International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 3, May-June 2016 **ISSN 2278-6856**

Analytical View of Sentiment Analysis on Unsupervised Big Data

[1]Ran vijay singh,[2]Mr. Sheenu Rizvi

[1]Amity University Uttar Pradesh Lucknow

[2]Assistant Professor Amity University Uttar Pradesh Lucknow

Abstract

Sentiment examination can be performed at three diverse levels: document, sentence and aspect level. The document level sentiment analysis goes for arranging the whole record as positive or negative. The point of study is to assess the execution for sentiment classification in wording of accuracy, precision and recall. The sentence level sentiment investigation is firmly identified with subjectivity examination. Here machine learning classification approaches with different feature selection schemes to obtain a sentiment analysis model for the movie review dataset has been explored. Numerous analyses were done utilizing distinctive capabilities and parameters to acquire greatest accuracy. In the present work machine learning techniques are used to detect the sentiments of movie reviews.

1. Introduction

The advancement of web innovation has prompted a tremendous measure of client generated content and has significantly changed the way we manage, organize and interact with information. Because of the expansive measure of client suppositions, audits, remarks, criticisms and proposals it is essential to explore, analyze and organize the content for efficient decision making. In the previous years opinion investigation has developed as one of the well known strategies for information retrieval and web data analysis. Sentiment analysis, also known as opinion mining is a subfield of Natural Language Processing (NLP) and Computational Linguistics

(CL) that defines the area that studies and analyses people’s opinions, reviews and sentiments.

Sentiment analysis characterizes a procedure of separating, distinguishing, breaking down and portraying the estimations or sentiments in the form of textual information using machine learning, NLP or statistics. A fundamental feeling examination framework performs three noteworthy errands for a given document. Firstly it recognizes the sentiment communicating part in the report. Secondly, it identifies the sentiment holder and the entity about which the sentiment is expressed. Finally, it identifies the polarity (semantic orientation) of the sentiments.

Sentiment examination can be performed at three diverse levels: document, sentence and aspect level. The document level sentiment analysis goes for arranging the whole record as positive or negative, (Pang et al, [2]; Turney, [3]). The sentence level sentiment investigation is firmly identified with subjectivity examination. At this level each sentence is analyzed and its opinion is determined as positive, negative or neutral, (Riloff et al, [4]; Terveen et al, [5]). The aspect level sentiment analysis aims at identifying the target of the opinion. The basis of this approach is that every opinion has a target and an opinion without a target is of limited use, (Hu and Liu, [6]).

In this paper we explored the machine learning classification approaches with different feature selection schemes to obtain a sentiment analysis model for the movie review dataset.

Natural Language Processing and Machine Learning approaches were used for the process. Numerous analyses were done utilizing distinctive capabilities and parameters to acquire greatest accuracy.

2. Data Used

The proposed work is evaluated by running experiments with the polarity dataset V2.0, available at http://www.cs.cornell.edu/people/pabo/movie-review-data. Sentiment model has been built using supervised learning. For this a set of 200 movie review data available from Pang and Lee at Cornell University has been used. It has 200 positive reviews, 200 negative reviews and 200 unlabelled reviews for testing of the model.

3. Problem Statement and Proposed Technique

This area displays the proposed method to analyze sentiments in a movie domain. The proposed approach utilizes a mix of NLP procedures and administered learning. In the primary stage a pre-preparing model is proposed to optimize the dataset. In the second stage tests are performed utilizing the machine learning techniques to acquire the performance vector for various feature selection schemes.

Here Rapid Miner Studio 6.0 software with the text processing extension, licensed under AGPL version3, and Java1.6 has been used. Rapid Miner supports the design and documentation of overall data mining process. Model implementation has been carried out using Decision Tree Machine learner.

First step for implementing this analysis is Pre-Processing the document from data i.e. extracting the positive and negative surveys of a movie and putting away it in various polarity. At to start with, both positive and negative surveys of a certain movie are taken. The greater part of the words are stemmed into root words. At that point the words are put away in various extremity (positive and negative). Both vector wordlist and model are made.

Next, the required list of movies, an unlabelled data, as input to the model validation. Model analyzes every last word from the given list of movies with that of words which come under different polarity stored earlier. The movie survey is evaluated in view of the dominant part of number of words that occur under a polarity.

4. Stages in Sentiment Analysis of Movie Review Data

A basic task in sentiment analysis is classifying an expressed opinion in a document, a sentence or an entity feature as positive or negative. The work here gives the list of movies and its review such as Positive or Negative. This program implements Precision and Recall method. Precision is the probability that a (randomly selected) retrieved document is relevant. Recall is the probability that a (randomly selected) relevant document is retrieved in a search. Or high recall means that an algorithm returned most of the relevant results. High precision means that an algorithm returned more relevant results than irrelevant.

STEP -I

The learning processes usually optimize the model they generate to make it fit the training data as well as possible. If we test this model on some independent set of data, mostly this model will not perform that well as it performed on the training data that generated it. This is called over-fitting. The X-Validation operator predicts the fit of a model to a hypothetical testing data, which can be especially useful when you don't have separate testing data.

STEP – II

Here both the model and vector wordlist obtains in Step-I are retrieved from the repository using store operators. Output from the retrieved wordlist is fed to the process document operator. Further, the Apply Model operator takes a model from a Retrieve operator and unlabeled data from Process document as input and outputs the applied model to the 'lab' port, which has been connected to the Report (result).

5. Result Analysis

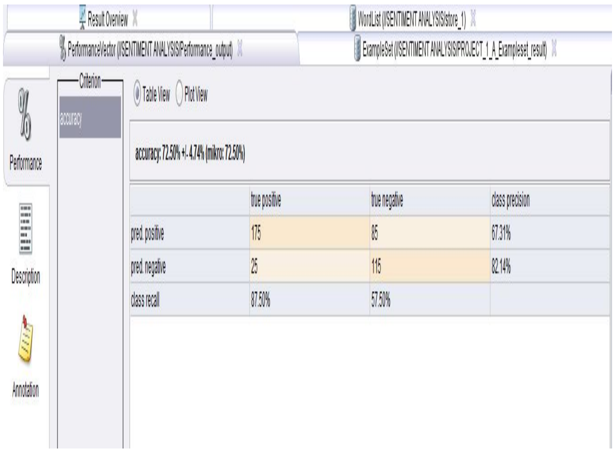
The dataset consists of 200 reviews equally divided into 100 positive and 100 negative. The dataset used for the experiments was divided into two classes, positive and negative. For a given classifier and a document there has four possible outcomes: true positive, false positive, true negative and false negative. If the document is labelled positive and is classified as positive it is counted as true positive else if it is classified as negative it is counted false negative. Similarly, if a document is labelled negative and is classified as negative it is counted as true negative else if it is classified as positive it is counted as false positive. Based on these outcomes a two by two confusion matrix can be drawn for a given test set. This is shown in Figure 1 below.

The confusion matrix in figure 1 forms the basis for the calculation of the following metrics.

i. *Accuracy* = (tp+tn)/ (P+N)
ii. *Precision* = tp/ (tp+fp)
iii. *Recall/ true positive rate* = tp/P
iv. *F-measure* =2/ ((1/precision)+(1/recall))
v. *False alarm rate/ false positive rate* = fn/N
vi. *Specificity* = tn/ (fp+tn) = (1-fp rate)

The experiments show that Term frequency-Inverse document frequency (TF-IDF) scheme gives maximum accuracy using Decision Tree classification approach.

From the Decision Tree model built from our dataset with sentiment attributes, we can see that the most important decision attributes were WORST, which is one of the attributes detected from the set of negative and positive sentiment datasets.


Result Overview
WordList (//SENTIMENT ANALYSIS/store_1)
PerformanceVector (//SENTIMENT ANALYSIS/Performance_output)
ExampleSet (//SENTIMENT ANALYSIS/PROJECT_1_A_Exampleset_result)
Performance
Description
Annotation
Criterion
accuracy
Table View
Plot View

accuracy: 72.50% +/- 4.74% (mikro: 72.50%)

	true positive	true negative	class precision
pred. positive	175	85	67.31%
pred. negative	25	115	82.14%
class recall	87.50%	57.50%	

Figure 1: Snapshot of the Model Performance using Decision Tree Classification approach

From the Performance operator, we get various measures of the sentiment dataset, as seen from Figure 1 above. The various performance measures are as mentioned below.

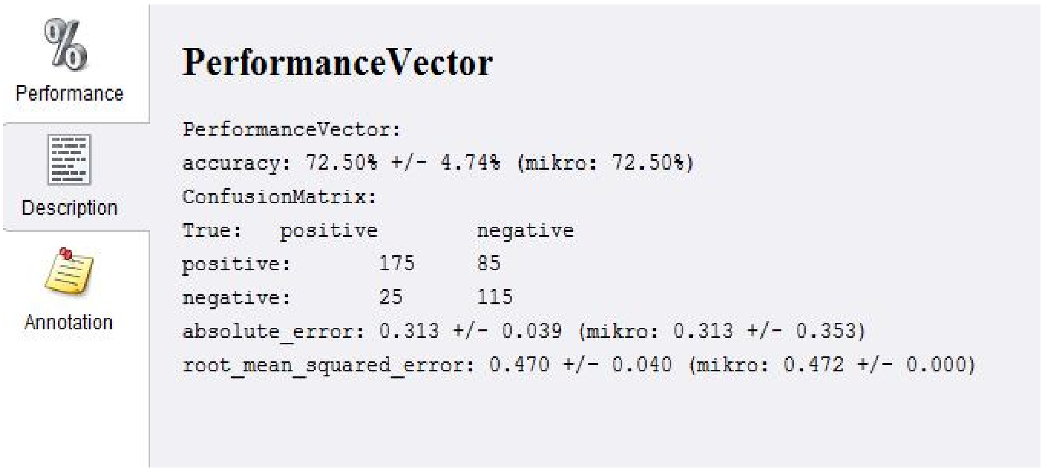
Accuracy is calculated by the percentage of correct predictions over the total number of examples. Correct prediction means examples where the value of the prediction attribute is equal to the value of the label attribute.

Precision of a class is calculated by taking the correct predictions of a label's value over the total predictions for the same label value (correct predictions + wrong predictions).

Recall of a class is calculated by taking the correct predictions of a label's value over the total of the real examples with the same label value (correct predictions + missed examples).

The performance of this decision tree has a better accuracy (Fig. 1 & 2), and especially the recall of the POSITIVE labelled comments (which are much less compared to the

NEGATIVE labelled comments) is 87.5%. Also the model performance has RMSE value of 0.47 and Absolute error of 0.313, which again clearly demonstrates the good predictive capability of the model.



PerformanceVector

```
PerformanceVector:
accuracy: 72.50% +/- 4.74% (mikro: 72.50%)
ConfusionMatrix:
True:   positive        negative
positive:       175     85
negative:       25      115
absolute_error: 0.313 +/- 0.039 (mikro: 0.313 +/- 0.353)
root_mean_squared_error: 0.470 +/- 0.040 (mikro: 0.472 +/- 0.000)
```

Figure 2: Snapshot of the performance of the model

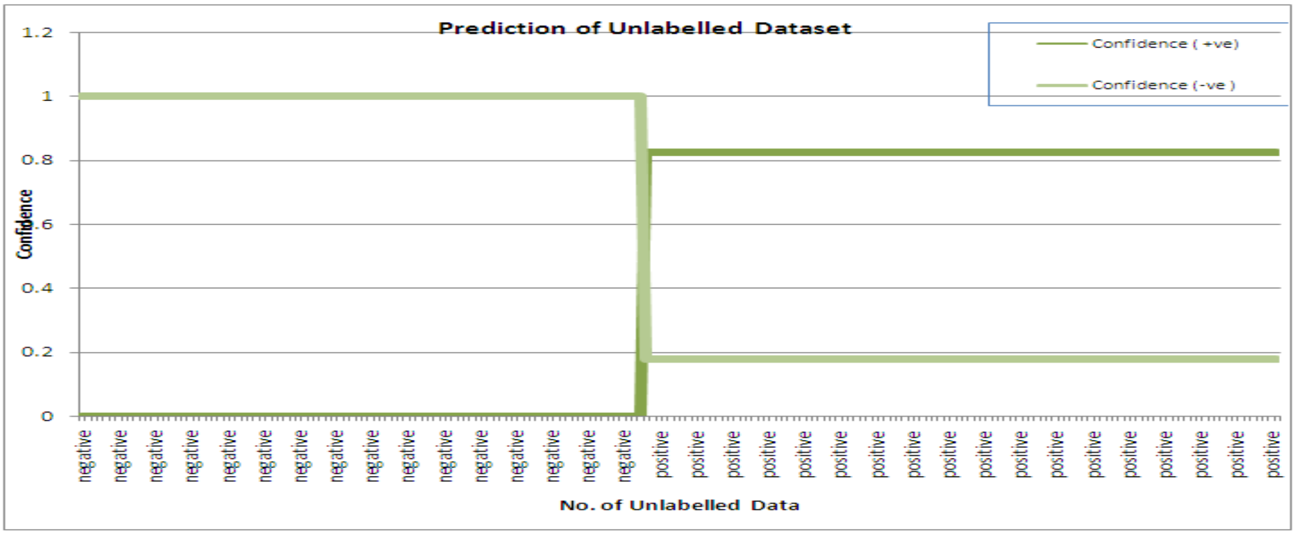


Figure 3: Graphical Plot of Sentiment Analysis of Unlabelled Data

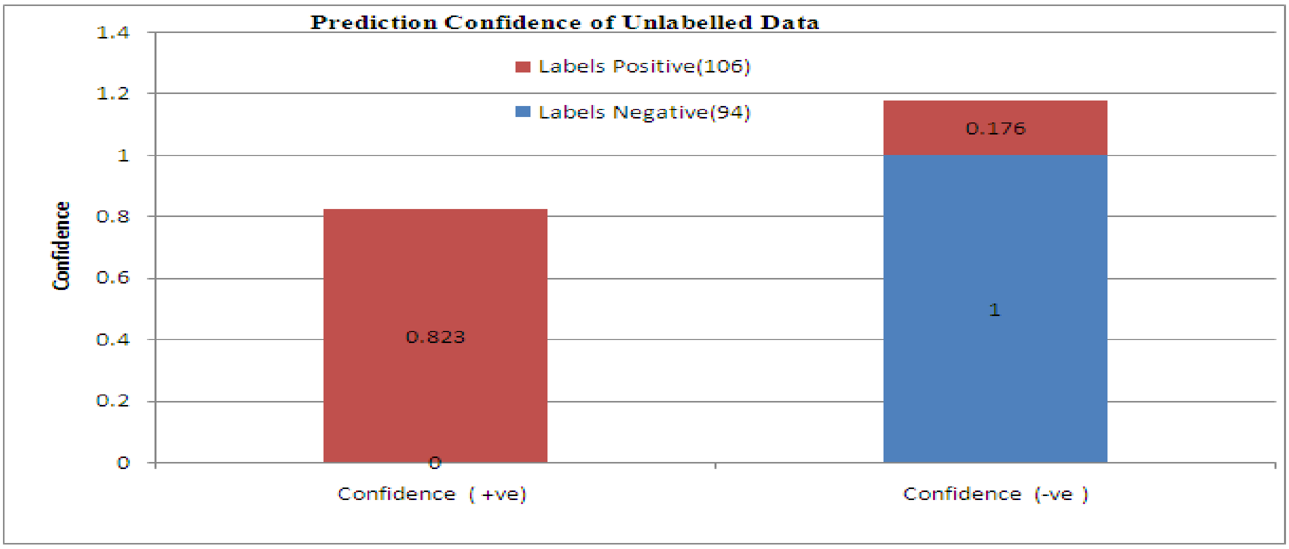


Figure 4: Prediction results of Unlabelled Data

From the above Fig. 4, it is clear that out of the total 200 unlabelled movie datasets fed to the developed Decision Tree Model, 106 labels have been classified as Positive sentiments and 94 as Negative Sentiments. The +ve confidence and –ve confidence value of '0' and '1' is for Negative Labels, whereas on the other hand it is 0.823 and 0.176 for Positive Labels. This can be clearly seen in Fig. 3 above.

6. Conclusion

Here machine learning classification approaches with different feature selection schemes to obtain a sentiment analysis model for the movie review dataset has been explored. Numerous analyses were done utilizing distinctive capabilities and parameters to acquire greatest accuracy. In the present work machine learning techniques are used to detect the sentiments of movie reviews. Polarity of the sentiment as to whether a review is positive or negative, as well as the intensity of sentiment are considered. Traditionally movie reviews only focus on polarity of a review and thus cannot distinguish between a good movie and a masterpiece. Decision tree proves to be good method to learn sentiments from a review, achieving considerable accuracy with very basic feature set. Overall, gaining further improvements using the supervised model of machine learning for the purpose of sentiment analysis is studied in Natural Language Processing and requires sufficient knowledge of Linguistics. This requires more sentence structure analysis and understanding how words used in different parts of speech can change the sentiment of the text. Extracting features based on semantic structure of the text will improve the accuracy of these classifiers. From the above section "Results and Discussions", it is seen that the performance of the decision tree has a better accuracy (Fig. 8 & 9), and especially the recall of the POSITIVE labelled comments (which are much less compared to the NEGATIVE labelled comments) is 87.5%. Also the model performance has RMSE value of 0.47 and Absolute error of 0.313, which again clearly demonstrates the good predictive capability of the model.

References

[1]. B. Pang et al, 2002, Thumbs up ?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 79-86.

[2]. P.D. Turney, 2002, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL), 417–424.

[3]. Riloff, E &Wiebe, J., 2003, Learning extraction patterns for subjective expressions, EMNLP‟ 03.

[4]. Loren Terveen et al, 1997, PHOAKS: A system for sharing recommendations, Communications of the Association for Computing Machinery (CACM), 40(3):59–62.

[5]. Minqing Hu and Bing Liu, 2004, Mining and summarizing customer reviews, Proceedings of the 10th ACM SIGKDD International conference on knowledge discovery and data mining.

[6]. Beaver, D., Kumar, S., Li, H. C., Sobel, J., & Vajgel, P. (2010). Finding a needle in hay stack: Facebook's photo storage. In Proceedings of the ninth USENIX conference on operating systems design and implementation (pp. 1–8). Berkeley, CA,USA: USENIX Association.

[7]. Cukier K., The Economist, Data, data everywhere: A special report on managing information, 2010, February 25, Retrieved from http://www.economist.com/node/15557443

[8]. Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032–2033.

[9]. Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual.

[10].Jiang, J. (2012). Information extraction from text. In C. C. Aggarwal, & C. Zhai (Eds.), Mining text data (pp. 11–41). United States: Springer