International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

RECOMMENDATION SYSTEM USING NAIVE BAYES CLASSIFIER

Pandian C. V[1] ,Venkatesh Pandian P. V[2], Vishal Kumar K.A V[3], Sachin Bharathi M V[4]

[1]Mr. Pandian C, Dept. of Information Technology, K.L.N. College of Engineering, Tamil Nadu, India
[2, 3,4]Student, Dept. of Information Technology, K.L.N. College of Engineering, Tamil Nadu, India

---***---

Abstract - *Sentimental Analysis is mainly meant for classifying the text. Opinion Mining is one of the major categories in sentimental analysis. Opinion of any user in buying a product or rating a movie contributes highly to the product or movie. For example: If a movie is given various levels of star rating, then the other audience who think of going for the movie might have a overview on the rating and then decide whether to prefer the movie or not. In this project Naive Bayes classifier technique is being used for the purpose of classification. The dataset used here is the youtube video comment dataset obtained from the famous youtube videos. Every record in dataset is being analyzed using the Naive Bayes approach, this is a probabilistic approach and finally the result is displayed in the form of list. Finally, the recommendation system is built based on classified positive, negative comments.*

***Key W Keywords*: Naive Bayes, Sentimental Analysis, User Comments, Recommendation System.**

1. INTRODUCTION

Sentimental analysis is one of the important fields in text mining. The varying thoughts of different users in any of the particular category are gathered as a single dataset and are being considered for the analysis. Generally, the analysis could be done with the help of any of the machine learning techniques. It helps in effective classification of the text.

Companies are receiving more information than ever before via social media, surveys, online reviews, emails, and other channels. But, did you know that 80% of this data is unstructured? So, if you want to analyze this data to get insights about your customers opinions, you'll need to sort it first.

For example, let's imagine you just launched a new product and carried out a survey to see what your customers thought. The survey was really successful, and you received 1000 open-ended responses. That's great, but you have limited resources, so it's gonna take forever to read and analyze them.

Performing sentiment analysis, however, can save you time, effort, and resources when analyzing your survey responses. It's far more efficient than manually sorting data because it's completely automated, meaning your team can focus on more urgent and important tasks.

Sentimental analysis Working

These systems don't rely on manually crafted rules, but on machine learning techniques, such as classification. Classification, which is used for sentiment analysis, is an automatic system that needs to be fed sample text before returning a category, e.g. positive, negative, or neutral.

There are two stages involved in implementing automatic systems:

1. Training
2. Prediction

In the training stage, a sentiment analysis model learns to correctly tag a text as *negative, neutral* or *positive* using sample data. The feature extractor then transforms the text into a feature vector, creating pairs of feature vectors and tags (e.g. positive, negative, or neutral) that are fed into the machine learning algorithm to generate a model.

In the prediction process, the feature extractor is used to transform unseen text into feature vectors, which are fed to the model, enabling it to make sentiment predictions.

Applications of Sentimental Analysis:

- Facebook analysis
- News Analysis
- Blog Analysis
- Movie Analysis, etc

Methods used in sentimental analysis:

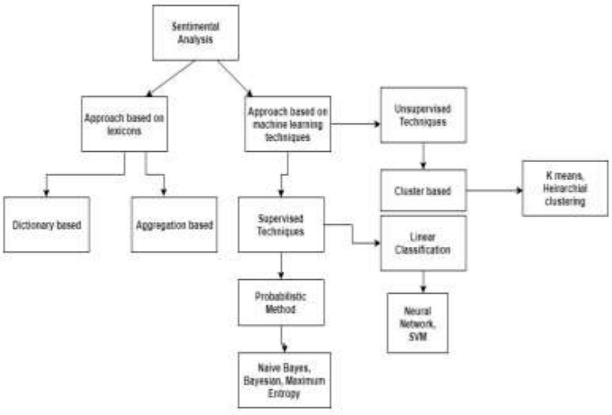

Fig. 1. Available methods for analysis

2. RELATED WORK

Bo pang proposed that Sentiment analysis seeks to identify the view point underlying a text span. It is classifying a movie review as “thumbs up” or “thumbs down”. This sentiment polarity proposed a novel machine-learning method that applies text-categorization techniques to just the subjective portions of the document extracting these portions can be implemented using efficient techniques for finding minimum cuts in graphs. This greatly facilitates incorporation of cross-sentence contextual constraints.

Abbsai concern when incorporating large sets of diverse n-gram features for sentiment classification is the presence of noisy, irrelevant, and redundant attributes. These concerns can often make it difficult to harness the augmented discriminatory potential of extended feature sets and they proposed a rule-based multivariate text feature selection method called Feature Relation Network that considers semantic information and also leverages the syntactic relationships between n-gram features. FRN is intended to efficiently enable the inclusion of extended sets of heterogeneous n-gram features for enhanced sentiment classification. Experiments were conducted on three online review test beds in comparison with methods used in prior sentiment classification research. FRN outperformed the comparison univariate, multivariate, and hybrid feature selection methods; it was able to select attributes resulting in significantly better classification accuracy irrespective of the feature subset sizes. Furthermore, by incorporating syntactic information about n-gram relations, FRN is able to select features in a more computationally efficient manner than many multivariate and hybrid techniques.

Prem Melville proposed the explosion of user-generated content on the Web has led to new opportunities and significant challenges for companies, that are increasingly concerned about monitoring the discussion around their products. Tracking such discussion on weblogs, provides useful insight on how to improve products or market them more effectively. An important component of such analysis is to characterize the sentiment expressed in blogs about specific brands and products. Sentiment Analysis focuses on this task of automatically identifying whether a piece of text expresses a positive or negative opinion about the subject matter. Most previous work in this area uses prior lexical knowledge in terms of the sentiment-polarity of words. In contrast, some recent approaches treat the task as a text classification problem, where they learn to classify sentiment based only on labeled training data. In this paper, we present a unified framework in which one can use background lexical information in terms of word-class associations, and refine this information for specific domains using any available training examples. Empirical results on diverse domains show that our approach performs better than using background knowledge or training data in isolation, as well as alternative approaches to using lexical knowledge with text classification.

Vasileios Hatzivassiloglou proposed the Subjectivity is a pragmatic, sentence-level feature that has important implications lhr text processing applications such as information extraction and information retrieval. We study the effects of dynamic adjectives, semantically oriented adjectives, and gradable adjectives a simple subjectivity classifier, and establish that they are strong predictors of subjectivity. A novel trainable method that statistically combines two indicators is presented and evaluated, complementing existing automatic techniques for assigning orientation labels.

3. MODULE DESCRIPTION

3.1 Dataset Collection (Youtube comment dataset)

The dataset is being collected from the online repository and stored in the local storage. The dataset collected for this work is the youtube comment dataset. The collected dataset is then converted into a CSV file and then it is further used for classification. Then the dataset is being imported and applied for the classification process.

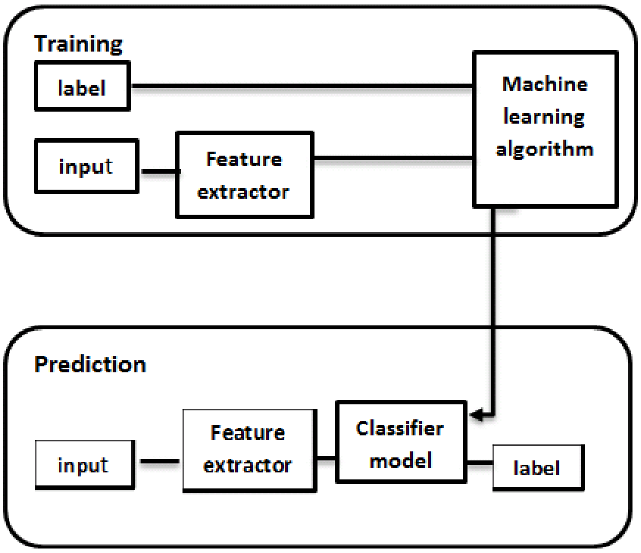

Fig-1: Architecture of machine learning approach

3.2 Preprocessing (Removing punctuation, stop words)

The pre-processing is being done to make the data still more efficient for the classification. Pre-processing or data clean-up is very much necessary in order to avoid the overloading and storage problem, Since the dataset consists of many information, and hence before the processing the unnecessary values and other symbols which is not at all required for classification purpose is being removed.

Since the dataset consists of many information, unwanted information for classification is being removed. This includes:

- Removal of Stop words
- Punctuations & symbols Removal
- Removal of numerical values

Labels	Count
Positive count	100
Negative count	50
Over positive count	20
Over negative count	30

Table -1: Results after classification

3.3 Review Classification

The review set is classified by the help of Naive Bayes Algorithm. Reviews are separated into single words or tokens. The training is done with the provided training set and a feature set is created. Feature set consists of the frequency of every word in review. Feature set is created only for training set. Based on the occurrences of the tokens in the feature set, positive and negative probability of the review is calculated using the following formula

- P(A/B) = P(B/A) P(A) / P(B)

P(A/B) – Posterior Probability

P(B/A) – Likelihood

P(A) - Prior Probability.

The higher probability is chosen as the label of the review based on the four label. Four labels are Positive, Negative, Over Positive, Over Negative. All the reviews are classified and count for each label is determined for building recommendation system.

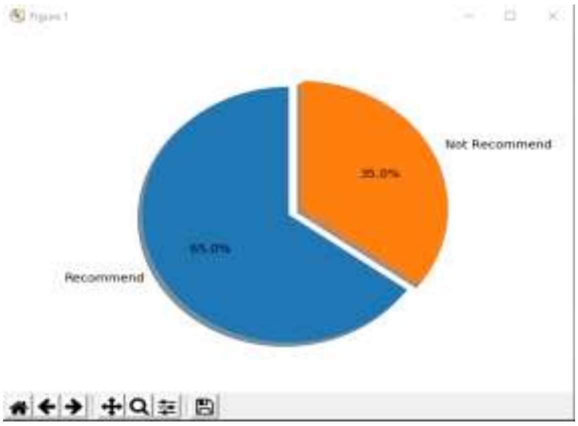

Fig-2: Output of recommendation system.

3.4 Score Aggregation

To build a recommendation system for the reviews, we made a method that gives score for each labels to calculate the recommendation percentage. We allot 1 for positive review, 0 for negative review, 1.5 for over positive, - 0.5 for over negative. By considering the unique score for each labels, it is efficient to develop a recommendation system. Recommendation system built using this method finally results in recommendation percentage of the video.

4. FUTURE WORK

Future studies should focus on analysis for sarcastic comments as it is tougher to predict and reviewers sometimes tries to express their emotions through emoji and it can be implemented as future work. It can also be extended to classifying the review comparing both the emoji and the actual review. This also needs to concentrate on identifying the silly spelling mistakes that user makes, it should be trained to autocorrect those spelling mistakes and finally analyzing the short version of words will also used to improve the accuracy of our model prediction.

5. CONCLUSION

The classifier is highly scalable, requiring a small number of parameters. The total number of positive and negative reviews in the given review set is calculated using Naive Bayes Classifier. Naive Bayes is a simple classifier technique. There are many other machine learning techniques available, Naive Bayes is one of the efficient methods which provides better level of accuracy and good results after classification using statistical approach and probabilistic techniques. The proposed approach could be applied to any kind of review dataset to analyse the sentiment of. The work could also be extended for higher level of input set.

REFERENCES

[1]”SentiFul: A Lexicon for Sentiment Analysis” Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka, Member, IEEE Transations on Affective Computing, vol 2, no. 1, Jan-Mar 2011

[2] “Sentiment Classification using Machine Learning Techniques” Bo Pang and Lillian Lee Department of Computer Science Cornell University Ithaca, NY 14853 USA Shivakumar Vaithyanathan IBM Almaden Research Center 650 Harry Rd. San Jose, CA 95120 USA

[3]A. Abbasi, S. France, Z. Zhang, and H. Chen, “Selecting attributes for sentiment classification using feature relation networks,” IEEE Trans. Know. Data Eng., vol. 23, no. 3, pp. 447–462, Mar. 2011.

[4] “Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification” Prem Melville IBM T.J. Watson Research Ctr. P.O. Box 218 Yorktown Heights, NY.

[5] “Effects of Adjective Orientation and Gradability on Sentence Subjectivity” Vasileios Hatzivassiloglou Department of Computer Science Columbia University New York, NY 10027, Janyce M. Wiebe Department of Computer Science New Mexico State University Las Cruces, NM 88003.

[6]VandanaJha, Savitha.R ,P.DeepaShenoy, ArunKumarSangaiah, Venugopal KR “ A novel sentiment aware dictionary for multi domain sentiment classification” Journal of Computers and Electrical Engineering”, Volume 69, July 2018, Pages 585-597.

[7] Xiaojiang Lei, Xueming Qian, Guoshuai Zhao “Rating Prediction based on social Sentiment from textual reviews”, IEEE Transactions On Multimedia, Volume:40, Issue:4, June 2017.