ISSN No. 0976-5697

Volume 11, Special Issue I, May 2020

International Journal of Advanced Research in Computer Science

CONFERENCE PAPER

Available Online at www.ijarcs.info

SENTIMENTAL ANALYSIS FOR MOVIE REVIEWS

Kumar Abhishek[1], Mayank Mehiral[2], M. S. Sathvik Murthy[3]
[1,2,3] Computer Science Department, REVA University, Bengaluru, India
Corresponding Author: raghavendrareddy@reva.edu.in

Abstract- - Sentimental analysis is defined as the use of computational linguistics, natural language, text analysis and biometrics to recognise,extract,test and study useful attributes and their information. We considered two different datasets both pre-dominantly pertaining to IMDB as source. One of the considered datasets composed only textual content which was processed by removing unnecessary contents and distributed into two categories namely positive and negative. We further divided the data into training dataset and testing dataset. Using more relevant training algorithms such as logistic regression and decision tree algorithm, we had more relevant attribute which helped us in training our model to predict if a review is positive or negative. When this analysis is linked with other attributes of any product of interest, we can accurately pin point or predict a product's rating even before it sees broad day light.

Keywords - Sentimental Analysis, Tokenization, Data pruning, Accuracy, Polarity.

INTRODUCTION

Sentimental analysis of movie reviews helps in understanding the extent to which the movie has impressed its audience. And it also gives the individuals hoping to see the movie, a heads-up weather to watch it or not. In the bigger picture sentimental analysis may be applied to a larger set of scenarios like products, people and things.

Sentimental analysis is nothing but the process of tokenizing the input in terms of given sequence of words checking if they are positive, negative or neutral in context because in general the reviews are used as a tool to keep productivity in check. And since the all have common review in terms of phrases of words the trained model can cater to larger set of audience with little or no changes.

Internally the trained model is trained to decide if a word belongs to any of the above mentioned category by extensive exposure to the dataset content where it learns to effectively distinguish between what is good, bad or neutral, after which it sums up the polarity of every token in order to arrive at a single value, in order to provide a cumulative polarity.

Among the two datasets used, first dataset was pruned and cleaned from all the anomalies in order to effectively train the model under purview. The model was initially trained with plain linear regression model in order to test if the model is trainable and later it was trained with logistic regression which yielded us 88 percentiles of the accuracy.

The second dataset enables us to predict the score of the movie by considering attributes like the technicalities, production, resources involved and many more. It could accurately predict a movies score by comparing all the specified parameters. Unlike first dataset, the second dataset contains more logical accumulation of interrelated data which can provide meaningful insights in understanding how the movie will be received in the future of its introduction.

The paper tries to establish a way for the reader to understand the various approaches to tackle the problems arising the umbrella of Sentimental Analysis. The goal of this paper is to draw out the various comparisons between various classifiers used to train ML model and compare them using the accuracies or confusion matrices in particular. This will allow the readers to arrive at conclusion as to choose the right algorithms for the sentimental analysis process. Further paper finalizes by demonstrating the result in the form of the bar graph of the tokens which occur most often while dealing with reviews of any context in the real world.

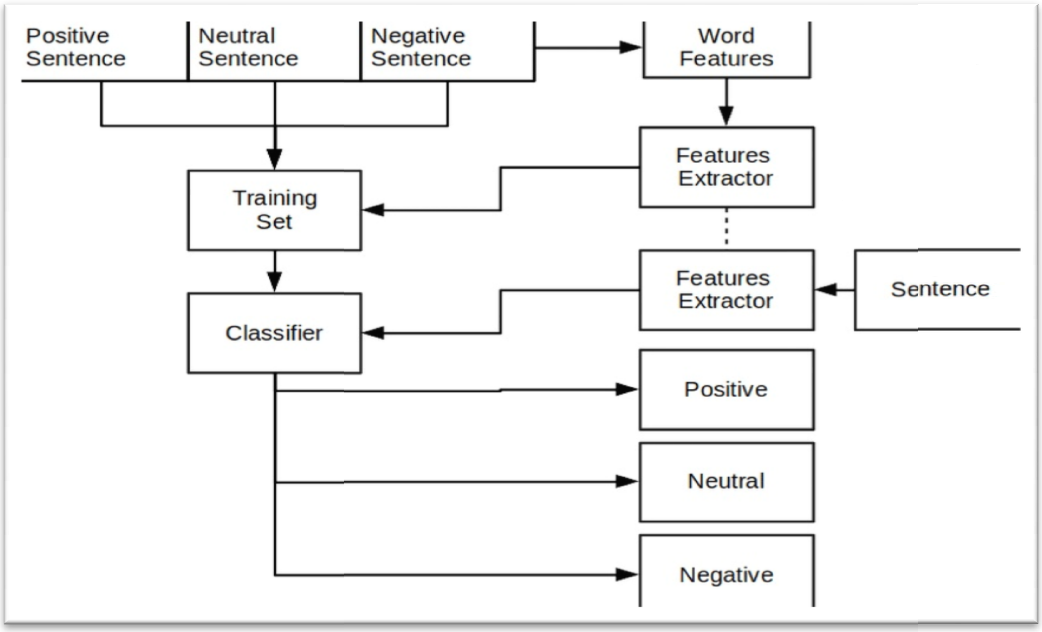
Literature Survey

Ref.	Methodology	Advantages	Disadvantages	Accuracy

[1]	SVM, Random Forest, RNN, Logistic Regression on Kaggle datasets.	Considers well known datasets from Kaggle which is more application friendly with IMDB.	Less efficient SVM classifier is used.Dataset restricted to product reviews from only one website.	71%
[2]	Keyword spotting, lexical affinity, SVM classifier and statistical methods.	Uses Naïve Baye's Classifier for training datasets. SKLearn Module of Python is used efficiently.	Provides more generic approach towards the problem. It prefers the readers to experiment on classifiers and then decide accordingly.	67%
[3]	Rains model using NB, SVM classifiers. Tokenizes the reviews in order to analyze them.	Uses well known datasets of Rotten-tomatoes as reference for training model.	Consists of many complex theoretical formulas which need prerequisite knowledge.	62%
[4]	Use NLP and Polarity classification on the results.	Takes into account Heuristics of structured data. Provides deep understanding of classification.	Since it works on mostly unstructured data, the training overhead of model will be high.	57%
[5]	Natural Language Processing with respect to emotions. Subjectivity of sentiments.	Introduces the subjectivity related to text well. Gives insights into the NLP of sentiments.	Shows moderate scope for implementation.	45%
[6]	Machine learning (ML) techniques such as Naive Bayes, SVM, and Max Entropy	Gives good incites on ambiguity handling of word polarity and identification of polarity.	Gives too much attention to theoretic backdrop of the machine training process.	53%
[7]	Text tokenization, training with twitter dataset.	Provide polarity for the messages being exchanged on twitter.	Too much variety in data and complexity in training model with this data set.	65%
[8]	Tensor factorization, hybrid factorization	Predict the rating of the movie using textual reviews.	Hybrid complex factorization using multi-dimensional matrices.	78%
[9]	NLP sub-tasks, subjectivity detection, concept extraction.	Generalize the given review into a numeric value.	Uses NLP strategies like multimodal fusion on 'Mandarin' language.	45%
[10]	Feature Optimization, Genetic Algorithm.	Lexicon based approach gives better results.	Requires proper preparation of training dataset.	67%
[11]	General approach to Sentimental analysis.	Outlines the way to understand sentimental analysis.	Gives a vague idea of how sentimental analysis is meant to work.	34%

[12]	CNN, RNN, Long Short-Term Memory Network (LSTM).	Provides a way to induct in depth approach to Sentimental analysis.	Use of CNN and RNN will have higher computation overhead on the training process.	51%
[13]	K-means Clustering, Peak-search Clustering.	Makes use of domain specific dataset of Amazon reviews.	Restricts the research to a specific domain, limiting its usage with respect to other domains.	68%
[14]	Feature extraction - Bag of words, Part of speech tags.	Parameters of dataset used for training are simple.	Uses Naïve Bayes Classifier	42%
[15]	Recursive auto-encoder, stacked auto-encoder	Compares and contrasts the results obtained by different classifiers.	Lacks the technical explanation of how to implement the technology.	33%

METHODOLOGY



The methodology of implementation begins by using the training set in order to teach the NLP model using the data mining classifiers like logistic regression or SVM. The classifier is used to enable the model to teach itself with the assist of features. The feature extractor goes through the training dataset to discover the tokens which occur most often than not. Later when the model tests, it searches for these same tokens. These tokens are then used to assign polarity and predict if the review is positive or negative. In order to understand how our model works we'll have to know that a sentimental analysis extensively deals with tokenization of day as today's models are not advance enough to obtain contextual/subjective knowledge without doing so.

Thus, we will have to train the ML model with the relevant datasets in order to make it familiarize it with various knowledge base of texts and other csv formats of available data. The thoroughly pre-processed dataset without any anomalies is used to train our model. When the model is fed with such datasets it tends to breaks

down each of the text in the form of word features by assigning them relevant tags which can be used by it in the future. Later these tags are used to extract features from the same tags in the form of word tokens which we read about in our abstract section of the paper. These features are used to train the model which is undergoing learning in the other end of the flow. Later the classifier uses the same set of features which were extracted while predicting if a sentence into the appropriate category. This is the general flow of processes in any sentimental analysis project.

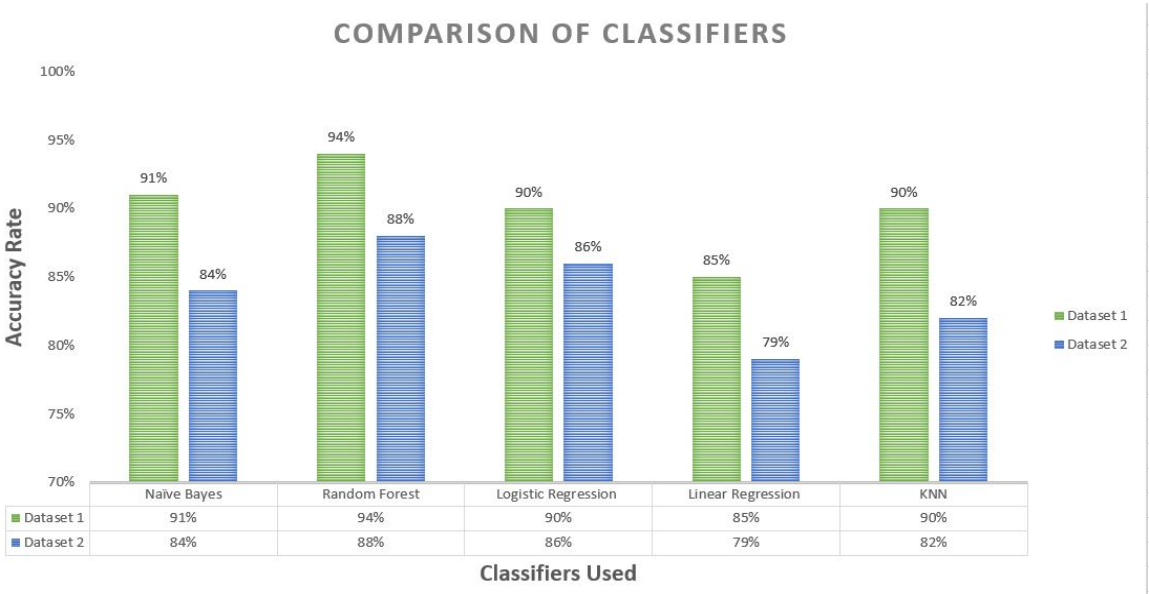
Similarly, our implementation of the sentimental analysis project the ML model being trained breaks down a given movie review paragraph into individual tokens which are compared with the tokens learnt by the model. After tokenizing the contents, it identifies and assigns a value to each token ranging between -1 to 1. Negative value represent negative token and positive value represents the positive token. At the end it just sums up the value of each token to arrive at a conclusion weather a given review is positive or negative as whole in the same way it classifies if a token is positive or negative.

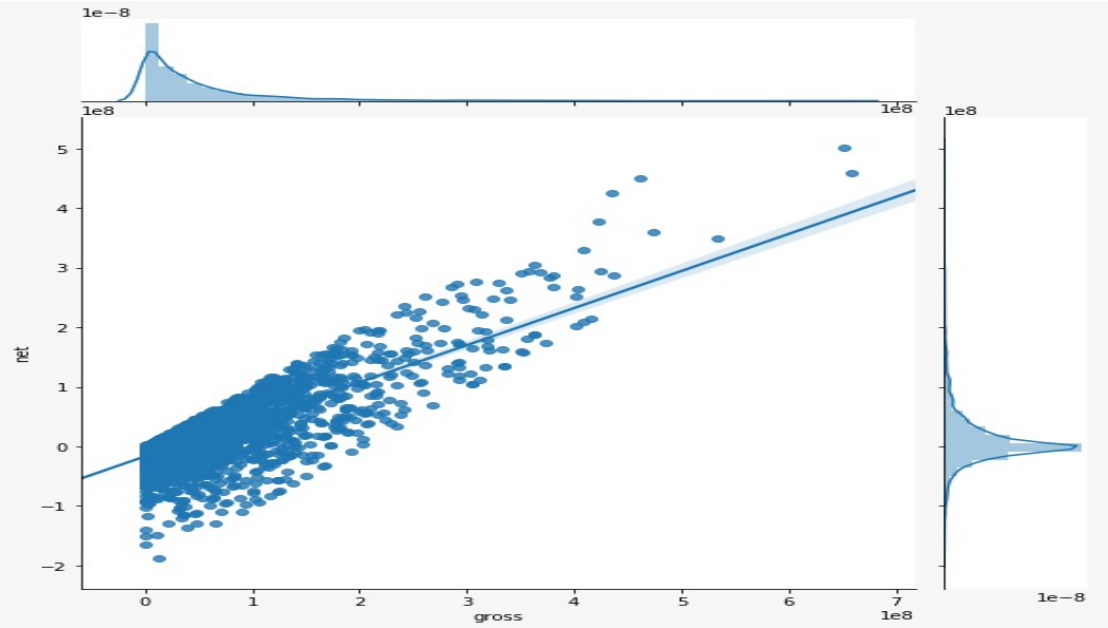
Result

Result in the form of bar graph depicts the identified tokens which were set aside while the model was being trained. These tokens are the ones which are found to be repeating most often. The tokens represented red bars were most often used in negative folder of training dataset and the trained model associated them with negative reviews. And the tokens represented by blue bars were often found in positive folder and thus will be associated with the positive aspects of a review imperatively. Further they are used to predict if a given review entered is positive or not.

For example, consider the token 'the worst' token which is found in the result. When the classifier comes across this token in a sentence it doesn't have to do much in order to find its polarity as it has already found the polarity of this particular token which is around -2.4 approximately. Thus, when it is added against the polarity of the other tokens after tokenization the review can be categorized based on the final sum of the polarity.

The accuracy of trained model which was found to be 0.88 or in other words 88 %. It can be increased by finding more relevant and useful training dataset.





CONCLUSION

The proposed system can achieve a very good accuracy with regards to the prediction of movie reviews when relevant attributes of that movie provided. Thus, the with respect to the results produced we can deduce that the model is good enough to effectively categorizes the reviews into positive, neutral or negative. Anybody interested in knowing how its product would perform can use this model by gathering data of similar products of its rivals and use it to compare it against its own product to get meaningful insights of where their product stands.

In other words, the results obtained from this project can be used as a leverage to draw business strategies in the field of movies in specific and product of interest in general. For example, the reviews drawn from datasets with respect to the products of amazon can be used to train similar model and can be used to decide if the products are satisfying the customers who have bought

product. Thus, it can affect various decisions about to be made in the future like 'Should the sellers continue to sell that product?' or 'Can we provide better services to customers by combining it with any other product in order to make the customer happier?'. This model has two-way benefits one being predicting if a section of text has positive or negative polarity and other being predicting the rating of product on the scale of one to ten. The trained model provides an accurate and dependable way to understand product in better way.

References

1. HadidPour Ansari, Saman Ghili, Stanford University, "Deep learning for sentimental analysis of movie reviews", Vol-2 pp 123-130, 2015.
2. Ankit Goyal, Amey Parulekar, "Sentimental Analysis for Movie Reviews", IJEC journal Sl no: 20301, pp 80-110, 2017.
3. [4]Pang, Bo; Lee, Lillian from Vaithyanathan Cornell University, Six outcomes of machine learning on sentimental analysis, 2015.
4. [2]Cambria, Erik; Schuller, Björn; Xia, Yunqing; Havasi, Catherin, "New Avenues in Opinion Mining and Sentimental Analysis.", 2016.
5. [5].J. Wiebe, T. Wilson, and C. Cardie,"Annotating Expressions of Opinions and Emotions in Language.", 2017.
6. [6][7]Mamatha M, Thriveni, Venugopal, "Techniques of Sentimental Classification: A Comprehensive Review", 2016.
7. [7]Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentimental Analysis", 2018.
8. [8]Xiaojian Lei, Xuening Qian, Member, IEEE, "Rating Prediction Based on Social Sentimental from Textual Reviews", 2016. DOI: 10.1109/TMM.2016.2575738,2016.
9. [9]Haiyung Peng, Erik Cambria, Amir Hassain, "A Review of Sentimental Analysis Research in Chinese Language",2017. DOI 10.1007/s12559—17-9470-8.
10. [10]Farkhund Iqbal, Jahanzeb Maqbool, "A Hybrid Framework for sentimental Analysis

using Genetic Algorithm based feature reduction", 2019. DOI 10.1109/2019.

11. [4][5]Doaa Mohey, El-Din Mohammed Hussein, "A survey on sentimental analysis challenges". ES 2016/1018-3639. 2016.
12. [10]Jin Zheng, Limin Zheng, "A Dictionary-based Convolutional Recurrent Neural Network Model for Sentimental Analysis", 2019. DOI 10,1109/CISCE.00142 .2019.
13. [2]Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In EMNLP, vol. 14, pp. 1532-1543.2014.
14. [13][14]Chantal Fry, Sukanya Manna, "Can we Group Similar Amazon Reviews: A Case Study with Different Clustering Algorithms", 2016. DOI 10.1109/ICSC.2016.
15. [13]Callen Rain, "Sentimental Analysis in Amazon Reviews Using Probabilistic Machine Learning", AES 2017/2017.
16. [15]Hanen Ameur, Salma Jamousei and Abdelhamid Ben Hamadao, "A New method for Sentimental Analysis using Contextual Auto-Encoders", 2018/ SPR-2018.
17. [2][3]W. Medhat, A. Hassan and H. Korashy, "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal, Vol 5, Issue 4, Pp. 1093-1113, 2018.
18. [1]Rafael M., D'Addio, Marcos A., Domingues, Marcelo G., and Manzato, "Exploiting feature extraction techniques on users reviews for movies recommendation", Journal of the Brazilian Computer Society, Vol.23, Pp-7, 2019.
19. [4][9]P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction," 2018.
20. [13]H. Saif, M. Fernández, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter using K-means," in Proceedings of the 9th International Conference on Language Resources and Evaluation 2019.
21. [14][15]S. A. Alasadi and W. S. Bhaya, "Review of data pre-processing techniques in data mining, using Encoding techniques" J. Eng. Appl. Sci., 2018.
22. [8][9]Bhumika M. Jadav M.E. Scholar, L. D. College of Engineering Ahmedabad, India-Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis, International Journal of Computer Applications Volume 146 –No.13, July 2019.
23. [7]Zhao Jianqiang1, Gui Xiaolin1- Deep Convolution Neural Networks for Twitter Sentiment Analysis IEEE ,2018.