International Journal of Pure and Applied Mathematics
Volume 118 No. 20 2018, 3277-3284
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Sentiment Analysis of Movie Reviews Using Machine Learning Algorithms - A Survey

J Sai Teja, G Kiran Sai, M Druva Kumar,R.Manikandan

School Of Computing, SASTRA Deemed to be University, India.

Abstract

Sentimental analysis also known as "Opinion mining" deals with analyzing the emotions and classifying the opinions from text. It refers to the use of natural language processing, text analysis and computational linguistics. With the rapid development of web technology, a huge amount of data is being generated. Social networking sites like Facebook, Twitter, YouTube and Instagram are gaining a lot of popularity as they allow users from different parts of the world to share their views upon various topics through comments, posts, tweets and tags. The research article provides a survey of existing techniques for sentimental analysis like machine learning and lexicon based approaches. Using different algorithms like Naïve Bayes, Max Entropy, SVM and Ensemble classifier a research on different data streams has been provided.

Keywords: Sentiment Analysis, Feature Extraction, Opinion Mining, Logistic Regression

1. Introduction

The advancement in the field of web technology has changed the way in which people can express their views. People depend upon this user generated data for analysis of products while shopping online or while booking movie tickets for watching movies in theatres. The users are connecting together through posts, Facebook, tweets and hash tags. The amount of data is so huge that it is difficult for a normal human to analyze and conclude.

Sentiment analysis is mainly concerned with the identification and classification of opinions. It is broadly classified in the two types first one is a knowledge based approach and the other machine learning techniques [1]. First approach requires a large database of predefined emotions and an efficient knowledge representation for identifying opinions. On the other hand the Machine learning approach makes use of a training data set and a test data set to develop a classifier. It is rather simpler than Knowledge base approach.

Since the development of algorithms several challenges were faced in the field of Sentiment analysis. The first is that an opinion word can be positive or negative depending upon the situation. The second challenge is that people don't always express opinions in the same way. Opinion mining helps to understand the relationship between textual reviews and the consequences of those reviews.

2. Sentimental Analysis

Sentimental analysis can be used to differentiate customers and followers based on their attitude towards a brand or a movie or a product with the help of reviews. One can find whether the product review is positive or negative or whether the user email is satisfied or not.

Major steps involved in sentimental analysis:

- Preprocessing of datasets

- Feature Extraction & Feature Selection
- Classification Model

2.1 Preprocessing of datasets:

Incomplete and noisy data are common properties of real world databases. Data needs to be cleaned before it is processed for better efficiency and accurate outputs. Table 1 provides different sources from where datasets can be collected. There are a lot of preprocessing techniques and when it comes to Sentimental analysis it is the techniques include cleaning and preparing the data for classification. Other preprocessing techniques include Integration and Transformation, Aggregation and Discretization.

Preprocessing of tweets or posts include removal of URL's, punctuations, symbols, emoticons and stop words.

Table 1. Sources where the data sets can be found

Tweets	demeter.inf.ed.ac.uk
Opinions	patientopinion.org.uk
Tweets	inclass.kaggle.com/c/si650winter11/data
Opinions	www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html

2.2 Feature Extraction& Feature Selection

Rohini S. Rahate [2] has proposed 4 categories of Feature Extraction.

- Syntactic Feature
- Semantic Feature
- Link based Feature
- Stylistic Feature

The most commonly used features are the first 2 features. Syntactic feature uses word tags, patterns, phrases and punctuations. On the other hand Semantic feature works on the relation between words, signs and symbols. Linguistic semantics can be used to understand human expression through language accurately.

Akshay Amolik [3] proposed a method where Feature Extraction can be done in 2 phases. Features from tweets are extracted in two phases. In the first phase, features related to twitter are extracted. They have replaced all the hashtags "#" with the exact same word by removing # sign. The Twitter specific features may not be present in all tweets. So they further extracted the tweet to obtain more features.

Zena M. Hira and Duncan F. Gillies [4] in their paper discussed about three Feature Selection methods and also compared differences between Feature Selection and Feature Extraction methods as shown in Table 2. The selection methods include the following:

- Filter
- Wrapper
- Embedded technique

Filter methods can be generally used as preprocessing technique and no algorithms are involved. Features in this method are selected based upon their scores in various statistical tests for their correlation with the outcome variable.

In Wrapper methods, the model is trained using a subset of features. These methods are computationally expensive. Some common types of wrapper methods are forward feature selection, backward feature elimination and recursive feature elimination.

In embedded methods, the features of filter and wrapper methods are combined. It's implemented by algorithms that have their own built-in feature selection methods.

Table 2. Differences between Feature Selection and Extraction [4]

Method	Advantages	Disadvantages
Selection	Preserving data characteristics for interpretability	Discriminative power Lower shorter training times Reducing overfitting
Extraction	Higher discriminating power Control overfitting when it is unsupervised	Loss of data interpretability Transformation maybe expensive

Tirath Prasad Sahu[5] proposed a new methodology for feature extraction. They extracted features that have a strong impact on determining the polarity of reviews. They also applied computation linguistic methods for preprocessing. Among the few classification techniques they used, Random Forest had the highest accuracy of 88.95%.

2.3 Classification Model

Classification is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes. This is the learning step and is done using the training data, where a classification algorithm builds the classifier by learning from that training set made up of either database tuples and their associated class labels or simple text. It is also known as "Supervised learning". Below are the few examples of classification algorithms in Machine Learning:

1. Linear Classifiers: Logistic Regression, Naive Bayes Classifier
2. Support Vector Machines
3. Decision Trees
4. Random Forest
5. Neural Networks

The output of the analysis depends upon the algorithm chosen and also the feature vector selected. A classification model draws some conclusion from observed values. Given an input to a classification

model, it will try to predict the outcome. Outcomes are labels that can be applied to a dataset. For example, when filtering service providers “good” or “bad”, when looking at product reviews data, “success”, or “failure”.

Neethu M S and Rajasree R [1] used four different algorithms Naive Bayes, SVM, Maximum Entropy and Ensemble classifiers. All these classifiers had almost similar accuracy. The feature extraction technique was similar to that of Akshay Amolik’s [2]. Naive Bayes has better precision compared to the other three classifiers, but slightly lower accuracy and recall. SVM, Maximum Entropy Classifier and Ensemble classifiers have similar accuracy, precision and recall. They obtained an accuracy of 90% whereas Naive Bayes has 89.5%.

Akshay Amolik and Dr.M.Venkatesan [3] in their paper used two machine learning classifiers Naïve Bayes and Support vector machine. Naïve Bayesian and Support vector machine performs well and also provide higher accuracy. The results show that they got 75 % accuracy form SVM and 65% accuracy form Naïve Bayesian classifier. They concluded that the accuracy of classification can be increased by increasing the training data.

Nagamma P and Pruthvi H.R [6]in their paper focused on classification using clustering and it resulted in good accuracy. Using clustering with a classification model reduces the data used for prediction. Hence the classification with or without clustering gave the same result as the data used for prediction is small. They concluded that the effect of using clustering with classification model could be seen predominantly if the dataset used is large.

3. Proposed Model

3.1 Regression

Regression analysis is one of the statistical techniques that can be used for sentimental analysis. Given a label y, the relation between label and data features X can be expressed as: $y=f(X)$. There are a lot of regression techniques available and few of them are:

- Linear regression
- Logistic Regression
- Nonparametric regression

3.2. Logistic Regression

Logistic regression can be binomial, ordinal or multinomial. Binomial,otherwise known as binary logistic regression deals with situations where the observed outcome for a dependent variable has only two possible types, "0" and "1" (which may represent, "yes" vs. "no" or "true" vs. "false").The second type, Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., "Movie A" vs. "Movie B" vs. "Movie C"). The third type, Ordinal logistic regression deals with dependent variables that are ordered. A logistic function is used to determine the relationship between categorically dependent and one or more independent variables.

Advantages of Logistic Regression:

- It is much more robust to correlated features.
- If two features f1 and f2 are perfectly correlated, regression will simply assign half the weight to w1 and half to w2.
- It is discriminative.

- It works well on large datasets when compared with naïve bayes.

Any classification algorithm works well and produces better output only when the data is consistent. So preprocessing step in sentimental analysis plays a major role in training and testing of data. If data is not processed or partially processed the accuracy may differ. A good Preprocessing step must implement all the following to every tweet to produce better output-

- Convert @username to USER_MENTION.
- Remove the unnecessary white spaces.
- Replace #HASTAG with only the word by removing the HASH (#) symbol.
- Replace all the numeric terms.
- Remove all the STOP WORDS.
- Replace emoticons with either EMO_POS or EMO_NEG.
- Replace all Punctuations.
- Convert words with more than two letter repetitions to two letters. Eg: happppy to happy.

Other steps that can be followed to improve the accuracy are:

- Using an exhaustive stop word list- Treating these types of words as a feature in classification would not result in better outcome. So these words can be ignored and removed during the preprocessing.
- Eliminating features with very poor frequency- Keywords that occur in lesser frequency usually does not play a role in text classification.User can get rid of these, resulting in better accuracy.
- Normalizing the dataset- Words are often used with different variations. If these words are reduced to their root words, it would result in efficient performance of the algorithm. Lemmatization and Stemming are techniques that are commonly used for normalizing the dataset. For example, words:
    - Smile
    - Smiling
    - Laugh
    - Laughing

  All the above words can be reduced to the root word “smile”.

- Looking beyond Unigrams into Bigrams and Trigrams.

Algorithm Tuning: Also known as “Hyperparameter Optimization”is the process of choosing optimal hyperparameters that produces the best output for a learning algorithm. It can be treated like a search problem through model parameter space. The same algorithm might require different parameters to produce different data patterns. Some of the approaches include-

- Grid Search
- Random Search
- Gradient-based optimization
- Evolutionary

Ensembles: Ensemble methods are concerned with combining results from multiple algorithms to get better results. It works well when there are multiple algorithms that specialize in various parts of the problem. It can be achieved in the following ways:

- Bagging
- Boosting
- Blending

4. Results Analysis

The Dataset "Deadpool.txt" is fetched from twitter using twitter API. Three major algorithms namely

Naïve Bayes, SVM and Logistic Regression are compared based on accuracy precision and recall as shown in Fig 1.

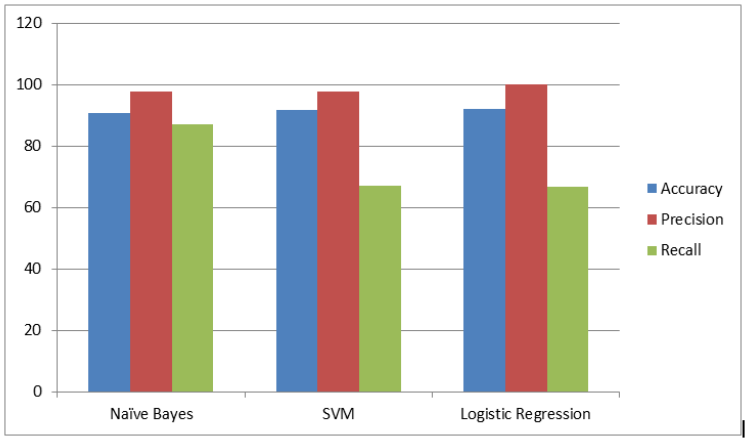

Fig 1. Performance of three algorithms

The result analysis claims that highest accuracy was achieved by Logistic Regression(92.15), followed by SVM and Naïve Bayes.

5. Conclusion

The paper mainly addresses the comparative study of different machine learning algorithms that can be used to extract sentiments from text. They are simpler and efficient. Naïve bayes and SVM with the best accuracy can be considered as a benchmark for all the other algorithms. It provides enough information that could help to improve the predictions in further research work. It can be concluded that cleaner the data, better the performance of an algorithm in predicting the success rate of the movies.

References

[1] Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-5). IEEE.

[2] Rahate, R. S., & Emmanuel, M. (2013). Feature selection for sentiment analysis by using svm. International Journal of Computer Applications, 84(5).

[3] Amolik, A., Jivane, N., Bhandari, M., & Venkatesan, M. (2015). Twitter sentiment analysis of movie reviews using machine learning techniques. International Journal of Engineering and Technology, 7(6), 2038-2044.

[4] Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.

[5]Sahu, T. P., & Ahuja, S. (2016, January). Sentiment analysis of movie reviews: A study on feature selection & classification algorithms. In Microelectronics, Computing and Communications (MicroCom), 2016 International Conference on (pp. 1-6). IEEE.

[6]Nagamma, P., Pruthvi, H. R., Nisha, K. K., & Shwetha, N. H. (2015, May). An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In Computing, Communication & Automation (ICCCA), 2015 International Conference on (pp. 933-937). IEEE.

[7] Trupthi, M., Pabboju, S., & Narasimha, G. (2017, January). Sentiment analysis on twitter using streaming API. In Advance Computing Conference (IACC), 2017 IEEE 7th International(pp. 915-919). IEEE.

[8] Ahmed, E., Sazzad, M. A. U., Islam, M. T., Azad, M., Islam, S., & Ali, M. H. (2017, March). Challenges, comparative analysis and a proposed methodology to predict sentiment from movie reviews using machine learning. In Big Data Analytics and Computational Intelligence (ICBDAC), 2017 International Conference on (pp. 86-91). IEEE.

[9] Wankhede, R., & Thakare, A. N. (2017, April). Design approach for accuracy in movies reviews using sentiment analysis. In Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of (Vol. 1, pp. 6-11). IEEE.

[10] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.

[11] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010).

[12]Krouska, A., Troussas, C., & Virvou, M. (2016, July). The effect of preprocessing techniques on Twitter Sentiment Analysis. In Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on (pp. 1-5), IEEE.

[13] Bahrainian, S. A., & Dengel, A. (2013, December). Sentiment analysis and summarization of twitter data. In Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on (pp. 227-234),IEEE.

[14] Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: A survey of techniques. arXiv preprint arXiv:1601.06971.

[15] Bahrainian, S. A., & Dengel, A. (2013, November). Sentiment analysis using sentiment features. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03 (pp. 26-29). IEEE Computer Society.