| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 11, November 2019 |

Predicting Sentiment from Movie Reviews Using Machine Learning Approach

Binita Verma, Ramjeevan Singh Thakur

B.U Bhopal, India

MANIT, Bhopal, India

ABSTRACT: With the rapid growth of Text sentiment analysis, need of automatic classification of digital documents has increased. Predicting of events is also a challenging task in the near and distance further. People are interested in forecasting weather, prediction economic, political and social changes and movies outcomes. Movie reviews play an important role in the task such as user submitted reviews, ratings and user movies recommendations. The ability to predict the sentiment of a movie would be useful considering these aspects. In this paper we propose a model to predict the sentiment of the text based movie reviews and applied on two different text mining supervised machine learning algorithms. We found that SVM is the most appropriate to work with our proposed model. The achieved result show significantly increases in accuracy compared to earlier method.

KEYWORDS: Supervised machine learning, Movie reviews, Text classification.

I. INTRODUCTION

On a variety of online platforms, such as review sites, blogs, as well as social services such as Twitter, Facebook, Instagram, Internet users produce vast amount of opinionated text of movie reviews, travel experiences, product reviews, opinions about news and others [1][2]. Automatic opinion mining is the ability to produce large amounts of opinionated text information from online sources without human interference [3]. A lot of research works have been done to gather and calculate the sentiment of posts and tweets and also a good number of text mining algorithms have been designed to analyze the sentiments.

Increasing the number of users and the data related to them has provided a good impetus to every company or organizations to mine these micro-blogging sites to collect information about people's opinion about their services or products. Due to this increase in user interaction, the future sales of any product or services depends a lot on the sentiments and perceptions of the previous buyers [4]. Therefore, it is necessary to have an efficient way to predict user sentiments about a product or service.

The solution of these problems is to classify the text using a strong machine learning algorithm. Human's face many decisions on a daily bases and sentiment analysis can automate the process of coming to a decision based on past outcomes of that decision. For example, if someone has to buy tickets for a movie, then rather than manually going through all the ling reviews, a sentiment classifier can predict the overall sentiment of the movie. Based on positive or negative sentiment a decision can be taken whether or not to buy tickets. Text classification can be used in many different areas such as:

- Understanding audience sentiments from social media,
- Detection of spam and non-spam emails [5]
- In medical science, analyze and categorizes reports, and hospital records [6].
- Auto tagging of customer queries [6]
- Categorization of news articles into defined topics

Sentiment analysis is utilized in order to determine the polarity of opinions like positive, negative or neutral or the emotional charge of opinions across a range of possible emotions like joy, surprise, anger and fear etc. In this paper, we have to build models for predicting the sentiments of movie reviews dataset. The performance of proposed techniques with SVM [7] and Logistic regression [8] classifiers are shown at end.

II. RELATED WORK

The field of sentiment analysis has recently witnessed a large amount of interest from the scientific community[9][10][11].

Pang et al. [12] have set a standard for machine-learning based sentiment analysis. They compare the sentiment analysis to topical categorization of documents and their approach is applying machine learning techniques to the

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 11, November 2019 |

problem of sentiment analysis. The authors define the main challenge of sentiment analysis as the fact that the sentiment of a message is conveyed with much more subtle linguistic cues than in topic categorization. They also conduct their research on Movie Reviews domain.

In addition to sentiment analysis, research into the prediction of sentiment was conducted by number of researchers [13][14][15][16]. Pang, Lee and Vaithyanathan propose sentiment classification on machine learning models SVM, Logistic regression and Naïve Bayes for sentiment analysis on unigrams and bigrams of data [12]. They found that SVM paired with unigrams produced the best results.

III. MACHINE LEARNING APPROACHES

Machine learning has multiple algorithms, techniques and methodologies that can be used to build models to solve real world problems using data. Machine learning techniques are classified into supervised and unsupervised learning techniques [17].

The unsupervised learning techniques mainly use lexicon based approach where they use existing lexical resources like WordNet and language specific sentiment seed words to construct and update sentiment prediction [18]. Although unsupervised learning algorithms do not require a corpus of previously classified data and generates a general sentiment, they fail to capture context or domain specific information of the document.

The supervised learning techniques use machine learning on a previously classified to be almost accurate. These pre-classified datasets are often domain specific, therefore the model it generate can work only for a particular domain. These datasets are first converted into intermediate models where documents are represented as vectors and then the intermediate representations are fed to the machine learning algorithm. Through our research we have found out that Support Vector Machines and Logistic regression are the popular choice of algorithm for sentiment analysis. Fig 1. shows the detailed workflow for building a standard text classification system with supervised learning (classification) models.

In dataset, documents are represented as a vector and every word is converted into a number. The number can be binary (0 or 1) or it can be any real number in case of TF-IDF model. The BOW model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier [19][20][21]. If a word appears in a document it gets a score 1 and if the word does not appears it gets a score 0. The term frequency-inverse document frequency (TF-IDF) approach [22] is commonly used to weight each word in the text document according to how unique it is. In other words, TF-IDF approach captures the relevancy among words, text documents and particular categories. So, the document vector can be a list of any numbers which are calculated using term frequency-inverse document frequency method.

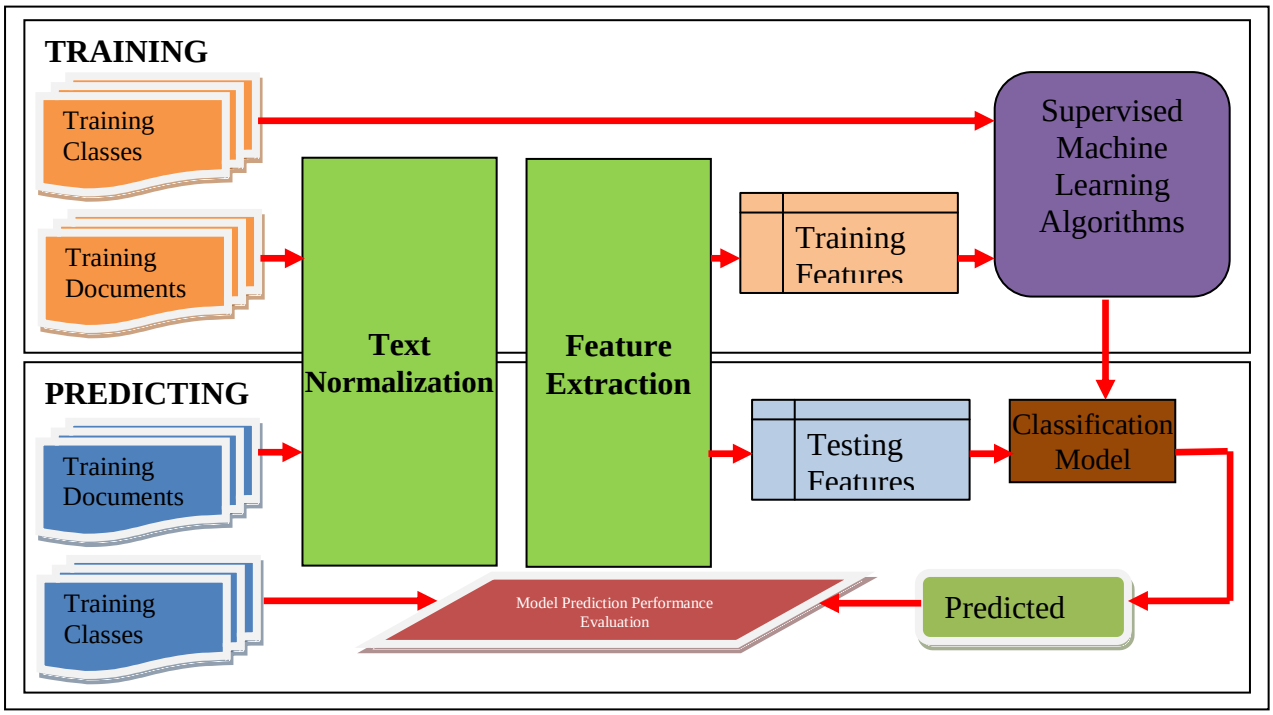

Fig 1. Text Classification system with Supervised Learning Models [23]

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 11, November 2019 |

IV. PROPOSED WORK

The proposed model consists of the various steps: collection of dataset, data preprocess/normalization and feature engineering, model selection and training, model prediction and evaluate the performance and finally deploy the model.

The overview of our proposed model is shown in fig 2. In the first phase, we have collected the dataset from the specific domain, then define the problem and preprocess that data after removing stop words, punctuations. In the next stage, feature engineering techniques are applied to input dataset, then the proposed model is prepared using the Logistic regression and Support Vector Machine. Finally, we get the evaluation model performance. The proposed work focuses in particular to predict sentiment of new incoming data.

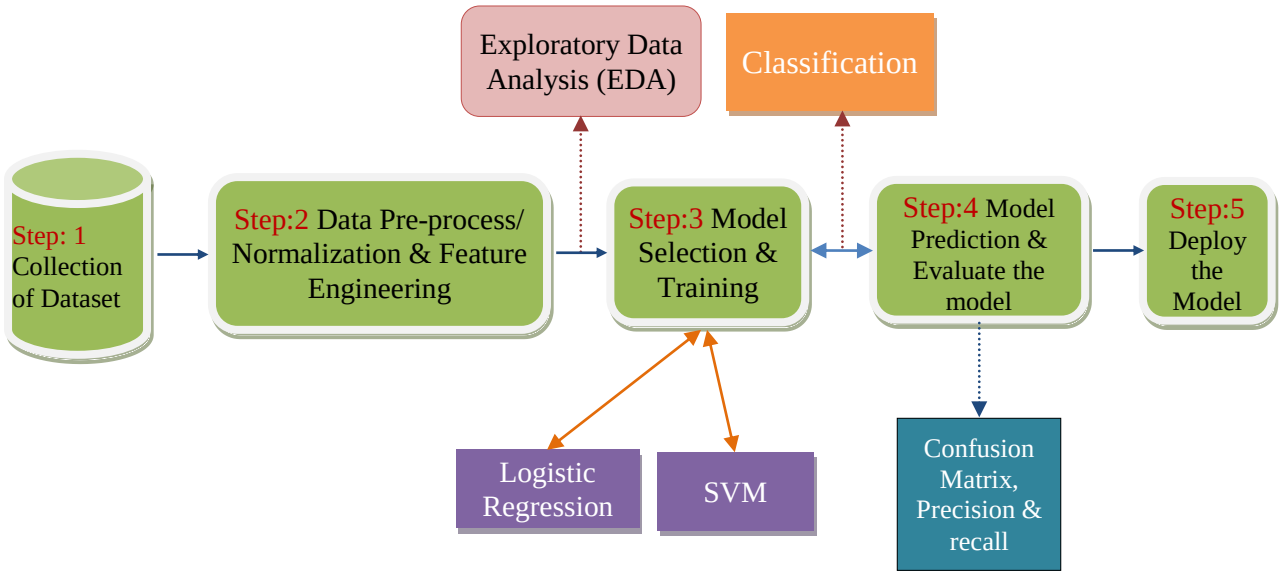

Fig 2. Proposed Model

Step : 1 Collection of Dataset

We use the movie review dataset obtained from internet movie database(IMBD), provided in [24], which is publicly available on Kaggle. For analysis, we obtained 25,000 movie reviews that have been prelabelled with 0 and1 class labels based on reviews, of which 15,000 issued for training purpose and 10,000 issued for testing purpose. The dataset contains id sentiments and reviews. We preprocessed the dataset to denote the sentiments by 0 as negative and 1 as positive. We evaluate model performance on the testing data.

Step: 2 Data Preprocess/Normalization and Feature Engineering

Preprocessing helps to remove the redundant information and transforms the information into uniform format. The process of preprocessing starts from Removal of Stop words and special symbols, parsing the sentences and stemming. Parsing includes Tokenization, Lemmatization, and POS tagging. Pre-Processing and normalize text documents movie reviews dataset are now prepared and normalized, so we can proceed for text classification. we use feature engineering techniques based on Bag of Words and TF-IDF model, engineer features using both these models on our train and test datasets.

We take into account word as well as bi-grams for our feature sets. We can use some supervised machine learning algorithms which work very well on text classification. We build models using Logistic regression, Support Vector Machines.

Step: 3 Model Selection and Training

The Logistic model is a supervised machine learning model used for classification. In this model we try to predict the probability that a given movie reviews will belong to one of the discrete classes (binary classes). The function used by the model for learning is represented here [25].

$$P(y = \text{positive} \mid X) = \sigma(\theta^T X)$$
$$P(y = \text{negative} \mid X) = 1 - \sigma(\theta^T X)$$

$$\frac{1}{1+e^{-z}}$$

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 11, November 2019 |

Where the model tries to predict the sentiment class using the feature vector X and σ(z) = , where is popularly known as the sigmoid function or logistic function. The main objective of this model is to search for an optimal value of such that probability of the positive sentiment class is maximum when the feature vector X is for a positive movie reviews and small when it is for a negative movie reviews. The logistic function helps model the probability to describe the final prediction class. The optimal value of can be obtained by minimizing an appropriate cost\loss function using standard methods like gradient descent. Logistic regression is also popularly known as Max Ent (maximum entropy) classifier.

SVM model is another supervised Machine Learning algorithm that can be used for classification. It is an example of a maximum margin classifier, where it tries to learn a representation of all the data points such that separate categories or labels are divided or separated by a clear gap, which is as large as possible.

V. EXPERIMENTAL RESULTS

We have used movie review dataset for the experiments. For training and testing purposes we have split the dataset into two parts training and testing. Further, we have conducted several experiments with different classification algorithms. We use python for experimentation. Python is one of the best programming languages when it comes to machine learning and textual analytics. It is easy to learn, open source and effective in catering to machine learning requirements like processing large data sets [26]. The experiments are done on 2 GB RAM, Pentium(R) Dual - Core CPU with 3.00 GHz having window 7 Operating system and 100 GB hard drive. In this section, we present the experimental details of proposed system. The proposed model has predicted the sentiments from movie reviews test dataset, and evaluated performance.

Experiment-1: Comparison of Model Performance of two algorithm with Bag of Word feature:

Classifiers	Accuracy (%)	Precision (%)	Recall (%)
SVM	87	87	86
Logistic Regression	88	88	89

Table 1. Model performance of ML algorithm with BOW feature

In this experiment, we have compared the model performance of the two most popular algorithms to train the classifier with BOW feature. The accuracy rate precision recall from each of the algorithms is displayed in the table 1. We found that accuracy of Logistic regression produce better result than Support Vector Machines.

Experiment-2: Comparison of Model Performance of two algorithm with TF-IDF feature:

Classifiers	Accuracy (%)	Precision (%)	Recall (%)
SVM	91	90	91
Logistic Regression	87	86	89

Table 2. Model performance of ML algorithm with TF-IDF feature

In this experiment, we have compared the model performance of the SVM and Logistic regression with TF-IDF feature. The accuracy rate, precision, recall from each of the algorithms is displayed in the table 2. We found that accuracy of SVM produce better result than Logistic Regression.

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 11, November 2019 |

Experiment-3: Accuracy comparison from existing model:

Model	Accuracy
Proposed model	91%
Existing model [25]	89.91%

Table 3 Accuracy comparision of our proposed model and existing model.

The experimental result of proposed model using SVM based on TF-IDF has been compared with existing model [27]. The experimental results of both the models are shown in Table 3. The accuracy of our proposed model works a little bit better than existing model shown in fig 3.

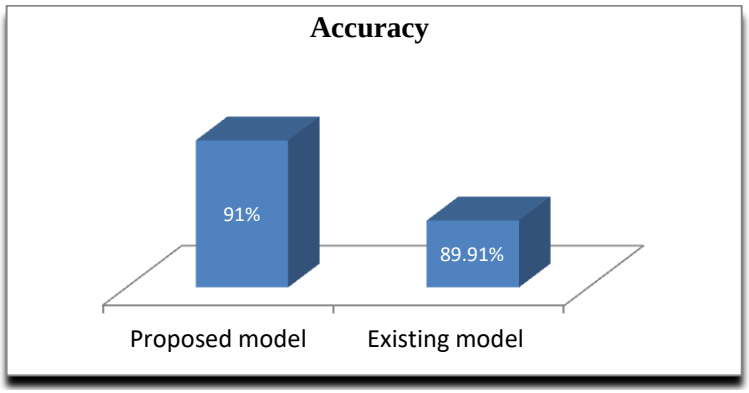

Fig 3. Comparision of Proposed Model and Existing Model

IV. CONCLUSION AND FUTURE WORK

In this paper, we have conducted experiments on IMBD movie reviews dataset, in experiment-1, we can build a SVM and Logistic regression model on BOW feature with accuracy of 87% and 88% respectively. In experiment-2, we can build a SVM and Logistic regression model on TF-IDF feature with accuracy of 91% and 87% respectively. After comparing the result of our proposed model using SVM based on TF-IDF with existing model [29]. We obtain higher accuracy than existing model. Thus we can see that how effective and accurate these supervised Machine Learning classification algorithms are in building a text sentiment classifier. In future, we can seek to improve the accuracy of this model by working on negation word.

REFERENCES

[1] Pushpendra Kumar and Ramjeevan Singh Thakur, "Recommendation system techniques and related issues: a survey", International Journal of Information Technology, Vol.10 (4), pp. 495–501, 2018.

[2] Pushpendra Kumar and R. S. Thakur, "A Framework for Weblog Data Analysis Using HIVE in Hadoop Framework", In: Proceedings of International Conference on Recent Advancement on Computer and Communication, Lecture Notes in Networks and Systems 34, 2018, https://doi.org/10.1007/978-981-10-8198-9_45

[3] Vinod Kumar, Pushpendra Kumar and R.S. Thakur, "A brief Investigation on Data Security Tools. and Techniques for Big Data", International Journal of Engineering Science Invention, Vol. 6(9), PP. 20-27, 2017.

[4] Georg Lackermair, Daniel Kailer, Kanan Kanmaz, "Importance of Online Product Reviews from a Consumer's Perspective", Advances in Economics and Business, Vol 1(1), pp. 1-5, 2013.

[5] W.A. Awad, S.M. Elseuofi, "Machine Learning Methods for Spam Email Classification", International Journal of Computer Science and Information Technology, Vol.3, pp.1-12, 2011.

[6] Mrs. Manisha Pravin Mali, Dr. Mohammad Atique, "Applications of Text Classification using Text Mining", International Journal of Engineering Trends and Technology, Vol.13(5), pp. 1-4, 2014.

[7] Steve Gunn, "Support Vector Machines for Classification and Regression", Conference Proceedings, pp. 1-66, 1998.

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 11, November 2019 |

[8] Amit Gupte, Sourabh Joshi, Pratik Gadgul and Akshay Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of computer science and Information Technology, Vol. 5 (5) , pp. 6261-6264, 2014.
[9] Go, L. Huang and R. Bhayani, "Twitter Sentiment Classification using Distant Supervision", The Stanford Natural Language Processing Group, vol. 150, 2009.
[10] Montoyo, P. Martínez-Barco and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments", Booktitle-Decision Support System, Vol. 53, pp. 675-679, 2012.
[11] S. Tan, X. Cheng, Y. Wang and H. Xu, "Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis," In *31th European Conference on IR Research on Advances in Information Retrieval*, Toulouse, France, pp-337-349, 2009.
[12] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," In *ACL-02 Conference on Empirical methods in natural language processing, Vol-10*, Philadelphia, USA, pp-79-86, 2002.
[13] G. P. C. Fung, J. X. Yu and W. Lam, "News Sensitive Stock Trend prediction", In *Advances in Knowledge Discovery and Data Mining : 6th Pacific-Asia Conference*, Taipel, Taiwan, 2002.
[14] G. P. C. Fung, J. X. Yu and W. Lam, "Stock prediction: Integrating text mining approach using real-time news," In *Computational Intelligence for Financial Engineering*, Hong Kong, 2003.
[15] R. Balasubramanyan, W. W. Cohen, D. Pierce and D. P. Redlawsk, "What pushes their buttons? Predicting comment polarity from the content of political blog posts", In *Workshop on Language in Social Media*, Portland, Oregon, USA, 2011.
[16] R. Balasubramanyan, W. Cohen, D. Pierce and D. Redlawsk, "Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News?", In *6th International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 2012.
[17] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," Intelligent Systems, IEEE , Vol.28, no.2, pp. 15-21, 2013.
[18] S. Baccianella, A. Esuli, and F. Sebastiani," SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", In Proc. of 7th Int'l Conf. on Language Resources and Evaluation, pp 2200-2204, 2010.
[19] Dani Yogatama and Noah A. Smith, "Making the Most of Bag of Words: Sentence Regularization with Alternating Direction Method of Multipliers", Proceedings of the 31st International Conference on Machine Learning, Beijing, China, JMLR: W&CP, volume 32, pp.656-664, 2014
[20] Cordelia Schmid, "Bag-of-features for category classification", www.cs.umd.edu/~djacobs/CMSC426/BagofWords.pdf.
[21] Jialu Liu, "Image Retrieval based on Bag-of-Words model", jialu.cs.illinois.edu/technical_notes/CBIR_BoW.pdf.
[22] Z. Yun-tao, G. Ling and W. Young-cheng, "An improved TF-IDF approach for text classification", School of Electronic & Information Technology, Shanghai Jiaotong University, Shanghai, China, 2004
[23] D. Sarkar, "Text Analytics with python", Apress, pp. 181, 2016.
[24] Andrew Lee Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng and Christopher Potts, "Learning Word Vectors for Sentiment Analysis", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol 1, pp. 142-150, 2011.
[25] Chao-ying, Joanne Peng and Tak-Shing Harry So, "Logistic Regression Analysis and Reporting : A Primer", Understanding Statistics, vol.1, no.1, pp.31-70, 2002.
[26] I.V. Shravan, "Sentiment Analysis in Python using NLTK", *Open Source For You*, December 2016.
[27] B. Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation", *CoRR* , Vol. abs/1806.06407, 2018.