Sentiment Analysis Of Movie Reviews using SentiWordNet

Naga Sindhu Korlapati

Data 294 – Research Paper

San Jose State University

08

**Fall**

Table of Contents

Abstract 2

Introduction 3

Purpose 3

Literature Review: 7

Research design and Methods 8

Classification 12

Results 14

Discussions 15

Recommendations 15

Future Work 16

References 16

## Abstract

Movie reviews are used as an important measure to determine the performance of a film. However, providing the star ratings of the film can help in knowing the success or failure of a movie. A collection of movie reviews from distinct users provides us deep insights on different elements of a movie. The primary objective of this study is to use sentiment analysis on a collection of movie reviews provided by viewers to predict overall movie rating. This paper used the SentiWordNet strategy to know the suitable polarity of textual reviews. This research work may be of interest to the movie industry as well as online platforms like Netflix and Amazon to facilitate better user experience to the customers.

**Keywords**: Classification, NLP, Feature selection, Data mining, accuracy.

## Introduction

The Internet movie database (IMDb) is the most popular comprehensive database having information about movies, actors and production. Furthermore it allows users to post comments and provides the aggregated ratings about the movie. Given that massive amount of data and degree of interactions between the users, IMDB can be used as good source for Data mining use cases. Sentiment analysis is major topic in Natural Language processing (NLP) which aims at extracting insights from the textual reviews. It also helps in identifying the polarity of the textual content or mind-set of reviewer with respect to multiple topics. Using sentiment analysis, we can find reviewer’s state of mind, while reviewing and understanding whether the person was "happy," "sad," "angry," etc.

## Purpose

The main goal of this research paper is to predict the overall rating of a viewer’s comment about a movie using Sentiment analysis. Based on these results, movie viewers can decide whether to watch newly released movie or not, and also it is useful for the movie industry, what kind of movie the average viewer will usually like. This paper can also be used to develop a recommendation system by recommending movies to the viewers based on previous reviews.

**Problems on Sentiment analysis**

Sentiment analysis is considered as a more challenging task to classify a review as positive or negative according to the overall sentiment expressed by the writer. By simply looking at the words in a review, the feelings that have been expressed are often very difficult to identify. In below review example:

“*This movie is supposed to be excellent. It sounds like a brilliant plot, the actors are first grade, and the supporting cast is also good, and Tony is trying to deliver a good performance, but it can't keep it up.*”

Words such as excellent, brilliant and good are viewed as positive sentiment, but in the above review there is an anaphor (“it can’t keep up”). What does “it “ refers in the sentence, whether the movie or Tony’s performance is not clear. This makes it hard to determine the overall polarity of the review. And most reviews does contain anaphora, abbreviations, emoticons, punctuation marks and poor spelling. This paper uses the SentiWordNet algorithm, to determine the sense of opinion and exact contextual polarity of the text.

**Definitions & Terminologies**

**Natural language processing:** Natural language processing is also known as NLP technique, is a branch of Artificial Intelligence that deals with the interaction between computers and humans using natural language [Garbade, G. &. (n.d.)]. NLP's main goal is to read, decipher and understand human languages in a valuable manner. Most NLP techniques rely on Machine learning algorithms to understand meaning of human languages.

**Python:** Python is widely considered as a most popular language for building Machine-learning models. It is an open source, user-friendly language with huge amount of inbuilt libraries. Practically it’s most popular in building Deep Learning Frameworks like Tensor flow and keras.

**Web Scrapper:** A web scrapper is software that extracts data from websites. For this research work, data is parsed using python inbuilt library popularly known as beautiful soup**.**

**Confusion Matrix:** Confusion Matrix is used to describe the performance of a classification model. Each column in the matrix table represents the predicted class instances, whereas each row represents the actual class instances [*Python Machine Learning Tutorial*. (2019, 05 13]. Using this confusion matrix, it is useful to measure accuracy and misclassification rate

**Background**

**SentiWordNet:** WordNet is a publicly available database of words containing a semantic lexicon for the English language that organizes words into groups called synsets (i.e., synonym sets). A synset is a collection of synonym words linked to other synsets according to a number of different possible relationships between the synsets (e.g., is-a, has-a, is-part-of, and others) [Allotey, 2011]. It is an open lexical resource used mainly for sentiment analysis.

Each sysnet s contains three numerical scores, which represents the terms present in the sysnet are positive, negative, neutral. Consequently, distinct meanings of the same word may have distinct opinion related characteristics. For example, a word "good" has 4 meanings of the noun "good" has 21 meanings of the adjective and "good", has 2 meanings of the adverb. So overall, it will have three positive and negative scores.

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Average positive score | Average Negative score | Word |
| Noun | 0.513 | 0 | good |
| Adjective | 0.5 | 0 | good |
| Adverb | 0.188 | 0 | good |

Sentiwordnet make use of POS (parts of speech) tagger to tokenize each word in a sentence. POS tagger is software that reads text as input and assigns parts of speech to each word present in the text as noun, verb and adjective etc. Detailed overview of POS tagger is explained in the methodology section.

**Machine learning:** Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [Expert system.com]. The learning process begins with data, in order to find insights or patterns present, and make better decisions in future based on events. Machine learning algorithms are often sub divided as supervised or unsupervised techniques. Supervised algorithm does contain labeled data and based on past experience it predicts the future events. Whereas, unsupervised algorithm mostly used to find hidden patterns present in the unlabeled data. Here in our problem, we are mostly working on supervised techniques.

**Support vector machines:** It is a supervised Machine-learning framework used for both classification and regression problems. The main objective of the SVM technique is to find a hyper plane in an N-dimensional space that distinctly classifies the data points. For example in the figure below red circles belongs to one class and a light grey circle belongs to other class, a hyper plane is used to clearly distinguish both the classes. The major advantage of support vector machines is, it works well when dealing with high dimensional data. But major drawback with SVM model is, it won’t give accurate results when working with imbalanced data.

|  |
| --- |
|  |

Fig 1: Hyper plane classifying the classes

**Decision Trees:** Decision tree is the most popular Machine-learning algorithm used for both classification and regression problems. A decision tree is a tree where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value) (Sanjeevi, M. (2017, 09 06)). This algorithm makes use of Gini index and information gain methods to determine which attribute should be present in the root node.

**Random Forest:** Random forest is one of the most popular algorithm applied for both classification and regression tasks. It is simple to use and doesn’t require any hyper parameters to tune. Random Forest is an ensemble classifier, which builds multiple decision trees and combines them together to obtain accurate results. The main difference between decision trees and Random forest is Decision Trees sometimes prone to over fitting while, Random forest prevents over fitting, by selecting random subset of features.

**K Nearest Neighbors**: KNN is a Non-parametric, lazy algorithm used for both classification and Regression problems. The definition of Non-parametric means, the algorithm doesn’t make any prior assumptions based on underlying data distribution. KNN work by calculating the number of nearest neighbors of a given type, the type with the most neighbors is the prediction. Prior to classification, we need to take two important decisions, one is value of K i.e., number of neighbors and the other one is distance metric [*A Quick Introduction to K-Nearest Neighbors Algorithm*. (n.d.)]. For example in the figure below the diamond is the new value to be predicted, then the algorithm calculates it’s nearest neighbors using distant metric and assigns red dots as it’s predicted neighbors. This algorithm is used in wide applications like anomaly detection, semantic searching etc.

|  |
| --- |
|  |

## Fig 2: Different classes for KNN

## Literature Review:

It has been over 20 years, that researches from various backgrounds used algorithms to predict movie ratings. There are techniques like kernel methods and model trees that had been implemented to solve the problem of movie score prediction using information related to movie’s production and content. Based on findings the accuracy of kernel method is slightly better than model trees. And the accuracy achieved through kernel method is 84%. Zhu, et all built a regression model by considering reviewer and product information to predict ratings.

Beset, et all developed a model where higher order sentences were converted to low dimensional embedding’s, in which a classifier was used to predict the sentiment of a sentences.

Also, there is another problem called the Netflix challenge, which is similar to the one we described in this research paper. Netflix released the dataset which contains 100 million entities as an open challenge to develop a framework for movie rating prediction for individual users. Adam Sadovsky, et all from Stanford University built a classification framework for this challenge using logistic regression. In the year 2012 oghina, et all implemented a prediction algorithm from social media. However, this research differs from previous works as it is only used for textual data coming from Twitter and cannot be implemented for non-textual data from other social media platforms.

Pang and Lee applied machine learning algorithms to classify reviews based on sentiment [Bo Pang and Lillian Lee. (2002)]. They used various algorithms such as Naïve Bayes, Maximum entropy and SVM classifiers and concluded that these algorithms didn’t perform well in doing classification tasks. However, SVM gave higher accuracy compared to Naïve Bayes and Maximum entropy. They observed that implementing unigrams as feature representations helped in achieving high accuracy, but when they try to use the bigrams as features, the accuracy was lower as compared to unigrams.

Gamon outlined that it is feasible to perform automatic sentiment classification on reviews. His work showed that implementing large feature vectors in combination with feature reduction helps in achieving high accuracy of Machine learning frameworks. He also stated that addition of linguistic analysis features such as stop words removal can contribute to improve classification accuracy.

## Research design and Methods

**Dataset Description**

The dataset having movie reviews are from IMDB and the relative star ratings are obtained from <http://ai.stanford.edu/~amaas/data/sentiment/>. Since IMDb doesn’t have any open API, the data needs to be extracted using web scrapping tool. The whole dataset has 50,000 IMDB movie reviews with equal number of 25000 train and 25000 test sets. Each review was labeled with a rating on a scale of 1-10. Therefore for binary classification, ratings are categorized into two different classes either 1 or 0. If the ratings are higher than 5, class 1 label has been assigned, meaning that particular user liked the film, otherwise he did not like it.

|  |  |  |
| --- | --- | --- |
| **Class** | **Meaning** | **Rating** |
| 1 | Positive | >5 |
| 0 | Negative | <5 |

The histogram below shows that the data set contains the same number of positive and negative classes. Hence the data has equal distribution of reviews.

|  |
| --- |
|  |

Fig 3: Positive and negative review counts

Apart from review data, the dataset also contains movie attributes like movie ID, movie name, genre, cast details, budget, length, rating etc. Overall, It consists of categorical, numerical, time data types. The entire data set contains 2500 movies with 15 features. No movie has more than 30 reviews.

**Proposed Methodology**

At first this paper focuses on identifying the sentiment of the individual feature contained in the review using Natural language processing and Machine learning techniques. And second based on polarity present in the review, each feature score is calculated; the overall rating of a movie is calculated by combining individual feature scores.

|  |
| --- |
|  |

Fig 4: Proposed methodology flow diagram

Above figure describes the brief outline of the process of sentiment analysis and classification techniques. The process consists of three primary steps: data preprocessing, feature extraction and classification. In preprocessing phase, after the data has been extracted, reviews are filtered to remove html tags, stop words and punctuation marks. And part-of-speech (POS) tagger tags each term in the review by it’s parts-of-speech. In feature extraction step, the tags are then used by a lexical resource SentiWordNet technique, to quantify the sentiment score for each term. In classification step, the terms in each review, together with their sentiment scores are stored as feature vector and sent to the classifier for finding the polarity of the phrases. And finally overall movie rating is obtained by combining each individual term scores.

Data Preprocessing

First step in the process is data preprocessing. IMDb data set contains html tags, which doesn’t serve any purpose in analyzing the data and regular expressions were used to abstract such kind of tags present in the text. In addition to that it requires removal of punctuation marks, like ‘\’ , and emoticons were abstracted using python inbuilt string manipulation libraries.

1. **Porter Stemming:** Porter stemming algorithm is applied which helps in replacing a particular word with its root and so words like eating and eat, or birds and bird becomes the same. Stemming helps in improving the classification accuracy while performing sentiment analysis.

**2. Stopping:** It is a technique used to remove most frequent words in a sentence. Natural language processing library has list of stop words containing prepositions like above, after, then and determiners like a, an, the. For example the text “ *Tony was so excited about the concert*”. Will be processed has “ Tony excited concert”.

3. **Parts of speech tagging:** POS tagger parses string of words in a sentence and labels each word with its parts of speech such as nouns, verbs and adjectives etc. For example, the text: “ *And now for something completely different*”. Generates the output as:

[(‘And’, ‘CC’), (‘now’,’ RB’), (‘for’, ‘IN’), (‘something’, ‘NN’), (‘completely’, ‘RB’), (‘different’, ’JJ’)]

Here *and* is tagged as coordinating conjunction (CC); *now* and *completely* are tagged as adverbs (RB); *for* as preposition (IN); something as noun (NN) and different as adjective (JJ).

**Feature Extraction**

The second step in text analytics pipeline is to calculate the individual term scores using SentiWordNet technique. This method requires term and its parts-of –speech to produce a score. For a negative term score, the value is preceded by a negation operator, whereas positive term score does not contain any preceding operator. Terms that are not present in the SentiWordNet dictionary are scored as zero. In few cases when there is a term having more than one score, the average score is generated. Below is the sample example that is generated after parsing with SentiWordNet technique.

|  |
| --- |
| null happy : JJ: 0.2881  head: NN: 0.0035;  start: NN : 0.0067;  movie : NN : 0.0000;  kick : VV: -0.015  bastard: NN: - 2881; |

Each line does contain term, part-of-speech and term score. Where “JJ” is labeled as adjective, “NN” is labeled as noun; “VV” is labeled as verb.

After parsing each review in the dataset, the matrix (feature vector) of n by m is created, where m is the total number of reviews and m represents the number of features (terms) present in the dataset. Below example array shows the overview of sample review data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Term1 | Term 2 | | Term 3 | Term 4 | |
| Review 1 | 0.1201 | -0.2333 | | -0.256 | -0.956 | |
| Review 2 | 0.989 | 0.00 | | 0.2356 | -0.003 | |
| Review 3 | 0.000 | | -0.123 | 0.0235 | | 0.001 |

These array is being sent as the input to the Machine learning classification algorithms to classify sentiments as positive and negative reviews. For IMDB dataset the feature vector contains 50000 rows and 768945 columns.

## Classification

In this study, the main task is classifying reviews into two different categories. Therefore, for this classification, multiple algorithms are modeled on above feature representations. The algorithms such as K Nearest Neighbors (KNN), Decision trees, Random Forest and Support vector machine (SVM) are used. These algorithms used in this study are implemented using Python module Sci-Kit learn library. This library provides various options usable for both classification and regression tasks. The highest accuracy is achieved is using random forest.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithms | KNN | Decision Trees | Random Forest | SVM |
| Accuracy | 67.7% | 78.6% | **85%** | 75% |

From the above table it is clear that Random Forest has highest accuracy compared to all other algorithms, whereas KNN has least accuracy.

**Movie overall rating**

Here each movie contains list of positive and negative reviews. Below formula is used to calculate the each individual feature score based on the polarity of feature. The percentage is calculated by counting the number of positive, negative and classified reviews [User specific product recommendation and rating system by performing sentiment analysis on product reviews. (2017, 01)].

And the individual feature score is calculated using below formula

The overall movie rate is calculated by aggregating the each individual feature scores.

|  |  |
| --- | --- |
| Total Number of Movies tested | Correctly Rated Movies |
| 700 | 560 |

The IMDB test dataset is used to find the overall ratings. From the above results it is interesting to know that 560 movies are predicted correctly i.e, 80% of movie ratings are accurate.

**Key Challenges**

* Due to high volume, the dataset contains huge amount of noisy, missing and inconsistent data. The main issue with the dataset is that it contains missing fields. The central tendency method for the attribute was adopted to overcome this issue. As a central tendency both mean and median are used and the duplicate items present in the dataset are removed.
* It took so much effort to obtain the information from various sources as it needed to match each film with the respective scores. And also I am interested to combine reviews from micro blogging websites such as twitter in order to achieve better accuracy, as the data is not readily available, so I was not able to execute it.
* Working with text data requires so much of hard work, the problem is it has so many features to consider, and handling nearly 50000 features requires so much of processing power.

## Results

The work done by Pang & Lee [Bo Pang and Lillian Lee. (2002)] in this domain achieved a best accuracy of 87.2%. When compared with the previous work done on sentiment analysis of movie reviews, incorporating SentiWordNet in the pipeline helped in achieving the classification accuracy of 85%.

**Impacts and Improvement:**

There are multiple ways to explore regarding the dataset, algorithms and features. Based on results it shows that, merely using reviews form IMDb is not an optimal method for improving the accuracy. The accuracy can be improved by considering features beyond IMDb, i.e. considering reviews from disparate social networking sites like twitter. And also such a system can benefit from getting insights into the opinions of a different population of the Internet, rather than solely depend on IMDb users.

When dealing with polarity of reviews merely depending on SentiWordNet is not sufficient. There are few words missing in the database. Either using new lexical resource or creating new one might be a better choice. Few online reviews do contain smileys and they determine the overall polarity of the review irrespective of sentence. Adding a smiley evaluation module in the methodology helps in improving the situation

**Visualization tools**

This paper uses built-in python tools for visualization purpose. Jupyter Notebook and Google Collaboratory platforms are used to introduce Matplotlib and Seaborn plotting libraries.

## Discussions

**Significance**

The significance of sentiment analysis is very broad and powerful. The ability to extract insights from social data is a technique that is being adopted in world wide organizations. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market [*Sentiment Analysis: How Does It Work? Why Should We Use It?* ]. In 2012, Obama administration implemented sentiment analysis technique to extract insights from the public opinions and sent campaign messages and policy announcements to the targeted individuals prior to the presidential elections.

**Insights**

This technology is currently being used in array of platforms like e-commerce, online advertising, Netflix movie recommendation etc. Below are few applications used in this domain.

1. Decision Making: Customer’s reviews play a major role to make a decision. Before buying a product or going for a new movie users read the reviews of that particular product/movie and those reviews have great impact on user’s mind.
2. Improving the quality of products: Using sentiment analysis, manufacturers can gather feedback from customers about their goods, whether favorable or not, and they can enhance product performance.
3. Recommendation system: By evaluating and classifying customer reviews, the system can decide which movies be suggested and which one should not be recommended

## 

## Recommendations

**New Trends**

Sentiment analysis has gained a lot of attention in recent years with increase in micro blogging websites. Current trends in this field are emotion detection, Multilingual analysis, Aspect based analysis etc. The techniques, algorithms used in sentiment analysis field are also evolving rapidly. In recent years the popularity of deep learning techniques in other areas has led to its use in Sentiment analysis. Nowadays, Recurrent Neural Networks (RNN) algorithm is the most popular ones used for sentiment classification tasks.

## Future Work

Further study is to combine information from various sites such as twitter and YouTube comments in addition to IMDB user reviews. And also this work will continue to implement sentiment analysis technique on other domains like product reviews, newspaper articles, political forums etc. In next two semesters, this paper will focus on deep neural network methodologies specifically on Long-term-short-memory (LSTM) and Recurrent Neural Network.

## References

1. (n.d.). Retrieved from https://www.python-course.eu/confusion\_matrix.php

*A Quick Introduction to K-Nearest Neighbors Algorithm*. (n.d.). Retrieved 04 11, 2017, from Noteworthy - The Journal Blog: <https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7>

2. Allotey, D. A. *SENTIMENT ANALYSIS AND CLASSIFICATION OF ONLINE REVIEWS USING CATEGORICAL PROPORTIONAL DIFFERENCE.* UNIVERSITY OF REGINA, COMPUTER SCIENCE. 2011.

3. Garbade, G. &. (n.d.). *A Simple Introduction to Natural Language Processing* . Retrieved from <https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32>

3. Goyal, A. *Sentiment Analysis for Movie Reviews.*

4. Hamouda, A. *Reviews Classification Using SentiWordNet Lexicon.*

5. https://www.expertsystem.com/machine-learning-definition/. (n.d.).

6. Lee, B. P. (2002). Thumps up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing . *EMNP* .

7. M, V. K. *User specific product recommendation and rating system by performing sentiment analysis on product reviews.*

8. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP 02. (2002). *Thumbs up?*

*9. Python Machine Learning Tutorial*. (2019, 05 13). Retrieved from Machine Learning with Python: Confusion Matrix in Machine Learning with Python: <https://www.python-course.eu/confusion_matrix.php>

10. Sahu, T. P. *Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms.* National Institute of Raipur.

11. Sanjeevi, M. (2017, 09 06). *Decision Trees Algorithms*. Retrieved from Medium: https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1

12. Scaria, A. T. *Predicting Star Ratings of Movie Review Comments.* Stanford University.

*Sentiment Analysis: How Does It Work? Why Should We Use It?* (n.d.). Retrieved 13 05, 2019, from Brandwatch: <https://www.brandwatch.com/blog/understanding-sentiment-analysis/>

*13. Sentiment Analysis: How Does It Work? Why Should We Use It?* (n.d.). Retrieved 05 13, 2019, from Brandwatch: https://www.brandwatch.com/blog/understanding-sentiment-analysis/

14. Thomar, D. S. *A Text Polarity Analysis Using Sentiwordnet Based an Algorithm.*

User specific product recommendation and rating system by performing sentiment analysis on product reviews. (2017, 01). *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* .