International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 7 28 – 31

Polarity Categorization on Product Reviews

Mixymol V.K.
Assistant Professor,
Department of Computer Science,
Kanchi Mamunivar Centre for Post Graduate Studies,
Puducherry-605008,
India. *Email : mail2mixymol@yahoo.co.in*
Mobile : 9443430086

Abstract: People search for other people's opinions from the internet before purchasing a product, when they are not familiar about a specific product. With the help of reviews, ratings etc. online data presents useful information to customers for buying a product and for manufacturers to improve the quality of product. When an individual wants to make a decision about buying a product or using a service, they have access to a huge number of user reviews, but reading and analyzing all of them is a tedious task. Reading all of them is generally inefficient. There is a need for summarization in product reviews. Sentimental analysis helps customer visualize satisfaction while purchasing by simple summarization of these reviews into positive or negative two broader classified classes. The study aims to tackle the problem of sentiment polarity categorization. The data set is collected from amazon.com. The data set contains 376 instances of reviews of Nokia mobile in the form of a text file. Two classification algorithms namely Naïve Baye's and Support Vector Machine Algorithms are taken to classify the reviews as positive, negative or neutral.

Keywords*: Product review, Sentiment analysis, Classification, Polarity Categorization.*

*****

I. INTRODUCTION

In past days, buying of products was more based on getting product review from nearby neighbors, relatives etc. as products were purchased directly from merchants. But with change in technology, development of E-commerce industry with sites flooded by products from different brands made available to customers at the touch of one click.[1] The availability of product based sites with doorway delivery has made it convenient for customers to shop online. With so much change in shopping pattern, merchants providing customers with feedback option about the product. Customers write reviews from all parts of the world. A lot of reviews are very long, making it hard for a potential customer to review them to make an informed decision on whether the customer should purchase the product or not. A huge number of reviews also make it difficult for product manufacturers to keep log of customer opinions and sentiments expressed on their products and services. It thus becomes necessity to produce a summary of reviews. Summarization of reviews is done using sentiment analysis.[2]

Sentiment analysis tends to extract subjective information required for source materials by applying natural concept of natural language processing. The main task lies in identifying whether the opinion stated is excellent, good or bad. Since customers usually do not express opinions in simple manner, sometimes it becomes tedious task to judge an opinion stated. Some opinions are comparative ones while others are direct.[3] Sentimental analysis helps customer visualize satisfaction while purchasing by simple summarization of these reviews into positive or negative two broader classified classes. Comments are mainly used for helping customers purchase online and for knowing current market trends about products which is helpful for developing market strategies by merchants.

There are several reasons why potential buyers read product reviews, Common questions readers would like to know include:

- Is the product easy to use?
- Is it of a high quality?
- Is it geared towards somebody like me?
- Have others had a good experience with the product?
- What are the pros and cons of the product?
- What alternatives are available, and how do they measure up?
- Is the product worth my money?

II. PROBLEM STATEMENT

One of the most common applications of sentiment analysis is the area of reviews of consumer products and services. There are many websites that provide automated summaries of reviews about products and about their specific aspects. This work focuses on analyzing product based reviews by classifying them as positive, negative and neutral. Dataset for the study is collected from Amazon.com and the dataset contains 376 reviews of Nokia mobile. The dataset

contains unstructured data which is in the form of text file. Implementation of the work has done in R programming language. Various classification rules namely, Naïve Bayes Classifier and Support Vector Machine are used to train a sample dataset and measure the Accuracy value using R programming language.

III. DATA DESCRIPTION

The dataset for the project work is collected from Amazon.com. The data is collected from https://github.com/SamTube405/Amazon-E-commerce-Data-set. A subset of the dataset which consists of 376 reviews for Nokia mobile is considered for this study. Reviews are in the form of a text file and the reviews include product, user information, rating, plain text, and it contains special characters symbols.

Sample product reviews:

- best 4mp compact digital available
- camera is perfect for an enthusiastic amateur photographer
- macro[+3]##the pictures are razor-sharp , even in macro
- it is small enough to fit easily in a coat pocket or purse
- it is light enough to carry around all day without bother

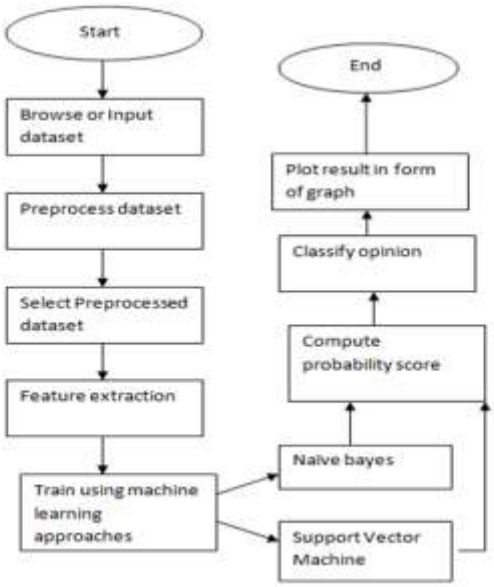
Experimentation and Result analysis

R programming Language is used to implement the classification algorithms. R is a programming language for statistical calculation and it is created by statistician for make statistical data analysis easier. Rstudio is an Integrated Development Environment (IDE) that helps to develop programs in R. And it is free and open source software. And also it is used for graphics representation and reporting.

Packages used for this study includes

- **NLP** - Natural Language Processing
- **tm** - text mining packages
- **RTextTools**-It is a machine learning package for automatic text classification
- RColorBrewer-It is an R packages that uses the work from to help you choose sensible color schemes for figures in R
- **Bitops**- Doing Bitwise operations
- **plyr** The package plyr is a set of tools that solves common problems by breaking down bigger problems into more workable pieces. The package then operates on each problem before reassembling the reworked pieces back together.
- **Stringr** stringr makes R string functions more consistent, simpler and easier to use by ensuring the function and argument names are consistent and all functions deal with NA's and zero length characters appropriately. Stringr also ensures that the data output from each function matches the input data structures of other functions.
- **ggplot2** This package provides an implementation of the grammar of graphics in R, combining the advantages of both base and lattice graphics. Plots can be built up step by step from multiple data sources.
- **e1071** this package provides function for latent class analysis, such as Support vector Machines, bagged clustering and Naive Bayes classification.

Figure 1 gives the step wise activities of the work done.



Pre-processing

The dataset is unstructured; it may contain repetitive words, large number of words that are not at all needed in summarizing of opinions. Pre-processing involves removal of stop words such as 'and', 'or', 'that' etc. followed by porter stemming which involves simplifying target words to base words by removal of suffixes such as – ed, ate, ion, ional, ment, ator, ssess, es, ance or conversion from ator to ate etc. For example, "replacement" is stemmed to replac; "troubled" to trouble ; "happy" to happi ; "operator" to operate. The raw data is pre-processed to improve quality.

Feature Extraction

Features in reviews are extracted so that it helps customer to know which feature has positive comment and which one has negative. Since, overall conclusion about product is much needed but there is also situation where customer requirements come into the scenario. Use of adjectives is done to classify opinions as positive or negative using unigram model. For example, "the Samsung camera I bought was good; it has got great touch screen, awesome flashlight." The feature extracted out of it would be like: Domain: Mobile; Product: Samsung; Feature: Camera; Adjective: Good.

Training and classification

Supervised learning generates a function which maps inputs to desired outputs also called as labels because they are training examples labelled by human experts. Here, naïve Bayes and Support Vector Machine techniques to carry out supervised learning on the dataset fetched.

Support Vector Machine:

Support Vector Machines (SVM) is binary classifies is able to classify data samples in two disjoint classes. The basic idea behind is classifying the sample data into linearly separable. Support Vector Machines (SVMs) are a set of related supervised learning methods used for classification and regression. In simple words given a set of training examples, each marked as belonging to one of two categories. An SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. SVM is a set of related supervised learning method for classification and regression.[4] SVM simultaneously minimize the empirical classification error and maximize the geometric margin. SVM is called Maximum Margin Classifiers and it can be efficiently perform non-linear classification using kernel trick.

Naïve Baye's Classifier

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. [5]A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Fgure 2 gives the classification of these input set into positive, negative and neutral and figure 3 shows the performances of these algorithms are measured by the accuracy value. The comparative study shows that SVM performs better than Naive Bayes for Sentiment analysis on product reviews.

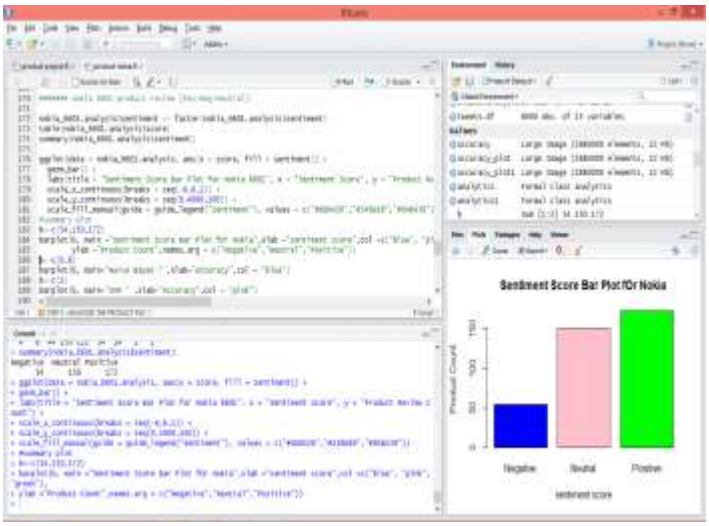


Figure 2 Classified as positive, negative and neutral

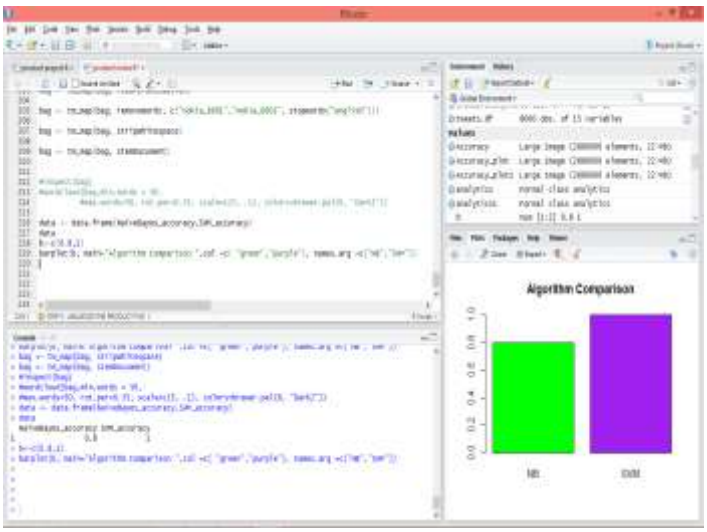


Figure 3 Algorithm comparison

IV. CONCLUSION and Future Directions

Sentiment analysis deals with identifying and aggregating the sentiment or opinions expressed by the users. Sentiment analysis is to classify the polarity of text in document or sentence whether the opinion expressed is positive, negative, or neutral. Here two approaches have been compared and a result for both approaches on the product review dataset has been done. SVM is found to give better accuracy that is 100% as compared to Naïve Baye's approach which is 80%. The study can be extended to create manual dictionaries, like gd,(for good, xlnt (for excellent), and using other sybols like !!!, * etc.,

REFERENCES

[1] Manvee Chauhan, Divakar Yadav "Sentimental Analysis of Product Based Reviews Using Machine Learning Approaches" Journal of Network Communications and Emerging Technologies (JNCET) Volume 5, Special Issue 2, December (2015) .

[2] Pravesh Kumar Singh1, Mohd Shahid Husain2 "Methodological study of opinion mining and sentiment analysis techniques" International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014.

[3] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam "Comparative Study of Classification Algorithms used in Sentiment Analysis" Amit Gupte et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6261-6264.

[4] https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code

[5] http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf.