ISSN 2395-1621

Report on Sentimental Analysis of Services

[#1]Atul Kamat, [#2]Snehal Chavan, [#3]Neil Bamb, [#4]Hiral Athwani
[#5]Prof. Shital A. Hande

[2]chavansnehal247@gmail.com

[#12345]Department of Computer Engineering,

Sinhgad Academy of Engineering, Kondhwa, Pune.

ABSTRACT

Reviews have been very valuable for many organizations to know about their performance. Efforts have been made to automate the task of analyzing sentiments using different techniques. Machine learning is a type of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The main work of machine learning techniques is building an algorithm that can take input data and use statistical approaches that predict an output value. Many approaches have been made to apply sentimental analysis on review data. Sentimental Analysis uses natural language processing to detect whether the particular text is positive, negative or neutral and is one of the most active research areas in natural language processing in recent years. Understanding the opinions behind user-generated content automatically is of great help for commercial and political use. The task can be conducted on different levels, classifying the polarity of words, sentences or entire documents. Generally, frameworks consists of extracting massive amount of genuine text reviews from a high level representation which captures the general sentiment clustering of sentences usually achieved through web scraping. Further, different machine learning techniques are used to generate a model. Recently, logistic regression has been proposed as an effective means for solving real time classification problems. The purpose of this paper is to survey a number of the more promising techniques.

ARTICLE INFO

Article History

Received: 10th January 2018

Received in revised form :

10th January 2018

Accepted: 12th January 2018

Published online :

13th January 2018

KEYWORDS: Machine Learning, Sentimental Analysis, Big Data

I. INTRODUCTION

Sentimental Analysis is the use of a natural language to extract information, predict patterns and determine the tone of the text. It has many applications in determining the reviews, getting the responses of online and social media. This kind of analysis has a huge importance in monitoring platforms as it allows the organizations to get the genuine opinion of the public. The applications of sentiment analysis are vast and dominant in many fields of work. By using Sentimental analysis, it quickly understands the positive or negative behavior of the reviews and provides the necessary information to the users. There are various approaches to sentiment analysis one of which is machine learning. Machine learning is a statistical approach that includes many algorithms which may be supervised or unsupervised. Machine learning algorithms such as naive baye's, clustering have been used in sentiment analysis of texts. Lexical approach is also one of many approaches which has been made. This includes maintaining of a dictionary of lexicons which are pre-tagged. Apart from using Machine Learning in Sentimental Analysis, front-end applications for representation of overall analysis are used. One such platform is Django. Django has been used by many others for creating applications for web. It is a software, which uses a database. It includes some kind of user interactivity, and operates through a web browser. A Framework provides a structure and common methods for making such a software.

II. RELATED WORK

From the past set of years, many articles, papers and books have been written on sentimental analysis. At the

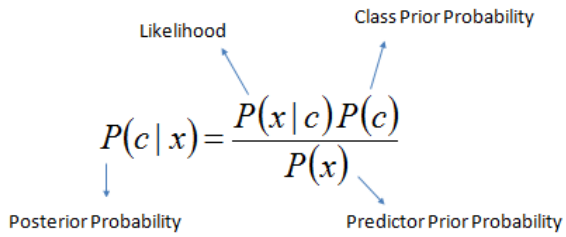
same time some researchers focus more on specific burden like finding the subjectivity expression, subjectivity clues, subjective sentence, topics, and sentiments of words and extracting sources of opinions, while others target is on assigning sentiments to whole document. All analyzers of sentiment analysis have adapted several methods to automatically predict the expression, sentiments of words or a document. The data set for sentimental analysis considered are movie, product review or social media data from the source of internet. They use pattern based approaches, Natural Language Processing (NLP) as well as machine learning techniques. Sentimental Analysis is a reference to the task of Natural Language Processing to determine whether a text contains subjective information and what information it expresses i.e., whether the attitude behind the text is positive, negative or neutral.

Many papers focus on several machine learning techniques which are used in analyzing the sentiments and in opinion mining. Sentimental analysis with the blend of machine learning could be useful in predicting the product reviews and consumer attitude towards newly launched products. These papers present a detail survey of various machine learning techniques and then compare them with their accuracy, advantages and limitations. On comparing it has been observed that 85% accuracy is attained by using supervised machine learning technique which is higher than that of unsupervised learning techniques.

Aspect based sentimental analysis was proposed [1] in which during experiment, four data sets were used to test the model. Authors proposed different approaches which bunch up the benefits of Senti WordNet, dependency parsing, and co reference resolutions are well organized for the purpose of sentimental analysis.

The Naive Bayesian Classification [2] represents a supervised learning method as well as a statistical method for classification. It is a probabilistic model that permits us to capture uncertainty about the model in a principled way by determining probabilities. It helps to solve diagnostic and predictive problems.

This Classification is named as Naïve Bayes after Thomas Bayes, who proposed the Bayes Theorem of determining probability. Bayesian classification provides useful learning algorithms and past knowledge and observed data can be combined. It helps to provide a useful perspective for understanding and also evaluating many learning algorithms. This helps to determine exact probabilities for hypothesis and also it is robust to noise in input data.



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

Fig 1

To implement Naïve Bayes algorithm a trained SentiWordNet [2] dictionary was taken which is available online. It consists of collection of different word with its synonym and its polarity. The synonym represents a word with similar meaning and also the same polarity. The polarity represents the positivity of the word in context of the sentence.

Two files are input to the mapper:

1) Dataset which contains the comments and review of the user.

2) SentiWordNet dictionary which contains the polarity of different words.

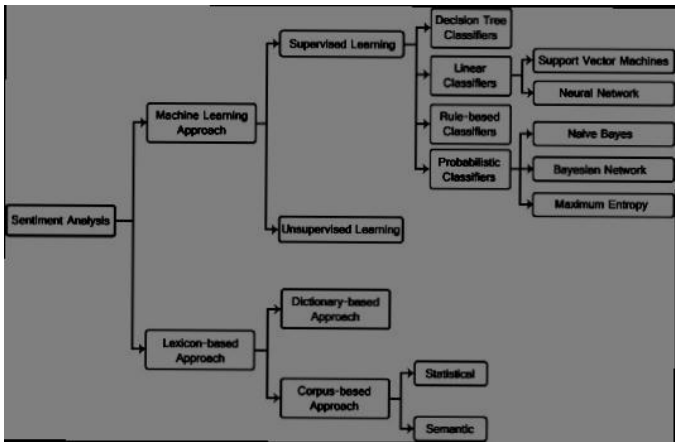


Fig 2

[3] Training data is most important part of the whole system as training of the system wholly depends on it and classification of testing data is done on the basis of this result only. While choosing training data, type of problem should be taken into consideration as similar type of training data should be taken so that it can provide more efficient and accurate results like if problem is related to a movie review then training data can be taken from IMDB or if problem is related to food review then it is better to use reviews of zomato or if problem is related to product review then data can be taken.

Emoticons are very useful symbols present in the text of Reviews of products, they emphasize the sentiments of

opinions and also help to find out the true sentiment of the text.

[4] Classification basically means categorizing data into different classes based on some computation which determines the sentiment behind the data. Number of classes depends on the type of problem. For example, for movie data the classes could be good , bad and average or simply positive or negative.
Many classifiers can be used for classification process like Naive Bayes classifier , Support vector machine , Baseline etc. Naive Bayes classifier has been used in previous papers for the classification process. Naive Bayes is mostly preferred for classification due to its speed and simplicity. Naive Bayes classifier assumes that the presence of a particular feature in a class is not related to the presence of any other feature . For example , a fruit may be considered to be an apple if it is red, round and about 4 inches in diameter. Even if these features depend on each other , all of these properties contribute to the property that this fruit is apple and thus the name "Naive". Mathematically , for a word w and class c , by Bayes Theorem

P(c/w)=[P(w/c)P(c)]/P(w) .....(1)

Where, P(c/w) is probability of class c given word is w.
P(c) is probability of class c and P(w) is probability of word w.

III. TRAINING A MACHINE LEARNING MODEL

Logistic Regression finds out tables of probabilities that are used to estimate the likelihood that the new data belongs to various classes. The probabilities are calculated using Bayes' theorem, which specifies how these events are related.

1. Bag of Words

Bag of words consist of unordered collection of words. Based on this, the Text classification can be done. The occurrence of each word is expressed as a feature for training a classifier .This occurrence is often referred to as the frequency of words. It also consists of stop words. Stop words are the list of words which are not required for training the classifier. These words are often used in the sentences and have no importance to determine the sentiment. For example 'a', 'an', 'the', 'for'.

2. Weighted Model Creation

For creating the bag of words model, the extracted words from a web scraper are considered. These words are broken down into tokens. The tokens are the simplest form of sentences which consists of one word. Let's consider a sentence 'The ambience is excellent'. This sentence gets broken down into tokens. The token form of the sentence would be 'The', 'ambience', 'is', 'excellent'. After this step, the stop words (unwanted words are removed from it).Finally, we have simplified set of tokens. These tokens are added into a dictionary which consists of unique words. For each of these words, the frequency is computed. Thus, we form a dictionary which consists of all the token words and their frequency. They are classified according to their sentiment (positive, negative and neutral). Considering the frequency i.e. the number of times they have occurred in a training set named as positive or negative class, each word is given some weight according to the model generated. This weight is considered as the polarity. The polarity consists of positive and negative values having a range from $+\infty$ to $-\infty$. Any token or word is more inclined towards the positive class if its value is greater than the some other token's value.

3. The Dataset

This weighed model is considered as the training data. The training data in logistic regression must be balanced. If the dataset is not balanced, then active learning approach is used for the text classification, which helps the system to overcome the problem of training data size. When the dataset is small, there is a possibility that the training data and testing data are not compatible and may give wrong results. Cross validation gives an average result but it doesn't help in training data selection. There is a definite statistical approach for building representative training dataset (training data selection). Once the training data is created, a real time text is given. This real time text is the review from the user. The polarity is then computed. It can update the service accordingly. There is no need of a management to look into each review to know about their performance. A system using machine learning can automatically find the review highlights from a set of reviews and update the information to the service after learning from this dataset.

4. Obtaining the Dataset

Web Scraping can provide us with a data set which can be used as the training task. Web scraping is a process of getting information from multiple web pages automatically.
A web scraper is basically a program that uses a Hyper-Text Transfer Protocol to send request and copying data from web servers to local databases. Since a web scraper fetches and downloads each page its main task is known as web crawling.
Web scraping is used in sentiment analysis to collect training sets and a web scraping program can be modified or

built to do so. A scraper for scraping data particular to a class, positive or negative in this case, can be coded to get data from websites which display reviews with stars. This training data collected is then arranged in separate databases according to their class. Work has been done on python with the famous Beautiful Soup library to scrape data which uses request function to access websites through HTTP. The HTTP page is then stored in a source code form. Specific tags from the source code can then be chosen to be stored on the database.

IV. SARCASM ANALYSIS

Research on classifying sarcastic words which cannot be classified based on their weights is an ongoing topic of interest for many scientists. Recognizing sarcasm is very much important for understanding people's genuine sentiment and beliefs. For example, if we consider a statement "The
journey was so good that everyone felt sick", will lead to a conclusion that the sentiment is positive towards the event of "felt sick".

1. Sarcasm as a Word Sense Disambiguation Problem

Here the task of checking if the 'sense' of a word is literal or sarcastic can be termed as [8] Literal or Sarcastic Sense disambiguation (LSSD). To tackle sarcasm authors have proposed a crowd-sourced task that relates to the task of creating a parallel database that groups words that can have a sarcastic meaning. Considering the above text, good in this context will be considered as bad or terrible and thus will be listed in the database.
Word embedding in a modified SVM achieve the best results among others.
The authors [8] also put forward unsupervised techniques to detect semantically opposite words or phrases. One of which is a co-training algorithm proposed by Barzilay and McKeown (2001). The co-training algorithm is used to remove paraphrases from a sentence. If we have a data base having sentences with tags named IM and SM where IM is the Intended meaning and SM Sarcastic meaning we can extract the opinion describing words as para-phrases.
Suppose,
SM1: The cycle was so good that it broke.
IM1: The cycle was so bad that it broke.
Here, the anchor words can be discarded and para-phrases can be found.

V. A FRONTEND

For an analysis to be useful a front-end is required. Sentiment analysis is an automated system to simplify our tasks. So, to present the processed information we make use of charts such as pie, bar or histograms. Different approaches have used different frameworks for this task.
Many planned to use the Django framework.
Django is a high-level web framework which was created for quick and transparent web project development.

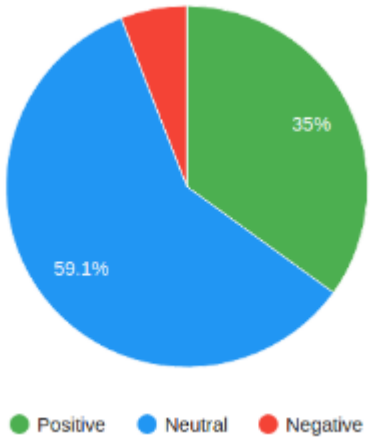


Fig 3

Django follows MVC-MVT pattern. MVC is the best way to design a client-server application. All of the best web frameworks are built around the MVC concept this design pattern is as follows:.
(M) Model: It is a representation of the data. It is an interface to the data and not the actual data. The model allows to pull data from a database without knowing the intricacies of the underlying database. It also provide an abstraction layer
with the database, because of that one can use the same model with multiple databases.
(v) View: The view is what the user can see. It's the presentation layer for the model. On a computer, view is a UI for a user's desktop app, or one can say, what we actually see on the browser for a website. The view also provides an interface to collect user input.
(C) Controller: The controller controls the flow of information between the view and the model. It uses a programmed
logic to decide what information is passed to the view and what information is pulled from the database via the model.
(T) Template: it is the presentation layer which contains the presentation related decisions, and how something should be displayed on a webpage.

VI. WHY NOT STARS

Sentiment analysis not only focusses on the user (customer), but also the service provider. The data to be reviewed or analyzed could be from any platform which may not be having a star rating associated with it. This makes the sentiment analyzer very useful to reduce a lot of human efforts. Feature extraction along with sentiment analysis can prove more useful than star based reviews as what feature or attribute in the natural language causing the opinion can be extracted at the same time.

VII. CONCLUSION

Sentiment analysis of reviews provides a generalized feedback of whether the review is positive or negative based on unbiased user reviews. It also suggests which particular product or service is better for a particular feature and its various pros and cons. Although many classifiers are available but Naive Bayes have been used because of its

speed. By use of word-net along with Naive Bayes for classification this accuracy can improved to a considerable extent as proposed by many authors in the above text.

REFERENCES

[1] Bhavitha B K, Anisha P Rodrigues, Dr. Niranjan N Chiplunkar, "Comparative study of Machine Learning techniques in sentimental analysis", 978-1-5090-5297-4/17/$31.00 ©2017 IEEE

[2] Ankur Goel , Jyoti Gautam , Sitesh Kumar "Real time sentiment analysis of tweets using Naïve bayes", International conference pp 235-244.

[3] J. Wang, K. Sasabe, and O. Fujiwara, "A siA. Kumar qnd T.M. Sebastian, "Machine Learning assisted Sentiment Analysis". Proceedings of International Conference on computer science and engineering (ICCSE'2012), 2012

[4] Hui Song, Yingxiang Fan and Xiaoqiang Liu ,"Extracting Product Features from online reviews for sentimental analysis".

[5] Christopher M. Bishop, Pattern Recognition and Machine Learning.

[6] Approaches for Sentiment Analysis on Twitter: A State-of-Art study.

[7] Janez Kranjcab Jasmina Smailovi Vid Podpe čanac Miha Grčara Martin Žnidaršiča Nada Lavrač, "Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the Clowd Flows platform".

[8] Debanjan Ghosh, Weiwei Guo and Smaranda Muresan, "Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words."

[9] "Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach."

[10] Star Ratings versus Sentiment Analysis - A Comparison of Explicit and Implicit Measures of Opinions

