Sentiment Analysis on Product Reviews using Machine Learning

Vishal Hulawale[1] Rushikesh Killedar[2] Shubham Gotal[3] Jayesh Chaudhari[4] Mahesh Shelar[5]

[1,2,3,4,5]Department of Information Technology

[1,2,3,4,5]AISSMS IOIT, Pune

Abstract— Now a day's use of e-commerce is increased. Peoples buy products and post their opinions, suggestions about topic or product. Also before buying a product they may need to go through those reviews before making any decision. We are proposing a system that will help them to finalize their choice and take a decision whether or not to purchase the product. If the number of reviews is more then, going through all reviews will become time-consuming for users. They will not able to interpret all product reviews and might get confused. Our system is based on sentiment analysis than can be used to interpret reviews and summarize reviews in the user suitable format. It focuses on specific words or attributes that the customer will be interested in product. Product reviews will be classified based on emotions extracted from reviews. We will be using machine learning based approach for sentiment analysis of product reviews. To process product reviews we are using the use of machine learning algorithm called Support Vector Machine. This system significantly reduces the time needed for the customer for going through an intensive process of reading multiple reviews.

Key words: Mi Machine Learning, Sentiment Analysis, Product reviews

I. Introduction

In recent years use of the E-commerce is increased. E-commerce makes the incredible or revolutionary change in the financial process. People also have changed their perspective about e-commerce and started to rely on it. So, to increase in sales merchants enabled the customer to share their reviews about the product on their site to increase the interaction with the customer. If the customer wants to check the product all he has to do is to go through reviews but going through all these reviews it can be problematic in terms of an amount of time that customer has to spend to read all reviews, will be large.

The merchandise provides a rating system to reduce the time required for selecting the product. The review used to generate the ratings. Data mining technique can be applied to assess the information and their classification. Text mining consists of techniques to analyze human language. So, sentiment analysis can automate the process of rating based on summarization of reviews.

A. Sentiment Analysis:

Sentiment analysis is a process of computationally identifying, categorizing reviews expressed in the form of text, to determine user's attitude towards the product. Sentiment analysis aims to detect polarities regarding the product. The main problem is to distinguish it from topic based classification because topics can be identified by keywords and sentiments can be expressed in more delicately.

Sentiment analysis is a subfield of Artificial intelligence focused on parsing the given text and proposed its opinion in terms of positive, negative or neutral text. Feature - based opinion summarization identifies the features in the given review and expresses the sentiment relevant to that feature. A simple example to illustrate features in the sentence would be as follows:

– "The display quality of the phone is fantastic."
– "The battery life though is draining fast."

Here, "display" and "battery life" should be considered as features in the above sentences respectively. By using such summarization, a potential customer might be able to narrow down his choices of the product if he's interested in specific features and also ease him in comparing the products.

II. Recent work

A comparative study between different methodologies has been reviewed and analyzed including subjectivity detection, feature selection for opinion mining, and different machine learning approaches [1]. The various mechanism has been implemented until now, which includes bags of words, training corpus, document level, sentence level and feature-level opinion mining [3]. Different polarity measures exist according to the external system wherein sentimental analysis is utilized. The linguistics feature and domain relevant features are essential for providing the better classification of text [2]. Hence, in this system, consideration the gamut of keywords associated with the feature is essential for successful classification. The algorithm explained revolves around the expansion and better understanding of the model proposed by Dave, Lawrence, and Pennock [8].

III. Proposed techniques

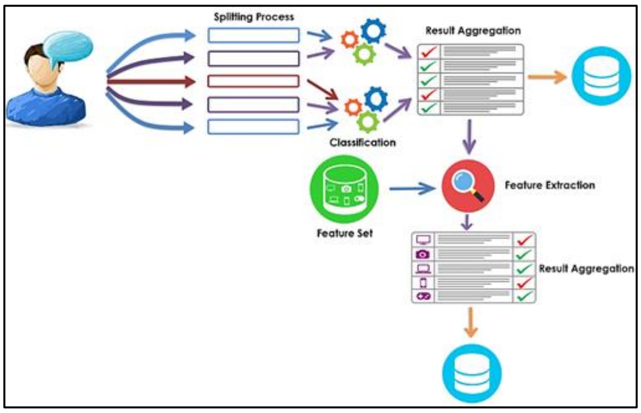

Fig. 1: Classification and Feature Extraction

Our system is based on reviews given by users on e-commerce sites. When a user submits a review it will be split into sentences with help of NLP framework. Once we get splitted sentences, each sentence will be given to classifier i.e. SVM (Support Vector Machine) algorithm. SVM will classify these sentences using pre-generated model and assign polarity i.e. positive or negative to every sentence. Once we get polarity for every sentence we will aggregate results and store them in HBase database. The output of classification will be given to feature extractor as input. In feature

]

extraction, every sentence will be scanned for features of that product using predefined keywords. And if we found a reference for the particular feature then respective count i.e. posCnt or negCnt will be increased.

IV. Methodology

A. *Support Vector Machine:*

In machine learning, Support Vector Machine is a supervised learning model with an associated learning algorithm that can analyze data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories; an SVM training algorithm builds a model that assigns new examples into one category or the other, which makes it a non-probabilistic binary linear classifier.

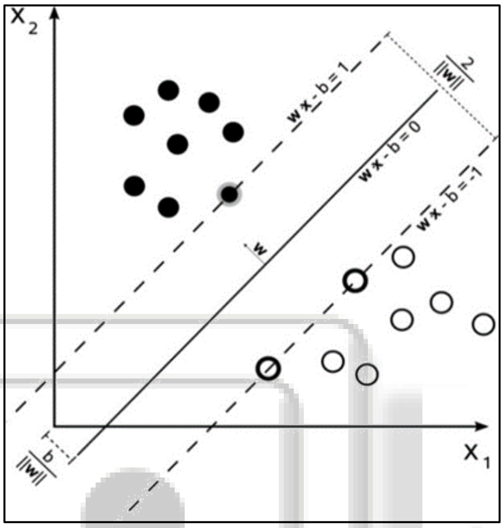

Fig. 2: Support Vector Machine

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into the same space and predicted to belong to a category based on which side of the gap they fall on. We are given a training dataset of n points of the form

$$(\overrightarrow{X_1}, Y_1), \ldots\ldots., (\overrightarrow{X_n}, Y_n) \qquad (1.1)$$

Where the Y_i are either 1 or -1, each indicating the class to which the point $\overrightarrow{X_i}$ belongs. Each $\overrightarrow{X_i}$ is a p-dimensional real vector. We want to find the "maximum-margin hyperplane" that divides the group of points $\overrightarrow{X_i}$ for which $Y_i = 1$ from the group of points for which $Y_i = -1$ is defined so the distance between hyperplane and the nearest point $\overrightarrow{X_i}$ from either group is maximized.

Any hyperplane can be written as the set of points $\vec{X}$ satisfying

$$\vec{w}.\vec{x} + b = 0 \qquad (1.2)$$

Where $\vec{w}$ is the normal vector to the hyperplane. The parameter $\frac{b}{\|\vec{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector.

B. *Phases of operation:*

1) Cleaning of data

Punctuations and special characters are to be removed from the sentences such that only alphabets and number are left in the sentences. The sentences are also entirely converted into lower cases.

2) Splitting Reviews into sentences

In this, the whole review is separated into small sentences. This is done using Apache OpenNLP.

3) Stop Words Removal

Stop words are considered as meaningless words, which are filtered out to reduce the processing time. This list consists of the preposition, conjunctions, articles, etc.

4) Classify Each Splitted Sentence

By classifying every sentence it gives result in the form of [Sentence][Class].

5) Feature Extraction

It represents searching for feature's related words in the sentence and then classifying in the same feature cluster. For example, the review data set is parsed for keywords/ feature (such as display: display, screen, gorilla glass, resolution, color, pixels) which are generated by finding frequent item set. In Feature Extraction each sentence is compared with a set of feature database. If the feature matches then count of that feature is increased whether it is positive or negative.

V. Result Analysis

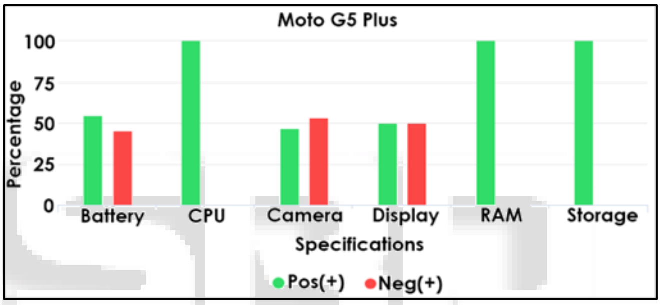

Fig. 3: Result in user suitable format i.e. chart

Results of analysis performed on product reviews will be represented in user suitable form using visualization techniques such as charts, graphs etc.

VI. Future Work

In future work, these techniques and rating process can be improved by taking into consideration the usage of slangs term and smiley symbols used by people. Features can also be clubbed together according to the score as good, neutral, and bad. Spam reviews can be detected and removed from the list to increase the overall efficiency. Neural networks can be implemented to improve efficiency of system.

VII. Conclusion

The system developed aims to achieve an efficient mechanism for summarizing the reviews posted by the customer to help other potential customers. It enables many e-commerce websites in the need of time to substitute their 'upvote' system for surfacing helpful reviews with the proposed system which doesn't involve manual intervention in the rating process. The system provides information in a graphical progress bar adjacent to each feature where each score is displayed by the length of the bar as shown in Fig3. We assert that usage of machine learning techniques for intensive processing of product reviews.

References

[1] R. S. Chandrakala, C. Sindhu. "Opinion Mining and Sentiment Classification: A Survey," ICTACT Journal on Soft Computing, October 2012, Volume: 03, issue: 01, ISSN: 2229-6956.

[2] Yun Niu, MSc, Xiaodan Zhu, MSc, Jianhua Li, MSc and Graeme Hirst, Ph.D. "Analysis of Polarity Information in Medical Text," AMIA 2005 Symposium.

[3] Nidhi Mishra, C. K. Jha, Ph.D. "Classification of Opinion Mining Techniques," International Journal of Computer Applications (0975 – 8887), Volume 56– no.13, October 2012.

[4] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, "Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," LREC 2010.

[5] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N, Prasad. M. R., "Analysis of Big Data using Apache Hadoop and Map Reduce," Volume 4, Issue 5, May 2014, India.

[6] Minqing Hu, Bing Liu, "Mining Opinion Features in Customer Reviews," American Association for Artificial Intelligence, 2004.

[7] Othman Yahya, Osman Hegazy, Ehab Ezat, "An Efficient Implementation Of Apriori Algorithm Based On Hadoop - MapReduce Model," International Journal of Reviews in Computing, ISSN: 2076-3328.

[8] Kushal Dave, Steve Lawrence, David M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," ACM 2003.

[9] R. Jose, V. Chooralil, "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Classifier Ensemble Approach," IEEE International Conference on Data Mining and Advanced Computing, 2016.

[10] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proc. 2002 Conf. Empirical Methods in Natural Language Processing, 2002.

[11] Neethu S., Rajashree R., "Sentiment Analysis in Twitter using Machine Learning Techniques," Fourth International Conference on Computing, Communications and Networking Technologies, 2013.

[12] E. Aydogan, M. Akcayol,"A Comprehensive Survey for Sentiment Analysis Tasks Using Machine LearningTechniques," International Symposium on INnovations in Intelligent SysTems and Applications, 2016.

[13] Kamal, M. Abulaish, "Statistical Features Identification for Sentiment Analysis using Machine Learning Techniques," International Symposium on Computational and Business Intelligent, 2013.