Sentiment Analysis of Movie Reviews

Sayali Pisal, Shruti Kothari

***Abstract*—Sentiment analysis is a sub-domain of opinion mining where the analysis is focused on the extraction of emotions and opinions of the people towards a particular topic from a structured, semi-structured or unstructured textual data. In this project, we try to focus our task of sentiment analysis on IMDB movie review database. We examine the sentiment expression to classify the polarity of the movie review as 0(disliked) or 1(liked) and perform feature extraction and use these features to train our bi-label classifier to classify the movie review into its correct label. Due to lack of strong grammatical structures in movie reviews which follow the informal jargon, an approach based on structured N-grams(in the range 1-2) has been followed.Word2Vec implementation has also been done for a comparative view of both the vectorization methods, Term Frequency-Inverse Document Frequency (TF-IDF), of unigram and bigrams and Word2Vec. In addition, a comparative study on different classification approaches such as Naive Bayes, Logistic Regression, Random Forest and Linear SVC has been performed to determine the most suitable classifier to suit our problem domain. We conclude that our proposed approach to sentiment classification supplements the existing movie rating systems used across the web and will serve as base to future researches in this domain. "Our approach using classification techniques has the best accuracy of 87.46 percent.[1]**

Keywords : Sentiment Analysis, IMDB Movie Reviews, Random Forest, Naive Bayes, Linear SVM, GridSearchCV, Logistic Regression, Word2Vec, Cross-Validation, SelectKBest, chi2

I. Introduction

THE advent of Internet has become a huge Cyber Database which hosts gigantic amount of data which is created and consumed by the users which has led to an increase in the amount of sentimental content available on the Web. The content has been growing at an exponential rate giving rise to a new industry filled with it, in which users express their opinions across channels such as Facebook, Twitter, Rotten Tomatoes and Foursquare. Such content is often found in social media web sites in the form of movie or product reviews. Opinions which are being expressed in the form of reviews provide an opportunity for new explorations to find collective likes and dislikes of cyber community. One such domain of reviews is the domain of movie reviews which affects everyone from audience, film critics to the production company. The movie reviews being posted on the websites are not formal reviews but are rather very informal and are unstructured form of grammar. Opinions expressed in movie reviews give a very true reflection of the emotion that is being conveyed. Understanding of the sentiments of human masses towards different entities and products enables better services for contextual advertisements, recommendation systems and analysis of market trends. The presence of such a great use of sentiment words to express the review inspired us to devise an approach to classify the polarity of the movie using these sentiment words. Thus, the focus of our project is sentiment focused web framework to facilitate the quick discovery of sentimental contents of movie reviews analysis of the same.[1]

Advisor: Gayathri Namasivayam , Professor for Department of Computer Science, San Jose State University, San Jose State University, San Jose, California 95192.

II. Literature Survey

Sentiment Analysis is a technology that will be very important in the next few years. With opinion mining, we can distinguish poor content from high quality content. With the technologies available we can know if a movie has more good opinions than bad opinions and find the reasons why those opinions are positive or negative. Much of the early research in this field was centered around product reviews, such as reviews on different products on Amazon.com, defining sentiments as positive, negative, or neutral. Most sentiment analysis studies are now focused on social media sources such as IMDB, Twitter and Facebook, requiring the approaches be tailored to serve the rising demand of opinions in the form of text. Furthermore, performing the phrase-level analysis of movie reviews proves to be a challenging task.

Sentiment analysis has been a good research area from a long time. Many have presented their proposed work in a large volume, they have been solved many problems like anaphora resolution, thwarting, sarcasm, word sense disambiguation, intensifiers, negation etc. In our study, we have focused on word and feature level sentiment analysis, problems occur in extraction of features and how to deal with complex sentences. Many researchers have used various types of tools like WordNet, Opinion Lexicon, MPQA, SentiWordNet etc for calculating the scores and used machine learning and rule based techniques for resolving their problems. Some have been proposed the systems for multilingual text and cross domain analysis also. We have summarized their work below.

[Peifeng Li et-al, 2011] proposed a Dependency Tree They used SVM Classifier to identify the polarity. The proposed sentiment word-based pruning strategy for dependency tree reduced the noise and showed good performance. [2]

[Zhongchao Fei, et-al, 2004] introduced the basic task of sentiment classification that how to get sentiment information in text using phrase patterns algorithms. They decided to use machine learning technology to evaluate the polarity and strength of phrase patterns. Their experiment achieved an 86 precision rate. [3]

[Farah Benamara, et-al, 2007] proposed an AAC-based (Adverb-Adjective Combinations) sentiment analysis technique that used a linguistic analysis of adverbs of degree. Their results showed that using adverbs and AACs produced significantly higher precision and recall. [4]

[Ann Devitt, et-al, 2007] presented a lexical cohesion based metric method of sentiment intensity and polarity in text. The results on polarity tag assignment, the best performers were the relation type and node specificity metrics using only modifiers, significant to the 0.05 level. [5]

[Bruno Ohana, et-al 2009] proposed a method for applying SentiWordNet to derive a data set of document metrics and other relevant features, and performed an experiment on sentiment classification of film reviews using the polarity data set. The proposed method yielded an overall accuracy of 65.85 percent in SentiWordNet features. [2]

Si Li Hao Zhang et-al, 2011 presented a novel combined model based on phrase and sentence level's analysis. The improvements over sentence-level from phrase-level are found to be statistically significant for precision, recall and F-measure with large margin. [6]

V.K. Singh et-al, 2013 had implemented two Machine Learning based classifiers (Nave Bayes and SVM), the Unsupervised Semantic Orientation approach (SO-PMT-IR algorithm) and the SentiWordNet approaches for sentiment classification of movie reviews. The accuracy of classification by NB was marginally better than the SVM and was close to the SO-PMI-IR algorithm. [7].

III. Design and Methodology

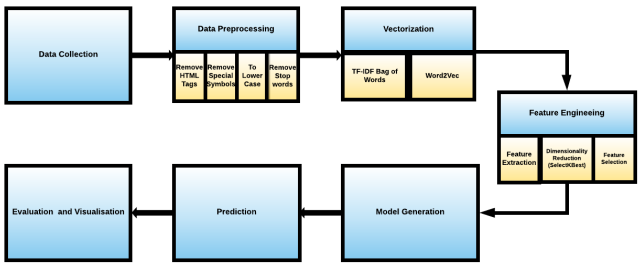

Fig. 1. **Architectural Design Work Flow**

A. Data Collection

A very large volume of user-generated text or reviews are available on internet in the form of various websites like imdb.com, twitter.com, amazon.com, flipcart.com, ndtv.com etc., We have used the data from imdb.com to collect the movie reviews. To determine the polarity of the sentences based on features, large numbers of reviews are collected from the www.imdb.com site for movies.

B. Data Pre-processing

After collecting the reviews, we did pre-processing before sentiment analysis. In this step, unnecessary noise was removed from the reviews which is not the part of analysis. We did following pre-processing steps for preparing our data:

1] Removal of HTML Tags
2] Removal of Special Symbols
3] Conversion to lower case
4] Removal of Stop Words

C. Vectorization

Vectorization is done using two techniques of TF-IDF and Word2Vec, which are used as follows:

1] TF-IDF
After collecting the data, all sentences or clauses are tokenized in the reviews by a blank space delimiter. Thus the tokens here are individual words. For each review, TF-IDF creates a vector where each element in the vector is the importance weight of that word calculated by combining frequency count and the document count for that particular word or token.

2] Word2Vec
A neural network for processing the reviews and grouping the vectors of similar words together in a vector space, creating a word embedding. Implemented Word2Vec using Googles pre-trained Word2Vec model.

D. Feature Engineering

Feature engineering is done by Dimensionality Reduction which is used for reducing the number of random variables under consideration by obtaining the set of principle variables. Dimensionality reduction is implented by the SelectKBest Algorithm combining it with chi2 methodology.

E. Model Generation

Four machine learning algorithms viz. Multinomial Naive Bayes, Random Forest, Logistic Regression and Linear SVC are implemented and tested for prediction of unseen dataset. The models are created with vectorization variation of TF-IDF and Word2Vec and are thus compared for better accuracy results for prediction.

F. Prediction

Prediction of movie review abstract as positive(1) and negative(0) is made by the algorithms implemented.

G. Evaluation and Visualization

The four machine learning classification models are evaluated based on the accuracy measure using TF-IDF and word2vec, precision-recall curve to summarize the trade-off between the true positive rate and the positive predictive value, confusion matrix to compare the FP(False Positive)-Type I error and FN(False Negative)-Type II error rates and execution time to show the time taken by each algorithm for execution.

IV. Implementation

The basic implementation methodology for classification is applying different classification algorithms on vectorized data to obtain the prediction results. The implementation details ae as follows:

A. Dataset Collection

We use the imdb movie reviews dataset which contains 50k reviews in total. These 50k reviews are split into two sets, training dataset and test dataset with 25k reviews in each set. Further, the 25k reviews in the training dataset are split into positive and negative review sets each with equal number of records with the count of 12.5k.
The dataset used is a complete balanced dataset with equal number of positive and negative review count. Each review is a collection of words which is combined to form a text abstract.

Training dataset:
The training dataset which consists of 25k reviews has three dimensions which are:
id: unique ID/number/count assigned to each review in the dataset in a ordered format)
review: the review text abstracts which are a collection of words, punctuations, symbols, numerals.
sentiment: sentiment assigned for each review based on the polarity of positive(1)/negative(0) sentiment contained by it.

Test dataset:
The test dataset which consists of 25k reviews has two dimensions which are:
id: unique ID/number/count assigned to each review in the dataset in a ordered format)
review: the review text abstracts which are a collection of words, punctuations, symbols, numerals.
The test dataset does not contain the attribute values for sentiment, as this is the predicted value assigned by the classification algorihm.

B. Data Pre-processing

Data Pre-processing is the preparation of the dataset before applying any algorithm on it. This is done to speed up the process of labeling the polarity of the reviews. Our approach for pre-processing of the dataset is as follows:

1] Removal of stop-words
Stop words were eliminated using nltk stopwords dictionary. Stop words are propositions, irrelevant words like ”is”, ”were”, ”the” etc which are not really important in classification.

2] Removal of HTML tags
Removed HTML tags so as to get all the review abstracts in a single common format containing only English letters. The HTML tags can generate sentence break error which is avoided by their removal.

3] Removal of Special characters and punctuation
Special symbols which occur in the form of punctuations and special ASCII characters were removed during the data cleaning process to obtain the plain text for algorithm processing in the later section.

4] Conversion of text to lowercase
All the cleaned text was converted to lower case to get it in a unified format for processing.

C. Vectorization

Two vectorization methods are used which are as follows:

1] TF-IDF Bag of Words model
Converting collection of raw review documents to a matrix of TF-IDF features.
Goal - To scale down the impact of tokens with frequent occurrence in the reviews that are less informative than the features that occur in a small fraction holding more importance.

2] Word2Vec model
A neural network for processing the reviews and grouping the vectors of similar words together in a vector space.
Implemented Word2Vec using Googles pre-trained Word2Vec model.

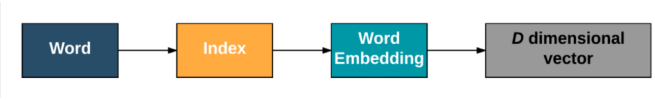

Fig. 2. **Word2Vec model flow**

D. Feature Engineering

Feature engineering mainly aims at taking whatever information you have about your problem and turning it into numbers that you can use to build your feature matrix.
Feature engineering is applied by Dimensionality Reduction which reduces the number of random variables under consideration by obtaining the set of principle variables. The SelectKBest algorithm is used which removes all but the k highest scoring features. Clubbing SelectKBest with chi2 for independence testing to determine the dependency of two features outperforms all the results giving higher scores.

E. Classification Models

1] Multinomial Naive Bayes
It is a conditional probabilistic distribution model which is suitable for classification with discrete features especially word counts for text classification.

2] Random Forest
An ensemble learning classifier which builds a number of decision trees on the training data and combines all their outputs to make the best predictions on the test data. It also

contains meta estimator that fits a number of classifying decision tress on various data sub samples.

3] Linear SVM using GridSearchCV
SVM is a discriminative classifier formally defined by a separating hyperplane which means given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.
Using GridSearchCV the optimised parameters and best cross validation score for linear SVM is obtained.

4] Logistic Regression
A predictive analysis model used for data description and to derive the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
The logistic regression not only gives a measure of how relevant a predictor is (coefficient size) but also its direction of association (positive or negative).

V. Results and Evaluation

The following results were obtained using the classification models with TF-IDF and Word2Vec :

1] *Accuracy scores tabulated for the four used algorithms for prediction, viz. Logistic Regression, Multinomial Naive Bayes, Linear SVC and Random Forest is as below:*

ALGORITHM	ACCURACY USING TF-IDF	ACCURACY USING WORD2VEC
LOGISTIC REGRESSION	0.86884	0.85223
MULTINOMIAL NAÏVE BAYES	0.86651	-
LINEAR SVC	0.87468	0.85988
RANDOM FOREST	0.83148	0.79380

Fig. 3. **Accuracy comparison using TF-IDF and Word2Vec**

We achieved the highest accuracy using TF-IDF with Linear SVC, and the lowest using Word2Vec with Random Forest. Random Forest did not yield good results for either of the methods as Random Forest makes the split based on randomly selected feature.Since, text data is sparse there are chances that Random Forest splits on irrelevant features. As Naive Bayes does not work with negative values, the above field for Naive Bayes using Word2Vec is blank.

2] *Confusion matrices for resulting Linear SVC using TF-IDF and Word2Vec.*
According to the results obtained from the confusion matrices, the FP (False Positive)-Type I error and FN (False Negative)-Type II error rates are slightly higher when using Word2Vec. While, better results are obtained using TF-IDF.

3] *Precision-Recall Curve for Logistic Regression using TF-IDF and Word2Vec.*
The Precision-Recall curve summarize the trade-off between the true positive rate and the positive predictive value for the Logistic Regression predictive model using different probability thresholds. As seen in the above graphs, the area under the PR curve which is the positive predicted values region, is almost the same using TF-IDF and Word2Vec, but slightly higher for TF-IDF.

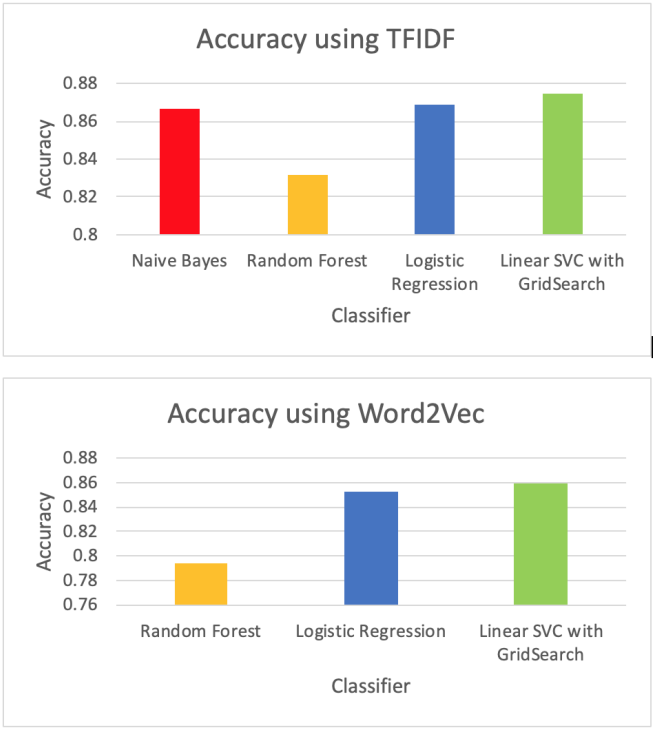

Fig. 4. **Accuracy Plot using TF-IDF and Word2Vec**

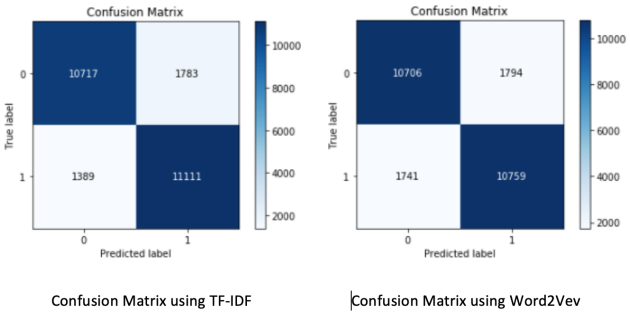

Fig. 5. **Confusion matrices for Linear SVC using TF-IDF and Word2Vec**

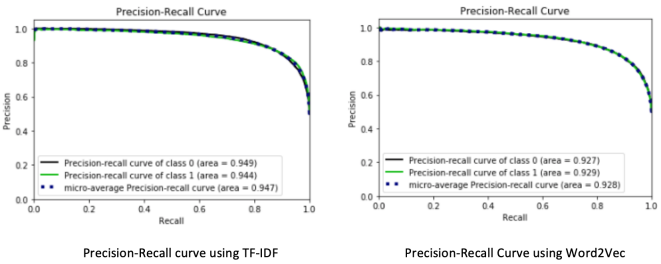

Fig. 6. **Precision-Recall Curve for Logistic Regression using TF-IDF and Word2Vec**

4] *Execution time for different algorithms using TF-IDF and Word2Vec.*

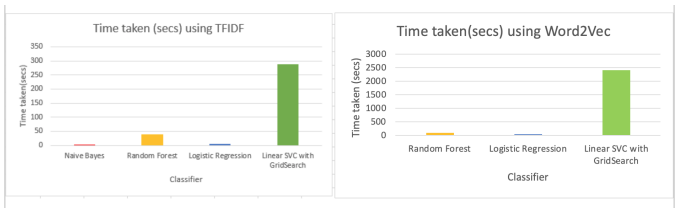

Fig. 7. **Execution time for different algorithms using TF-IDF and Word2Vec**

The execution time for implemented algorithms viz. Multinomial Naive Bayes, Random Forest, Logistic Regression and Linear SVC is seen in the graphs plotted which clear show that the time taken by TF-IDF is less as compared to Word2Vec.

VI. Conclusion and Future Work

The project presents an extensive study on the prediction of sentiments from movie review abstracts by four Machine Learning Classification Algorithms viz. Logistic Regression, Multinomial naive bayes, Random Forest and Linear SVC. The main contribution of this work is to effectively select the appropriate methods for the sentiment classification in the online movie reviews taken from IMDB website. The project attempts to determine which vectorization method out of TF-IDF and Word2Vec performs best for sentiment classification on movie reviews in combination with the four algorithms. We also investigate how feature selection methods contribute to improving the classification performance of the four machine learning classifiers. The result indicates that TF-IDF yields siginificantly improved results over the Word2Vec vectorizer, particularly with Linear SVC classifier, which when used with GridSearchCV achieves the best performance for the review sentiment classification, with an accuracy of 87.46 percent.
Our future work will focus on developing a Doc2Vec Movie Review sentiment analysis model investigating the implementation of Neural Network approaches and the implementation of some deep learning models, to address the objective of review sentiment classification.

VII. Acknowledgement

We would like to extend acknowledgment and express our gratitude to prof. Gayathri Namasivayam, instructor, CS-256 course for encouraging us to work on this application as a course project and making the guidance, outlook and help required available on very first request. We would also like to thank her for her timely counselling and boost to implement various methodologies for enhancement of the project.

References

[1] Tirath Prasad Sahu, Sanjeev Ahuja, "Sentiment analysis of movie reviews: A study on feature selection classification algorithms", International Conference on Microelectronics, Computing and Communications (MicroCom), 2016.

2] Peifeng Li, Qiaoming Zhu, Wei Zhang, "A Dependency Tree based Approach for Sentence-level Sentiment Classification", 12th ACIS International Conference on Software Engineering Artificial Intelliaence Networking and Parallel/Distributed Computing, 2013.

3] Zhongchao Fei, Jian Liu, Gengfeng Wu, "Sentiment Classification Using Phrase Patterns", Proceedings of the Fourth International Conference on Computer and Information Technology, 2004.

4] Farah Benamara, Carmine Cesarano, Diego Reforgiato, "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives", ICWSM Boulder. CO USA, 2017.

5] Ann Devitt, Khurshid Ahmad, "Sentiment Polarity Identification in Financial News: A Cohesion-based Approach", Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 984-991, June 2007.

6] Si Li, Hao Zhang, Weiran Xu, Guang Chen, Jun Guo, "Exploiting Combined Multi-level Model for Document Sentiment Analysis", 2010 International Conference on Pattern Recognition, 2010.

7] V. K. Singh, R. Piryani, A. Uddin, "Sentiment Analysis of Textual Reviews", 5th International Conference on Knowledge and Smart Technology (KST), 2013.

8] Pallavi Sharma, Nidhi Mishra, "Feature level sentiment analysis on movie reviews", 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016.