Volume 6, Issue 1, January 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Sentiment Analysis using Machine Learning

Jyoti Jain*, Ass. Prof. Archana Shinde, Prachi Panchal, Nihar Suryawanshi
Department of Information Technology, Sinhgad Academy of Engineering,
Maharashtra, India

Abstract- While purchasing products on an e-commerce website, a user has to go through multiple sites to find the product at its best in terms of prices as well as reviews. Going through complete reviews list to understand sentiments of other fellow users is a tedious task. Sentiment analysis with the blend of machine learning could be useful in predicting the product reviews or consumers attitude towards to newly launched product. Presently there is no such system which could aggregate the reviews and provide an un-biased insight of consumer sentiments over a product. Our system would be able to aggregate the reviews of a product from a web source and using machine learning algorithm would predict the sentiment of users towards it. This data analytic tool will provide insights to new consumers and product owner regarding the product. The system also aims to compare the similar products from various sources and provide deep analysis regarding same. The system once developed could be applied to compute the sentiment in any field where data analytics plays important role.

Keywords- Sentiment Analysis, Supervised Learning, Naive Bayes Algorithm, classifier, E-commerce websites.

I. INTRODUCTION

E-commerce caters to almost all the needs of the users today mainly because users expect to find products easily. Users want to take an informed decision of buying a certain product and that's why user convenience is of utmost importance. Usually while browsing for buying a product, users have to navigate through various sites and then make a choice regarding the same. For instance, while buying a mobile, a person would browse sites like Flipkart, Amazon, Ebay, etc. and then compare prices and reviews of the product and then come to a decision. The time required for this procedure is usually too much.

In this research project we aim at developing a portal which would reduce the overhead resulting from browsing. Some vendors provide better services in terms of quality than the others, and thus the need of analysing the sentiments is justified. The tool analysing the sentiments of these e-commerce websites will perform the function of classification of various reviews. The product user wishes to buy will be looked up from various websites, the reviews for the product will be extracted and then classified by the tool as positive or negative with the help of machine learning techniques. The aggregated polarity will be displayed to the user who now can make an easy choice.

II. LITERATURE SURVEY

A. Extraction and Classification

The reviews for the product a user wishes to buy can be obtained from various E-commerce websites through the process of extraction. HTTP requests are to be handled by the tool and then scrape the required reviews. The next task is of classification of the scraped reviews as positive or negative. This can be done using supervised machine learning techniques. Supervised machine learning techniques are the ones where the machine is trained to do a particular task based on a data-set. The trained machine is then tested for its performance by a new data-set called as testing data. We propose to train the machine using an existing data set comprising of good and bad reviews. During testing phase, reviews can be extracted from the websites as and when required and tested for their polarity.

B. Thumbs up and Thumbs down orientation

In Reference [1] the method of classifying a review as thumbs up (positive) or thumbs down (negative) is mentioned. Here, unsupervised learning algorithm has been used, where patterns are found in the given data. Classification of reviews is done by extracting only the adjectives from the sentence, i.e., classification are done by adjective processing. Initially, parts-of-speech tagger (POS tagger) is implemented for extraction of adjectives from a review. Then Pointwise-Mutual_Information (PMI) algorithm is used to estimate the polarity or the semantic orientation of the phrase. Then the review is classified based on its average orientation. The algorithm attains an accuracy of 74% for automobile reviews and 66% for movie reviews. The classification can be misleading though, because an adjective alone depicts subjectivity. The meaning of adjectives depends on the context where they are used. Also, adjectives alone cannot describe the entire meaning of the sentence.

C. Product Safety using Sentiment Analysis

Reference [2] collects public sentiments about a particular brand of drug or cosmetic products. The reviews of a sample population about these products can be monitored and thus product counterfeiting can be predicted. Social media

platforms like Facebook and Twitter are used for collecting the reviews. The data is analysed by both text mining and sentiment analysis techniques. A lexicon based classifier uses sentiment scoring function whereas Naive Bayes algorithm is used as another approach for classification. Out of these two approaches, the Naive Bayes classifier was found to be more effective with an accuracy of 83%.

D. Sentiment Analysis of Twitter Data

In Reference [3], classification of tweets from the social media platform Twitter is mentioned. The tweets are classified as positive or negative based on the sentiment they depict. Supervised learning methods are used to perform this task. Ensemble methods comprising of Naive Bayes algorithm, Maximum Entropy and Support Vector Machine have been used. Semantic analysis is used along with these algorithms using WordNet as a database, which further improves the accuracy of the model.

E. Emoticon based Sentiment Analysis

In Reference [4], classification of emoticons (smileys) has been described. Emoticons are widely used on the social media. The paper describes a system which processes Chinese tweets to understand their sentiments. The methodology used is Naive Bayes algorithm which classifies sentiments of emoticons into four types: angry, disgusting, joyful and sad.

III. PROPOSED SYSTEM

A. Disadvantages of Existing Systems

- Most models perform only word (adjective) processing.
- Accuracy of the algorithms varies when the data sets are changed.

B. Introduction to Proposed System

1) Modules

a) REQUEST AND RESPONSE HANDLER: This module will handle all the HTTP requests for fetching product reviews from the various E-commerce websites.

b) EXTRACTOR: Extractor will scrape the reviews from the sites and forward it to the classifier for further processing.

c) CLASSIFIER: Classifier will calculate the semantic orientation of the newly arrived data with respect to the data which it has been already trained upon. The reviews will be classified as positive or negative based on the sentiment they possess and the cumulative result will be displayed to the user.

2) System Architecture

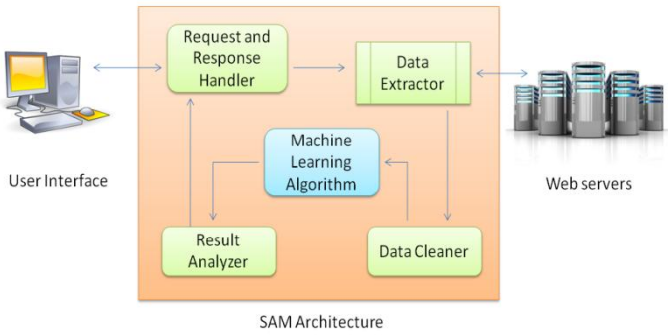
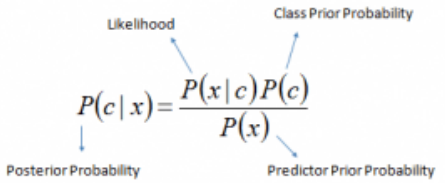


Fig. 1 System Architecture

The system architecture explains how the request from the user will be handled by a server and the extraction of reviews henceforth. The reviews will be analysed on the basis of a machine learning algorithm, in this case, Naive Bayes algorithm. The results will again be displayed to the user.

3) Naive Bayes Algorithm

It is an algorithm used for classification. It comes under probabilistic models of supervised learning. It is based on Bayes theorem which uses the concept of conditional probability. For example a person will be classified to have flu if he has cold and fever. The algorithm is called as Naive because all the properties which describe a victim of flu independently contribute to the probability. Although Naive Bayes is known to perform highly complicated and sophisticated classification tasks, it is easy to implement. The formula for calculating the posterior probability is as follows:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig. 2 Bayes Theorem

4) Steps of Naive Bayes Algorithm

Step 1: Building a frequency table from the data set.
Step 2: Building likelihood table.
Step 3: Calculation of class with highest posterior probability.

5) Advantages of Naive Bayes Algorithm

- Easy and fast.
- Performance is better compared to other algorithms.
- Categorical input variables are handled effectively.

IV. FEATURES

The system we are proposing provides simplicity to the user. Instead of traversing through various websites to read reviews about products, a cumulative output will be displayed directly to the user. The system acts as a recommendation tool. It saves the overhead of searching and then analysing reviews. Also, test cases like zero reviews for a product or if a website has stopped selling a product are handled by the system, which ultimately save the time of the user.

V. ADVANTAGES OF THE PROPOSED SYSTEM

- The system will be less time consuming.
- The system aims at better performance.
- The system will act like a recommendation engine for the end user.
- It aims to improve the accuracy of the algorithm.

VI. CONCLUSION

In this paper, we have analysed various classification algorithms. Machine learning techniques like supervised learning methods and unsupervised learning methods have been applied to a lot of existing models. After analysis of various machine learning algorithms, we found that Naive Bayes algorithm can be applied to classification tasks. It can perform the classification of complex datasets as well while providing high accuracy. Although it can perform sophisticated classifications, it is easier to build and understand compared to other algorithms. Also, supervised learning models are a better choice than unsupervised learning models when it comes to tasks like classification. The proposed system aims at increasing the existing accuracy of the algorithm. It will act as a recommender system for the end user. In future the model can be applied to various data sets which can analyse the sentiments.

ACKNOWLEDGEMENTS

We take this opportunity to express gratitude towards all the people who have been a part of this project, right from the inital phases. We extend our gratitude towards our internal project guide, Mrs. Archana Shinde, who has been a source of great help whenever needed. We thank our external guide Mrs. Neha Chattopadhyay for guiding us throughout the project phases. Our Head of Department Prof. Abhay Adapanawar has also been very helpful and we appreciate the support he provided us with. We would like to convey our gratitude to all the teaching and non-teaching staff members of Information technology department, our friends and families for their valuable suggestions and support.

REFERENCES

[1] Thumbs up? Sentiment Classification using Machine Learning Techniques. BoPang and LillianLee, Shivakumar Vaithyanathan [IBM, Cornell University].
[2] Emotions in product reviews – Empirics and models. David Garcia, Frank Schweitzer. Chair of Systems Design, ETH Zurich.
[3] Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. Geetika Gautam and Divakar Yadav [Jaypee Institute of Information Technology].
[4] R.Liu,R.Xiong,and L.Song, ”A Sentiment Classification Method for Chinese Document,” Processed of the 5th International Conference on Computer Science and Education (ICCSE), pp. 918 – 922, 2010.

