J48 algorithm. [end loop].

$S -> sentence$

Check Value(S)

Take initial value = 0

Step: Convert the sentence to lowercase

Step 2: Perform Filtering to remove extra symbols and unwanted words from the sentences.

Step 3: Each word is taken through stemming

Step 4: Finally extract the sentiment carrying words and compare them to lists of positive words, negative words, domain specific positive and domain specific negative words.

Step 5: Each match modifies the result value.

**Match (Word1, Word2)**

Step 1: Return true if word1 = word2

Step 2: $x = \text{soundex (word1)}$

$y = \text{soundex (word2)}$

Step 3: if $x=y$

Return true

Step 4: Now the synonyms of the words are found by Wordnet and results true when successfully matched.

**FLOWCHART OF PROPOSED TECHNIQUE**

The process flow diagram of our work is shown in the figure below.
6. CONCLUSION

NLP has been an area of wide interest for years with the researchers. With the advent and advancement of machine learning techniques, the results have become refined and accuracy level has gone considerably up. In this research we use machine learning classification techniques like Naïve Bayes and SVM(Support vector machines) to create a sentiment classification system with high degree of accuracy.

The proposed methodology can be used in other domains like social network data to classify the intent and sentiment of an end user. This can be helpful to put some accounts under scrutiny who consistently post some objectionable data. This can be a huge support to identify criminals and terrorists who use social networks for spreading hatred or their malicious messages.

7. ACKNOWLEDGEMENT

I owe my thanks to the almighty first and then to my guide Er. Ms. Richa Dutta who helped me to accomplish this research work successfully. I am also grateful to my family and friends for all their kind and unconditional support.

REFERENCES


https://nlp.stanford.edu/software/lex-parser.html


Corresponding Author

Gurjeet Kaur*

Research Scholar, M.Tech (CSE) Yamuna Group of Institutes, Yamuna Nagar