Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 02, February -2018

Classification of Sentiment Analysis on Reviews using Machine Learning Techniques

Nilesh Thakkar[1]

[1]Computer Engineering, Swaminarayan College of Engineering & Technology, Kalol-382721, Gujarat

Abstract — *Now a days citizens can raise their opinions publicly and this opportunity has been provided due to the opened doors of internet which uses social media as platform for information exchange. Social media has huge amount of data related to people's opinion which includes reviews about products, movies, and blogs. With the use of Natural Language Processing we can perform sentiment analysis in data mining. Natural Language processing is also helpful in many more tasks related to information processing. In today's era everybody loves to post their views on social media about products, movies, and blogs in terms of opinions and reviews. People's opinion about any product is precious and it is very important for an organization to process it before making any decision. All fortune companies are focusing precisely on opinion mining considering the growth of social media. Using the techniques of sentiment analysis we can process the reviews and after completion of that process we can extract information. Analysis and categorization of the reviews can be done afterwards so that we can have mainly two categories which are positive and negative.*

Keywords*- Opinion Mining, Sentiment Analysis, Natural Language Processing, Naïve Bayes, Multinomial Naïve Bayes, Movie Review*

I. INTRODUCTION

Sentiment analysis and opinion minion [1] is open research field with manifold real life applications. Blogs, Forum, Twitter, Facebook and other resources on internet are put to use by humans for expressing their opinions. Sentiment Analysis has been addressed at different levels of granularity like at the document level, sentence level and many others. The social media has bought the people around the world closer; communication is one click away. Movies are perhaps the best entertainment mankind has got and it is very usual that people watch the movies and express their views and opinions on them by going online either in social networking sites or their own blogs. These type of reviews have a materialistic impact on the movie makers and even on the other people who tend to go to the movie. So, rather than reading the enormous content posted by users we can analyze the textual records by sentiment Analysis and conclude what is the overall impact movie has created in the people.

II. THEORETICAL BACKGROUND

"Sentiment Analysis" is the specialized branch of Data Mining Stream which deals with the classification of statuses or textual reviews into positive, negative and neutral as well[2]. Numerous research work has been already done in field of sentiment analysis. But the informal tone of tweets has always been a challenge for the analysis. Sentiment analysis has given way to a wide range of researches ranging from document level classification [3] to sentence level leading to phrases.

Nowadays, to buy any product people depend on the reviews given by other people on the websites, blogs or forums. The amount of user generated content is too large for a normal user to analyze. So to automate this, various sentiment analysis techniques are used. Symbolic techniques or Knowledge base approach and Machine learning techniques are the two main techniques used in sentiment analysis. Knowledge base approach requires a large database of predefined emotions and an efficient knowledge representation for identifying sentiments. Machine learning approach makes use of a training set to develop a sentiment classifier that classifies sentiments. Since a predefined database of entire emotions is not required for machine learning approach, it is rather simpler than Knowledge base approach.

Machine learning is about teaching computers to recognize patterns. It can be classified in two branches: Supervised learning and unsupervised learning.

Supervised learning: In this we build the classifier based on labelled data i.e. predefined set of categories. Here the value of a specific target variable is predicted from the values of the input variables, given a model of the relationship between the input and target variables. The drawback of supervised learning techniques is that the model, whatever its nature, can only be estimated if a training sample exists in which the values of the target variable are known.

Unsupervised Learning: The learning mechanism is built without using the labelled data. Unsupervised techniques are based on unlabeled data, i.e. categories of interest are not imposed on the studied data. This class of technique aims at discovering patterns in data and represents them in a low-dimensional form, often accompanied by visualizations that make them easier to interpret.

Training Data: In machine learning, a training set is a dataset used to train a model. In training the model, specific features are picked out from the training set. These features are then incorporated into the model. If the training set is labelled correctly, the model should be able to learn something from these features.
Test Data: The test set is a dataset used to measure how well the model performs at making predictions on that test set. In the case of sentiment analysis, a test set is a dataset of reviews that are distinct from the reviews in the training set. Training Sets and Test sets are the crux of machine learning. In order to make any prediction, we need some original dataset that our model can learn from, and we need a test set to see how well the model actually does at making predictions. Without a test set, there is no way to know whether or not our model over fits or under fits the training data- either scenario indicates that our model was not properly tuned to learn from the training data.

III. LITERATURE SURVEY

Sentiment analysis classification:

1. Document level of sentiment analysis: Opinions are the expressions, sentiments or notions towards a component or an event. Numerous data in net or gatherings permit individuals to express their assessment as surveys and remarks. At the point when opinions are communicated as surveys, rather than a straightforward Positive or Negative, recognizing the genuine opinions would require a subjective examination of the words utilized as a part of the survey.
2. Sentence level of sentiment analysis: This method use to give useful data when we search because the polarity of sentence will made perfect. In this level of sentiment analysis go through those sentences which contain opinions and gives reviews as though it is negative or positive.
3. Aspect based sentiment analysis: Document level and sentence level sentiment analysis functions admirably when they allude to a single element. Then again, a great part of the time people talk about components that have various points of view or qualities. The aspect based sentiment analysis focuses on the acknowledgement of all reports within a given record and points to which the feelings.
4. Comparative sentiment analysis: The user's utilization to express diverse emotions on items or brands. Either it may be same type of movie or by same production. The objective of comparative sentiment is to discover supposition of relative sort sentence.
5. Sentiment lexicon acquisition: Sentimental analysis is a process which utilizes data to find opinions and expressions of that data. In Sentiment analysis, we will see two kinds of classes positive and negative. Given a statement: "Auto X is superior to anything auto Y". This statement doesn't show which class is that statement falls in. Likewise, these sorts of sentences/documents are analyzed using two systems: Manual methodology, dictionary based approach.
    a. **Manual methodology:** It's totally time taking process because we can't retrieval the data is in positive or negative.
    b. **Dictionary based approach:** This approach utilizes sentiwordnet to find the polarity of that sentence by POS tagging.

POS Tagging: POS Tagging is extremely valuable in Opinion Mining procedure [4]. When we have to examine a document or a sentence first we need to concentrate the subjective data from the record or that specific sentence. POS Tagging helps us to find parts of speech of that word. Subsequent to extricating these words we can perform different activities on these and we can reach a conclusion. POS Tagging is done by utilizing the HMM model which used to tokenize and Tag the words furthermore for naming elements.

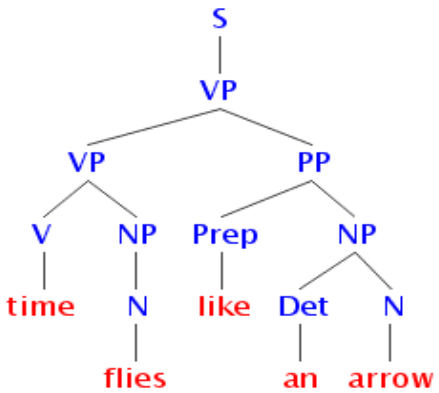

Figure 1. POS tagging

We take sentiment orientation (SO) of the examples are extracted. For instance we may have taken: Amazing + Phone which is: [JJ] + [NN] (or descriptive word took after by thing in human)
Sentiwordnet: Sentiwordnet is a lexical asset of opinion mining. Sentiwordnet allots a synset of word net in three scores: positive, negative, neutral. It extracts the parts of speech of that words that going hand in hand with we see the separated

words using POS Tagging from substance documents containing customer reviews. In the following figure the outline between number of reviews and number of words removed for diverse number of reviews.

Naïve Bayes Classifier : Naïve Bayes classifier [5] is based on the Bayesian theorem with the naïve assumption of independence between every pair of features. This classifier in spite of the apparently over-simplified assumptions has worked quite well in many real- world situations. It is very fast and has a good performance, better in some cases than more sophisticated methods. The Naïve Bayes model with Gaussian is equivalent to a mixture of Gaussians (GMM) with diagonal covariance matrices. The main advantages of this classifier are the conditional independence assumption, which helps to obtain a quick classification, and the probabilistic hypotheses

Bayes theorem can be computed by doing posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) does not depend on the values of remaining predictors. This is also known as conditional independence.

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

Where, $P(x|c) = P(x1|c) * P(x2|c) * \ldots\ldots * P(xn|c) * P(c)$

1. P (c|x) is the posterior probability.
2. P(c) is the prior probability.
3. P (x|c) is the likelihood.
4. P(x) is the prior probability of predictor.

Multinomial Naïve Bayes algorithm: Multinomial Naive Bayes [6] theorem is similar to naive Bayes algorithm except for the part that it is implemented for multinomial distributed data, and is a classic naive Bayes which is mainly used for text classification.

Natural Language Processing: The branch of computer science that researches on the development of systems that can communicate with humans in everyday language is called as Natural Language Processing [7]. In theory, it deals with the range of techniques that compute, analyze and represent naturally happening texts at multi-level analysis of languages for the purpose to make the machine process like human language for different disciplines and applications. NLP algorithms depend highly on machine learning with the majority being statistical. Older implementation of language- processing tasks normally required hard coding of big set of rules. By using machine learning, we can use normal learning algorithms usually in statistical inference, to learn rules by analyzing large corpora of real-world examples. A corpus is a set of documents that were annotated by hand with the correct values to be learned. Types of Natural Language Processing are:

1. Morphological processing:
2. Syntax and semantic analysis
3. Pragmatic analysis

IV. PROBLEM STATEMENT

The Objective is to design a mechanism to get reviews using machine learning technique is as follows.

- In today's era we are dumped with information including reviews, ratings, opinions from over communicative platforms like online social networking sites, forums, review sites, blogs, and micro blogs to share peoples' opinion on many products, services or reviews regarding any events, movies etc..
- This information, reviews or opinion consists so many things. Apart from their informative reviews, it also includes words considering all human emotions, repetitive words, negative / positive comments, and what not!
- It's very difficult task to find out straight forward information about any opinion on any product, service or reviews regarding any events, movies etc.
- Our objective is to elect the accurate decisive machine learning algorithm through which we can extract expressions of opinions describing a target features and classify its positive or negative polarity.

V. CONCLUSION

Social media such as Twitter and YouTube have been used for sharing contents and comments on all types of subjects by millions of people on a daily basis. It is clear that businesses have a strong interest in tapping into these huge data sources to extract information that might improve their decision making process. The topic of movies is of considerable interest in the social media user community. And for this researcher use machine learning technique to predicting revenues of movies. Machine learning techniques such as Naive Bayes classifier, multinomial naive Bayes classifier to know the sentiment of the reviews. But the main aim of this research work is to improve the level of accuracy by using best classifier from Machine learning techniques.

REFERENCES

[1] G. Kontaxis, I. Polakis, S. Ioannidis, and E.P. Markatos, "Detecting social network profile cloning, In Pervasive Computing and Communications Work-shops (PERCOM Workshops)", 2011 IEEE International Conference on, pages 295300, IEEE, 2011.

[2] Mr. B. Narendra, Mr. K. Uday Sai, Mr. G. Rajesh, Mr. K. Hemanth, Mr. M. V. Chaitanya Teja, Mr. K. Deva Kumar "Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies", IJISA.2016.

[3] Bo Pang and Lillian Lee, A sentimental education, "Sentiment analysis using subjectivity summarization based on minimum cuts, In Proceedings of the 42nd annual meeting on Association for Computational Linguistics", page 271, Association for Computational Linguistics, 2004.

[4] Davidov, Dmitry, Oren Tsur, Ari Rappoport,"Enhanced sentiment learning using twitter hashtags and smileys." Proceedings of the 23rd Inter-national Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010.

[5] T. Stein, E. Chen, and K. Mangla, "Facebook immune system" , In Proceedings of the 4th Workshop on Social Network Systems, SNS, volume 11, page 8, 2011.

[6] Huang, Jin, Jingjing Lu, and Charles X Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy," Data Mining, 2003, ICDM 2003,Third IEEE International Conference on. IEEE, 2003.

[7] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? sen-timent classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002.

