INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue3)
Available online at www.ijariit.com

Product Rating Based On Review Using Data Mining

Anurag Manni
Shri Ramdeobaba College of Engineering and Management, Nagpur
mannia@rknec.edu

Naman Jaiswal
Shri Ramdeobaba College of Engineering and Management, Nagpur
jaiswalnn@rknec.edu

Nayan Jaiswal
Shri Ramdeobaba College of Engineering and Management, Nagpur
jaiswalns1@rknec.edu

Abstract: The Online Shopping Experience has opened the new ways of business and shopping. As E-commerce trend is growing rapidly, it is very difficult to get exact reviews for any product as most of the reviews are not as per the requirements. Multiple reviews have different sentiments as positive, negative or neutral; also some reviews include good rating but bad reviews and bad rating but good reviews. This creates conflict and becomes very difficult to assess any product. This system needs intelligent sentiment analysis that can calculate exact sentiment and give an exact rating in order to get a correct review for the product. This paper will help to understand importance of sentiment analysis, language processing concepts in order to provide better intelligence to e-commerce systems.

Keywords - Sentiment Analysis, Rating, Review, Language processing, e-commerce.

I.INTRODUCTION

Sentiment analysis is the task of identifying whether the opinion expressed in a text is positive or negative in general or about a given topic.For example: "I am so happy today, good morning to everyone", is a general positive text, and the text: "Django is such a good movie, highly recommends 10/10", expresses positive sentiment toward the movie, named Django, which is considered as the topic of this text. Sometimes, the task of identifying the exact sentiment is not so clear even for humans, for example in the text: "I'm surprised so many people put Django in their favorite films ever list, I felt it was a good watch but definitely not that good", the sentiment expressed by the author toward the movie is probably positive, but surely not as good as in the message that was mentioned above.

A. Domain Introduction

The project heavily relies on the techniques of "Natural language Processing" in extracting significant patterns and features from the large data set of reviews and on "Machine Learning" techniques for accurately classifying individual unlabeled data samples (reviews) according to whichever pattern model best describes them.

Language based features are used for modeling patterns and classification. Language based features are those that deal with formal linguistics and include prior sentiment polarity of individual words and phrases, and parts of speech tagging of the sentence. Prior sentiment polarity means that some words and phrases have a natural innate tendency for expressing particular and specific sentiments in general. For example, the word "excellent" has a strong positive connotation while the word "evil" possesses a strong negative connotation. So whenever a word with a positive connotation is used in a sentence, chances are that the entire sentence would be expressing a positive sentiment. Parts of Speech tagging, on the other hand, are a syntactical approach to the problem. It means to automatically identify which part of speech each individual word of a sentence belongs to noun, pronoun, adverb, adjective, verb, interjection, etc. Patterns can be extracted from analyzing the frequency distribution of these parts of speech (either individually or collectively with some other part of speech) in a particular class of labeled reviews.

Classification techniques can also be divided into two categories: Supervised vs. unsupervised and non-adaptive vs. adaptive/reinforcement techniques. The supervised approach is done when the system has pre-labelled data samples available and the system uses them to train this classifier. Training the classifier means to use the pre-labelled data sample to extract features that best model the patterns and differences between each of the individual classes, and then classifying an unlabeled data sample

according to whichever pattern best describes it. Unsupervised classification is done when the system does not have any labeled data for training. In addition to this, adaptive classification techniques deal with feedback from the environment. In this case, feedback from the environment can be in form of a human telling the classifier whether it has done a good or poor job in classifying a particular review and the classifier needs to learn from this feedback. There are two further types of adaptive techniques: Passive and active. Passive techniques are the ones which use the feedback only to learn about the environment (in this case this could mean improving this model for reviews belonging to each of the three classes) but not using this improved learning in this current classification algorithm, while the active approach continuously keeps changing its classification algorithm according to what it learns at real-time [1].

II.REVIEW OF LITERATURE

Sentimental Analysis basic objective is to classify an opinion according to a polar spectrum. The extremes on the spectrum usually correspond to positive or negative feelings about something, such as a product, brand, or person.

Like all opinions, the sentiment is inherently subjective from person to person, and can even be outright irrational. It's critical to mine a large and relevant sample of data when attempting to measure sentiment. No particular data point is necessarily relevant. It's the aggregate that matters. An individual's sentiment toward a brand or product may be influenced by one or more indirect causes; someone might have a bad day and review a negative remark about something they otherwise had a pretty neutral opinion about. With a large enough sample, outliers are diluted in the aggregate. Also, since sentiment very likely changes over time according to a person's mood, world events, and so forth, it's usually important to look at data from the standpoint of time.

As to sarcasm, like any other type of natural language processing (NLP) analysis, context matters. Analyzing natural language data is the problem of the next few decades. It's an incredibly difficult issue, and sarcasm and other types of ironic language are inherently problematic for machines to detect when looked at in isolation. It's imperative to have a sufficiently sophisticated and rigorous enough approach that relevant context can be taken into account. For example, that would require knowing that a particular user is generally sarcastic, ironic, or hyperbolic, or having a larger sample of the natural language data that provides clues to determine whether or not a phrase is ironic [2].

III. RELATED WORK

Lina Zhou et al., [6] investigated product review mining using machine learning and semantic orientation. Supervised classification and text classification techniques are used in the proposed machine learning approach to classify the product review. A corpus is formed to represent the data in the documents and all the classifiers are trained using this corpus. Thus, the proposed technique is more efficient. Though the machine learning approach uses supervised learning, the proposed semantic orientation approach uses "unsupervised learning" because it does not require prior training in order to mine the data. Experimental results showed that the supervised approach achieved 84.49% accuracy in three-fold cross validation and 66.27% accuracy on hold-out samples. The proposed semantic orientation approach achieved 77% accuracy of product reviews. Thus, the study concludes that the supervised machine learning is more efficient but requires a considerable amount of time to train the model. On the other hand, the semantic orientation approach is slightly less accurate but is more efficient to use in real time applications. The results confirm that it is practicable to automatically my opinions from unstructured data.

Bo Pang et al., [3] used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Experiments demonstrated that the machine learning techniques are better than human produced baseline for sentiment analysis on product review data. The experimental setup consists of product-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews. Features based on unigrams and bigrams are used for classification. Learning methods Naive Bayes, maximum entropy classification and support vector machines were employed. Inferences made by Pang et al., is that machine learning techniques are better than human baselines for sentiment classification. Whereas the accuracy achieved in sentiment classification is much lower when compared to topic based categorization.

Zhu et al., [4] proposed aspect-based opinion polling from free-form textual customers reviews. The aspect related terms used for aspect identification was learned using a multi-aspect bootstrapping method. A proposed aspect-based segmentation model segments the multi-aspect sentence into single aspect units which were used for opinion polling. Using an opinion polling algorithm, they tested on real Chinese restaurant reviews achieving 75.5 % accuracy in aspect-based opinion polling tasks. This method is easy to implement and are applicable to other domains like product or movie reviews. Jeonghee Yi et al., [5] proposed a Sentiment Analyzer to extract opinions about a subject from online data documents. Sentiment analyzer uses natural language processing techniques. The Sentiment analyzer finds out all the references on the subject and sentiment polarity of each reference is determined. The sentiment analysis conducted by the researchers utilized the sentiment lexicon and sentiment pattern database for extraction and association purposes. Online product review articles for digital camera and music were analyzed using the system with good results.

IV. METHODOLOGY

This system uses text mining algorithm in order to mine keywords. It takes a review of various users, based on the review. A database of sentiment based keywords along with positivity or negativity weight in the database and then based on these sentiments mined keywords in a user review, the review is ranked. This system is a web application where the user views various products and purchases products online and can give a review about the merchandise and online shopping services. The system will help many E-commerce enterprises to improve or maintain their services based on the customer review as well as to improve the merchandise based on the customer review.

Below is the figure that explains the actual working of the system.

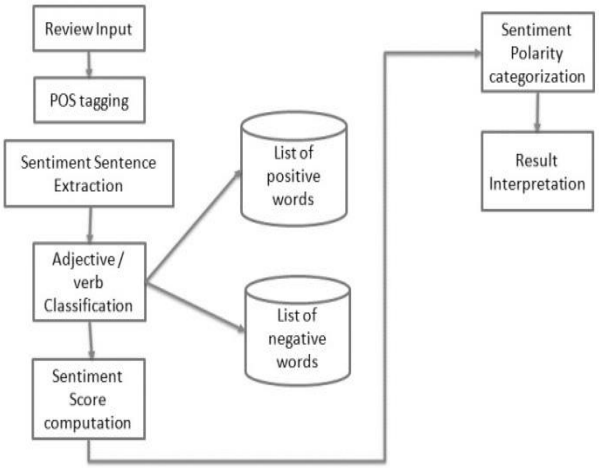

Fig. 1 Design Phase Details of the System

A. Calculation of Sentiment: (positive or negative)

The system reads the user reviews from data set and uses it for text processing. Text processing includes Part-of-Speech (POS) tagging and sentiment extraction which classifies it into adjectives. Following are the steps involved in the calculation of sentiment of a review:

1) *Duplicate Word Elimination:* Duplicate words in a sentence are removed in this step. A function is defined to remove duplicate words from user review. The function includes the regular expression which removes the continuously repeated word from the sentence.

   Regular expression used- \b (.+)(\s+\1\b

   Example sentence -"This Phone is very bad" Output- "this phone is very bad"

2) *Part of Speech Tagging:* Tagging is the process of automatic assignment of descriptors to the given tokens, the descriptor is a tag. The tag may indicate one of the parts-of-speech, semantic information, and so on. The process of assigning one of the parts of speech to the given word is called Parts Of Speech tagging. It is commonly referred to as POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories.

3) *Stemming:* Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

4) *Stop Words Elimination*: Stop words are words which are filtered out before or after processing of natural language data. Though stop words usually refer to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list.

5) *Removal of Antonym ('not + adjective'):* People review products with sentences such as for example: "The product is as such not bad". This „not" in the sentence with adjective reverses the polarity of the sentence and make it negative if it's from a positive point of view. Hence it becomes necessary to remove such negation words. Hence it is removed to make the sentence easier to understand to the classifier.

6) *Naive Bayes classifier:* Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.Input provided to naive Bayes classifier is a document with specified classes and training set of m-hand labeled documents. This set is the one which the system have to train; it may be set to positive, negative labeled sets. The output will be predicted class which will be either positive or negative based on training sets provided.

B. Rating each review according to the polarity

In this phase, the rating is done for each review and for the reviews stored beforehand as a result giving an overall rate of the product.

a) Create a .csv file (say pos.csv) for storing a list of positive words and number of times it occurred in the training data set.
b) Similarly, create a file for negative words (say neg.csv).

c) Tokenize the input text (as explained in the previous section)
d) Count the total number of positive words in pos.csv and similarly count a number of negative words in neg.csv. Store these values in variables Total_pos and Total_neg.
e) After tokenizing the input probability is calculated by comparing tokens to a total number of tokens in positive words file. The Same method is applied to calculate the negative probability.
f) At the end, the positive and negative rating is calculated by multiplying probability with a range within 1-5(frequency set) and rating is displayed along with review sentiment whether positive or negative.

CONCLUSION

In this paper, it is seen that sentiment analysis/opinion mining play a vital role in making a decision about product /services. Also, it is seen that soft computing techniques have not been extensively used in the literature. The work can be further extended to emerging areas like Mobile learning and investigation with soft computing techniques like a neural network.
The task of sentiment analysis is still in the developing stage and far from complete. So the system proposes a couple of ideas which the system feels are worth exploring in the future and may result in further improved performance.
In this research, the system is focusing on general sentiment analysis. There is the potential of work in the field of sentiment analysis with partially known context. For example, the system noticed that users generally use E-commerce websites for specific types of keywords which can divide into a couple of distinct classes, namely: products/brands, sports/sportsmen, and media/movies/music. So the system can attempt to perform separate sentiment analysis on reviews that only belong to one of these classes (i.e. the training data would not be general but specific to one of these categories) and compare the results the system get if the system applies general sentiment analysis on it instead.

ACKNOWLEDGEMENT

We sincerely thank Prof. A. R. Raipurkar for his continuous support, supervision motivation and guidance throughout the tenure of our project, who truly remained driving spirit in our project and his experience gave us light in handling project and helped us in clarifying abstruse concepts, requiring knowledge and perception, handling critical situations and in the understanding objective of work.
We also wish to express our sincere gratitude to Head of Department Dr. M.B. Chandak and Computer Science Department for providing an opportunity to learn new things in our field, to get experience in Project Development.

REFERENCES

[1]. Steven Bird, Edward Klein, Edward Loper;" Natural Language Processing with Python Analysing Text with the Natural Language Toolkit "; O' Reilly Media Publications, June 2009.
[2]. http://www.nltk.org/ (Accessed 26-10-2016 17:49)
[3]. Bo Pang, Lillian Lee, and ShivakumarVaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.
[4]. Zhu, Jingbo Wang, Huizhen Zhu, MuhuaTsou, Benjamin K. Ma, Matthew, "Aspect-Based Opinion Polling from Customer Reviews", IEEE Transactions on Affective Computing, Volume: 2, Issue:1 On page(s): 37. Jan-June 2011.
[5]. Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack: 2003, "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques", In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003). Melbourne, Florida
[6]. https://web.stanford.edu/class/cs124/lec/naivebayes.pdf (Accessed 02-11-2016 15:02)
[7]. Lina Zhou, PimwadeeChaovalit, "Movie Review Mining: a Comparison between Supervised and Unsupe

