Sentiment Analysis of Movie Reviews using Machine Learning Techniques.

Md. Bapiur Rahman Biswas.

Department of Information Technology,

University of Information Technology and

Sciences.

Baridhara, Dhaka, Bangladesh.

rahmanbapiur@gmail.com

Salma Akter popy.

Department of Information Technology,

University of Information Technology and

Sciences.

Baridhara, Dhaka, Bangladesh.

salmaakteruits@gmail.com

ABSTRACT :

- ABSTRACT Sentiment analysis is the analysis of emotions and opinions from any form of text. Sentiment analysis is also termed as opinion mining. Sentiment analysis of the data is very useful to express the opinion of the mass or group or any individual. This technique is used to find the sentiment of the person with respect to a given source of content. Social media and other online platforms contain a huge amount of the data in the form of tweets, blogs, and updates on the status, posts, etc. In this paper, we have analyzed the Movie reviews using various techniques like Naïve Bayes, K-Nearest Neighbour and Logistic Regression etc.

General Terms:

Reviews, Classification, Weka, Artificial Intelligence, Machine Learning, Sentiments.

Keywords:

- Sentiment Analysis, Logistic Regression , Movies Reviews, Naive Bayes, K-Nearest Neighbour, svm.

1. INTRODUCTION:

Movie reviews are an important way to gauge the performance of a movie. While providing a numerical/stars rating to a movie tells us about the success or failure of a movie quantitatively, a collection of movie reviews is what gives us a deeper qualitative insight on different aspects of the movie. A textual movie review tells us about the strong and weak points of the movie and deeper analysis of a movie review can tell us if the movie in general meets the expectations of the reviewer. Sentiment Analysis is a major subject in machine learning which aims to extract subjective information from the textual reviews. The field of sentiment of analysis is closely tied to natural language processing and text mining. It can be used to determine the attitude of the reviewer with respect to various topics or the overall polarity of review. Using sentiment analysis, we can find the state of mind of the reviewer while providing the review and understand if the person was "happy", "sad", "angry" and so on. In this project we aim to use Sentiment Analysis on a set of movie reviews given by reviewers and try to understand what their overall reaction to the movie was, i.e. if they liked the movie or they hated it. We aim to utilize the relationships of the words in the review to predict the overall polarity of the review.

2. LITERATURE REVIEW:

The original work on this dataset was done by researchers at Stanford University wherein they used unsupervised learning to cluster the words with close semantics and created word vectors. They ran various classification models on these word vectors to understand the polarity of the reviews. This approach is particularly useful in cases when the data has rich sentiment content and is prone to subjectivity in the semantic affinity of the words and their intended meanings. Apart from the above, a lot of work has been done by Bo Pang and Peter Turnkey towards polarity detection of movie reviews and product reviews. They have also worked on creating a multi-class classification of the review and predicting the reviewer rating of the movie/product. These works discussed the use of Random Forest classifier and SVMs for the classification of reviews and also on the use of various feature extraction techniques. One major point to be noted in these papers was exclusion of a neutral category in classification under the assumption that neutral texts lie close to the boundary of the binary classifiers and are disproportionately hard to classify. There are many sentiment analysis tools and software existing today that are available for free or under commercial license. With the advent of microblogging, sentiment analysis is being widely used to analyze the general public sentiments and draw inferences out of these. One famous applications was use of Twitter to understand the political sentiment of the people in context of German Federal elections.

3. MACHINE LEARNING METHODS:

3.1 Naïve Bayes:

It is a technique based on Bayes' Theorem. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This model is easy to build and particularly useful for very large datasets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

3.2 K- Nearest Neighbour:

K-NN is the simplest of all machine learning algorithms. The principle behind this method is to find a predefined number of training samples closest in distance to the new point and predict the label from these. The number of samples can be a user-defined constant or vary based on the local density of points. The distance can be any metric measure. Standard Euclidean distance is the most common choice for calculating the distance between two points. The Nearest Neighbours have been successful in a large number of classification and regression problems, including handwritten digits or satellite image processing and so on.

3.3 SVM:

Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

3.4 logistic regression:

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

4. EXPERIMENTAL SETUP:

The research started with analyzing different research and review papers on sentiment analysis and the summary for each paper was created by reading and understanding the paper. Study of frequently

used classification algorithms such as Naïve Bayes, logistic regression, k-nearest neighbour, Support Vector Machine was done.

i)   First collect dataset on movie review statement and sentiment analysis.
ii)   Then make code using algorithms(svm,knn,naïve bayes,logistic regression).
iii) Upload that dataset in code.
iv)   Then run the programme.
v)   If programme/code has any problem then solve it.
vi)   After running the code successfully it give us results.
vii) The results contain answer of all algorithm accureccy.
viii) It give answer all movies depend on sentimental analysis.

5. RESULTS:

```
Naive Bayes
Accuracy Score: 98.9161849711%
Confusion Matrix:
[[586  12]
 [  3 783]]

Logistic Regression
Accuracy Score: 99.3497109827%
Confusion Matrix:
[[593   5]
 [  4 782]]

Support Vector Machine
Accuracy Score: 99.0606936416%
Confusion Matrix:
[[592   6]
 [  7 779]]

K Nearest Neighbors (NN = 3)
Accuracy Score: 98.6994219653%
Confusion Matrix:
[[589   9]
 [  9 777]]

Analysis
No. of tokens:  294
No. of positive tokens:  143
No. of negative tokens:  151

Search Results for token/s: ['awesome']
     Token  Positive  Negative
11  awesome     896.0      1.0

Test a custom review message
Enter review to be analysed:  I really appreciate the details of the movie. It was an awesome
The review is predicted Positive
"""
```

6. CONCLUSIONS:

In this research, various techniques were used to identify the polarity of the tweets. The algorithms performed were Naïve Bayes, K-Nearest Neighbour, logistic regression,svm. The best results were given by Naïve Bayes classifier. The Naïve Bayes classifier achieved 98.92% accuracy, Logistic regression classifier we achieved 99.35% accuracy, K-Nearest Neighbour classifier achieved 98.70% accuracy, SVM classifier achieved 99.06% accuracy. As only few algorithms were tested , it is required to test other algorithms or create hybrid methods so that accuracy of the results can be increased. Finding the polarity of the reviews can help in various domain. Intelligent systems can be developed which can provide the users with comprehensive reviews of movies, products, services etc. without requiring the user to go through individual reviews, he can directly take decisions based on the results provided by the intelligent systems.

7. Future Plan:

Making a Sensor based movie review and recomandation system. To make this system we need a sensor in the glass which we can called smart glass, that smart glass contain a sensor which is check your mood and automaticcaly send a movie recomandation to your smart phone and smart tv as a notification and show its rating by sentimental analysis and that smart glass also contain a microphone which can receive voice from you, besides sensor you can tell what movie you want to see then it also give you that movie rating/review using sentimental analysis .

8. REFERENCES:

1] Joscha Markle-Hub, Stefan Feuerriegel, Helmut Prendinger. 2017 Improving Sentiment Analysis with Document-Level Semantic Relationships from Rhetoric Discourse Structures, Proceedings of the 50th Hawaii International Conference on System Sciences.

2] A. Hogenboom, F. Frasincar, F. de Jong, and U. Kaymak,. 2015, Using Rhetorical Structure in Sentiment Analysis, Communications of the ACM, vol. 58, no. 7, pp. 69–77.

3] Amir Hossein Yazdavar, MonirehEbrahimi, Naomie Salim, 2016, Fuzzy Based Implicit Sentiment Analysis on Quantitative Sentences, Faculty of Computing, UniversitiTechnologi Malaysia, Johor, Malaysia, Journal of Soft Computing and Decision Support Systems vol 3:4, pp.7-18.

4] Bhatia, R. S., Graystone, A., Davies, R. A., McClinton, S., Morin, J., & Davies, R. F. 2010, Extracting information for generating a diabetes report card from free text in physicians notes. Paper presented at the Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents.

5] Dhiraj Murthy, Twitter and elections: are tweets, predictive, reactive, or a form of buzz?, Information, Communication & Society, 18:7, 816-831, DOI:10.1080/1369118X.2015.1006659.

6] Kamal A., 2015, Review Mining for Feature Based Opinion Summarization and Visualization.

7] Kamal, A. 2013 Subjectivity Classification using Machine Learning Techniques for Mining FeatureOpinion Pairs from Web Opinion Sources. International Journal of Computer Science Issues 10(5), 191- 200.

8] Humera Shaziya, G.Kavitha, Raniah Zaheer, 2015, Text Categorization of Movie Reviews for Sentiment Analysis , International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue11.