RESEARCH ARTICLE OPEN ACCESS

Mining Web Content Using Naïve Bayes Classification Analysis

K.S. Jeen Marseline
Head, Department of Information Technology,
Sri Krishna Arts and Science College, Coimbatore
TamilNadu - India

ABSTRACT

The enormous growth of web usage is creating splendid growth in information which leads to difficulty in extracting useful knowledge. The voluminous data available on online, the World Wide Web is a broad field for data mining research. Web mining helps to improve the exposure of web search engine by identifying the web pages and classifying the web documents. The web mining has few mining categories such as web content, web usage and web structure. In this paper, it focuses on web content mining which do text mining on the user's feedback from the website. The Content mining here, does extract the essential information from the web page content and shows that whether the content has positive or negative impact most. In this, have concluded the result of accuracy using Naive Bayes classifier algorithm.

***Keywords*:-** Text Mining, Naïve Bayes classification

I. INTRODUCTION

In vast area world wide web, huge amount of contents is available in different format like text, videos, image and audio. Those data are publicly and also privately available and that constantly gets growing, a large number of texts expressing opinions are available in review sites, forums, blogs, and social media. With the help of sentiment analysis systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, or any topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, product reviews, product feedback and customer service. That comes under field named Sentiment Analysis also known as Opinion Mining, that builds systems that try to identify and extract opinions within text. Currently, sentiment analysis is a topic of great interest and development since it has many practical applications. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. Here, identifying or discovering patterns from large data sets and these patterns enable you to predict user opinions. The user feedbacks which have been classified and analyzed by using naïve bayes classification algorithm.

II. NAÏVE BAYES CLASSIFICATION

Naive bayes is a popular algorithm for classifying text. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. The classifier is a straightforward and powerful algorithm for the classification task. Multinomial Naive Bayes is a specific instance of Naive Bayes where the P (Feature*i*| Class) follows multinomial distribution like word counts, probabilities, decision making. The general term Naive Bayes refers the strong independence assumptions in the model, rather than the particular distribution of each feature. A Naive Bayes model assumes that each of the features it uses are conditionally independent of one another given some class. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of features in a learning problem. The classifier often performs much more complicated solutions. A naive bayes classifier works by figuring out the probability of different attributes of the data

being associated with a certain class. To understand the naive bayes classifier, need to understand the Bayes theorem. So let's first discuss the Bayes Theorem.

The formula for calculating the conditional probability, is shown below:

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)}$$

Here,

- P(H) is the probability of hypothesis H being true. This is known as the prior probability.
- P(E) is the probability of the evidence.
- P(E|H) is the probability of the evidence given that hypothesis is true.
- P(H|E) is the probability of the hypothesis given that the evidence is there.

Even though, there is a data set with millions of records with some attributes, it can evaluate by the Naive Bayes approach. The data set available publicly that can be used for the process.

III. IMPLEMENTATION

Even if we are working on Naive Bayes classifier gives great results when we use it for textual data analysis. Such as Natural Language Processing. It is special case of text mining generally focused on identifying opinion type in score, it doesn't give very accurate but it is useful. The classification focuses on two sentiments that is positive and negative. This work has implemented in Python language which runs on emerging platform named Anaconda. All of the nltk classifiers work with *featstructs*, which is to name the feature name and value. For the text classification, can use the bag of words where each word has a feature name and its value. In this, have taken datasets of user's feedback from publicly available data. The easiest way to generate features from text is to split the text up into words. Each word in the feedback will then be a feature that have worked with. In order to do this, first split the sentences based on whitespace. Then count up how many times each word occurs in the negative feedback and how many times each word occurs in the positive feedbacks. This will allow to eventually compute the probabilities of a new sentences belonging to each class. Now that the word counts, to convert them to probabilities and multiply them out to get the predicted classification. And, have done training and testing part which is from the whole collected datasets 70% texts will be trained and balance 30% will be for the testing part. To predict the probabilities on the user's feedback in test.csv and train.csv, import the file because the probabilities were generated from it and the algorithm has prior knowledge about the data it's predicting on. Give the label as, positive and negative as classification label, get to know the accuracy have used the nltk.classify.util.accuracy along with the coding specification. To do all this, we'll need to compute the probabilities of each class occurring in the data, and then make a function to compute the classification.

```
Feedback: And I have only utilized
staffing agency couple of times in
the past much to no available, one
of their service was absolutely
AMAZING!! They were energized and
efficient. To my surprise my
experience there was awesome. In
addition, I received all kind
support. No cost of effective and
lots of firm similar like them,
they do better this and no proper
maintain. Needs more work.
Negative prediction:
2.0050503623505e-221
Positive prediction:
8.3071546676793e-226
```

Fig.1

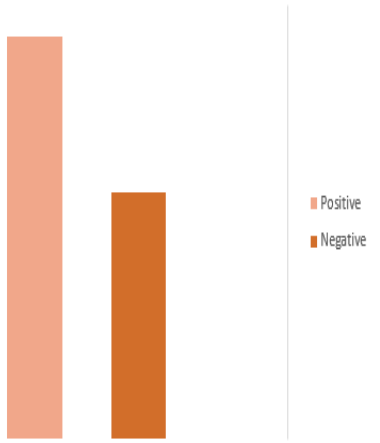

Fig2

The Fig.1 shows the sample output of the work, the user's feedback which has separated as positive and negative. Also which gives the accuracy score for both the positive and negative feedbacks. The user feedbacks which have been calculated and the accuracy scores have plotted as graph into positive and negative. It is implemented to know overall opinion of the people from their feedback.

According to the plot (Fig.2), by working with naïve bayes classifier with the complete user feedbacks datasets which conclude the overall opinion of the people are positive which we could identify by the accuracy calculation that attain from naïve bayes classifier.

IV.CONCLUSION

The decision has been made through the complicated feedback from number of people, those are analyzed and the voluminous data of people classified accordingly to feature label as positive and negative. The user feedback on a website, the datasets acquired from the data site and does text classification through the naive bayes classifier. That classifies the sentiment from the sentences as positive and negative by the features. After attain the accuracy scores of both the sentiment and then perform plot to know overall sentiment from the people's feedback.

REFERENCES

[1] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[2] Liu, B., Zhang, L., 2012. A Suvery of opinion mining and sentiment analysis, in: Aggarwal, C.C., Zhai, C. (Eds.), Mining Text Data.Springer US, Boston, MA. chapter 13, pp. 415–464.

[3] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in Information Communication and Embedded Systems (ICICES), 2013 International Conference on. IEEE, 2013, pp. 271–276.

[4] Analyzing social media remark using sentimental analysis, T.Devishree, K.S.Jeen Marseline, 2018 IJRTI , Volume 3, Issue 6.

[5] The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Author links open overlay panelMika V.Mäntylä[a]DanielGraziotin[b]MiikkaKuutila[a]