Sentiment Analysis on Movie Review Data Using Machine Learning Approach

Atiqur Rahman, Md. Sharif Hossen

Dept. of Information and Communication Technology
Comilla University, Comilla, Bangladesh
atikriyadict@gmail.com, sharif5613@gmail.com

2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 978-1-7281-5241-7/20/$31.00 ©2020 IEEE 10.1109/ICBSLP47725.2019.201470

***Abstract*—At present Sentiment analysis is the most discussed topic which is purposed to assist one to get important information from a large dataset. It centers on the investigation and comprehension of the feelings from the text patterns. It automatically characterizes the expression of feelings, e.g., negative, positive or neutral about the existence of anything. Various sources like medical, social media, newspaper, and movie review can be used in data analysis. Here, we have collected movie review data as well as used five kinds of machine learning classifiers to analyze these data. Hence, the considered classifiers are Bernoulli Naïve Bayes (BNB), Decision Tree (DE), Support Vector Machine (SVM), Maximum Entropy (ME), as well as Multinomial Naïve Bayes (MNB). Our analysis outlines that MNB achieves better accuracy, precision and F-score while SVM shows higher recall compared to others. Besides it also show that BNB Classifier achieves better accuracy than previous experiment over this classifier.**

Keywords— precision, sentiment, classifiers, recall, opinion

I. INTRODUCTION

Internet makes people easier to connect each other. They express their opinion using internet through social media, blog post, movie review, product review site etc. Everyday huge amounts of data are generated by user. Movies are probably the best form of entertainment for mankind and it is common that people watch the movies and express their opinions either in social networking sites. By analyzing movie review data we can learn about the strong and weak point of a movie and tell us if the movie meets the expectation of the user. When a person wants to watch a movie, he first checks the review and rating of the movie. Sentiment analysis (SA) helps in obtaining the review of that movie.

SA is the process of getting valuable fact from a big set of data. It automatically classifies the people opinion as a positive or negative view. According to [1], SA techniques are Sentence, Document, Aspect and User based. The first technique detects the sentiment of each sentence as positive or negative. The second technique detects the sentiment of full document as a single unit. The third technique focuses on all the properties of an entity. Last technique conducts the social interrelation having graph theory for different parities. Machine learning (ML) and Lexicon based techniques are the most common in SA where the first uses training and testing set to categorize data. While the second approaches are used as a dictionary which contain predefined positive and negative words [2]. Here, we have used ML technique to classify data. Using document based SA we find the accuracy of different classifiers where Multinomial Naïve Bayes (NB) has better accuracy (88.50%) than others while Bernoulli NB achieves 87.50%. The organization of the paper is following: The related investigation has been discussed in section II. Section III includes the analysis procedure of SA. Section IV discusses about various types of machine learning algorithm. Section V describes the experimental results. Section VI includes the summary and future endeavor.

II. RELATED WORK

This paper evaluates the opinion on movie review data. SA is handled by natural language processing (NLP) using several levels. Various approaches are available for doing this.

In [2] authors proposed a method using word embedding to create sentiment lexicon by word vector representation. Authors of [3] discussed about the usage of SVM in SA with different source of data. In [4], authors proposed a method based on a sentiment score vector for SVM. In [5] Hu and Liu research was churning out the product features and gave product based summary. Authors of [6] worked on feedback data from global support services survey.

Sentiment analysis has many usages on movie review dataset using different techniques. Authors of [7] broadly categorized SA techniques into ML and lexical-based. Yonas Woldemariam has done SA based on ML and lexicon based approaches. He used apache hadoop framework with lexicon based model [8]. Authors of [9] discussed about entity-level sentiment analysis of issue comment. Authors of [10] discussed about the sentiment lexicon dictionary enrichment based on word2vec. They enlarge the opinion words by using SentiwordNet. In [11], authors deal the view-level SA on e-commerce data. According to [12], SA can be applied to detect the polarity of customer reviews in several dimensions.

According to [13], SA can be done based on combined techniques. ML needs training and testing data set. There are two types of ML methods, namely, supervised method (SM) and unsupervised method (UM). SM is a method in which we at first teach the machine providing some data. It [14] has several algorithms to conduct the classification technique based on the trained data which are. Naïve Bayes (NB), SVM, ME, DE etc. There is no training data in unsupervised learning where data are unlabeled. Author of [15, 16] design advanced NB algorithm with parts-of-speech tagging. He also claim that NB algorithm is costly wastes a lot of resources. Authors of [13] discussed the NB and complement NB classifier algorithm on hadoop framework. According to [17] SVM classifier was also conduct to find the public user counsel on products. NB and SVM were used [18] on various medical forum data. Authors of [19] have done the sentiment analysis on Bengali data about Bangladesh cricket team using support vector machine. In [20], they done the sentiment analysis using KNN, Bagging, COCR, NB and

978-1-7281-5241-7/19/$31.00 ©2019 IEEE

Decision tree classifier. Authors of [21] used deep learning recurrent neural network method and decision tree for movie review data.

Two kinds of lexicons are available [22]. The 1st one is corpus based lexicon and second one is dictionary based lexicon. Corpus based lexicon, like as SenticNet [23] can acquire more appropriate sentiment outcome as it is context oriented instead of accord of words oriented. Semantic oriented concept was applied based on a concept net lexicon. In [24] the authors show a statistical approach to find the sentiments. Authors of [25] show the 2 dictionaries. The 1st one is word dictionary and the second one is topic modeling. According to [14], a small set of counsel words are culled manually with acquainted orientations in dictionary based approach. This paper follows the machine learning technique for sentiment analysis.

III. Analysis Procedure

In this section, we discuss the analysis procedure. Figure 1 show the steps and techniques used in this paper for classification our text.

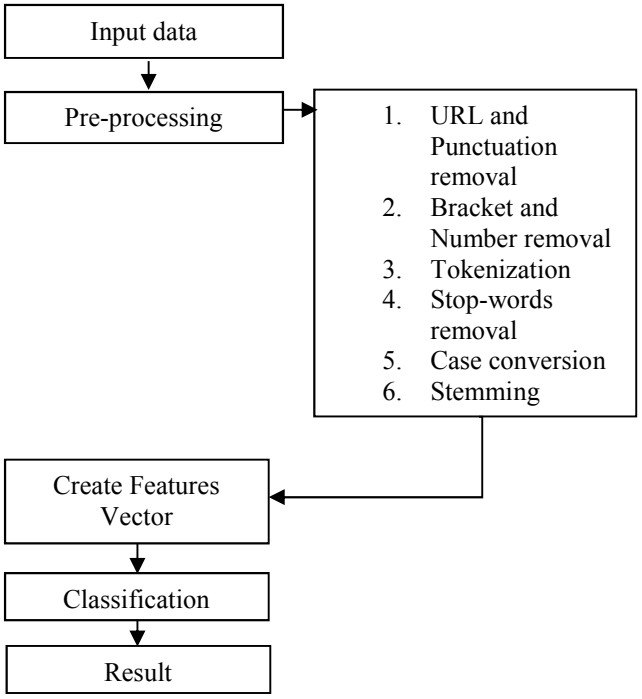

Fig. 1: Steps for classification in SA

First of all we created our dataset from the movie review data (Data source are mentioned in reference section [26]). We collected two thousands movie review posts where there exists equal number of positive and negative reviews. Then, we followed those steps for our sentiment classification.

A. Text Preprocessing

a) URL Removal: URL is a sharing link which is also known as HTML tag. Many texts contain URL or HTML tag. First of all we remove URL from the text.

b) Bracket and Number Removal: Bracket and numbers have no meaning in sentiment classification. These are treated as noise. We need to clean bracket and number from the text.

c) Tokenization: We divide our textual data into smaller components. This can be done by using tokenization. Tokenization provides us to turn text into sentences and sentences into words.

d) Omitting Punctuation: Quotation mark, semi-colon, colon, etc. are omitted as there are no data.

e) Case conversion: To remove the distinction between "Review" and "review" it is done.

f) Omitting stop words such as "I", "it", "you", "a", "an", "the" since they have no meaning in sentiment classification. We should remove those words from the input text.

g) Stemming: It is the method to reform the inflected words and removes derivational affixes from a word.

B. Feature Vector Creation

Feature is a measurable characteristic of a phenomenon. We classify each review as positive or negative. Every review is considered as a simple document. In unigram model, every document is denoted by its primary words where positive and negative primary words are used for respective positive and negative document. We also added parts of speech (PoS) tags to get the more accurate sentiment. Emotions are determined by opinion words. Positive and negative sentiments are maintained by respective sentiment scores. For this reason, these words are instrumental feature to sentiment analysis. How many features are there is termed as dimensionality. Positive and negative keywords are signed as pos and neg respectively for every document. Then, PoS are included with sentence. For this reason every word has a feature. Depending on the more variety of dimensions classifiers shows worst outcome.

IV. Considered Classifiers

A. NB Classifier

NB [27] is a SM which is the easiest and most recognizable used classifier. It considers that every feature is distinct from one another. NB classifiers are a collection of classifications algorithm which is not a standalone algorithm but a family of algorithm. The mathematical expression is as follows:

$$P(x|Y) = \frac{P(Y|x)P(x)}{P(Y)}$$

Where

$$P(Y|x) = P(y_1|x)P(y_2|x) \dots P(y_n|x)$$

Here, X is class variable and Y is a dependent feature vector. P(x) and P(x|Y) denote the respective priori and posteriori probability of x and Y.

We use two Bernoulli and Multinomial NB techniques where Multinomial NB is good for when features describe discrete frequency counts (e.g. word counts). If we want to estimate P(W | Z), the decision rule for this classifier become

$$P(W|\ Z) = \frac{count(this\ W\ in\ class) + 1}{count(all\ W\ in\ class) + count(all\ W)}$$

where W denotes sentiment. BNB is good for making predictions from binary features. Mathematically,

$$P(a_j \mid b) = P(j \mid b)a_j + (1 - P(j \mid b))(1 - a_j)$$

which differs from MNB approach.

B. SVM

SVM is another SM technique which is used to figure out each raw fact as a dot for fixed dimensions of features. Then, we choose a hyper-plane between two classes. There are several hyper-planes, but we choose one that maximizes the margin between the two classes. According to [14], text categorization is consummately appropriate for SVM due to the meager idea of content, where not many highlights are unimportant, yet they will in general be associated with each other and by and large planned into straightly particular classes.

C. ME Classifier

ME is an algorithm which uses probabilistic concept and does not guess that the characteristics are conditionally autonomous of each other. It is utilized when we cannot accept the restrictive autonomy of the features. Using the training data it makes a model which can prefer the trained data having higher entropy after a much time than other classifiers.

D. DT Classifier

DT differentiates those records having many features checking the property from the root and vertex in a tree. All terminal vertexes are assigned a class label pos or neg. The checker on property is the occurrence or non-occurrence of at least one word. Tree is partitioned until there exists least number of records.

V. Experimental Result and Analysis

A. Training and Testing Data

After creating feature vector we apply ML techniques for classification. We take 1400 and 600 movie reviews for training and testing data set where first sets are trained by a classifier and the accuracy of that is calculated using the second. We use five types of main classifiers, namely, MNB, BNB, SVM, ME and DT). All classifiers are implemented by using python. The dataset contains 2000 movie review where 1000 is negative and remaining is positive. We use the terms, namely, true positive (U), false positive (V), true negative (X), false negative (Y) for analysis. Here, first and second terms indicate that the review is really positive and negative respectively but both are featured as positive term. Third and fourth terms indicate that the review is really negative and positive respectively but both are featured as negative term. Accuracy, recall, precision, and F-score are determined from the above terms. Table 2 shows the analysis got from each classifier for data with label.

Table 1. Classifiers with four terms for the 2000 review

Method	U	V	X	Y
Multinomial NB	250	19	281	50
Bernoulli NB	259	34	266	41
SVM	268	44	256	32
Maximum Entropy	254	190	110	46
Decision Tree	247	66	234	53

Table 2. Performance statistics of several classifiers

Method	Accuracy	Precision	Recall	F-score
Multinomial NB	88.50%	92.94%	83.33%	87.87%
Bernoulli NB	87.50%	88.40%	86.33%	87.35%
SVM	87.33%	85.90%	89.33%	87.58%
Maximum Entropy	60.67%	57.21%	84.67%	68.28%
Decision Tree	80.17%	78.91%	82.33%	80.58%

Figure 2 shows the graphical representation of accuracy, precision and recall. We also show the recall versus precision graph for better comparison in Fig 3. From Fig. 2 and Table 2, we see that Multinomial NB has better accuracy compared to others. It obtains an accuracy of 88.50% while Bernoulli NB obtains 87.50%, SVM with 87.33%, Maximum Entropy with 60.67% and Decision tree with 80.17%. Multinomial NB also has high precision and F-score (Table 2), but the SVM has higher recall. Besides it also shows that Bernoulli Naïve Bayes Classifier achieves better accuracy than previous experiment over this classifier [27]. It has the good precision, recall and f-score Maximum Entropy classifier shows the low performance than other classifier. It has the low precision and f-score compare to the others. But, the precision is higher than Multinomial NB and Decision Tree. The above result shows the quality of features vector selected for movie review data.

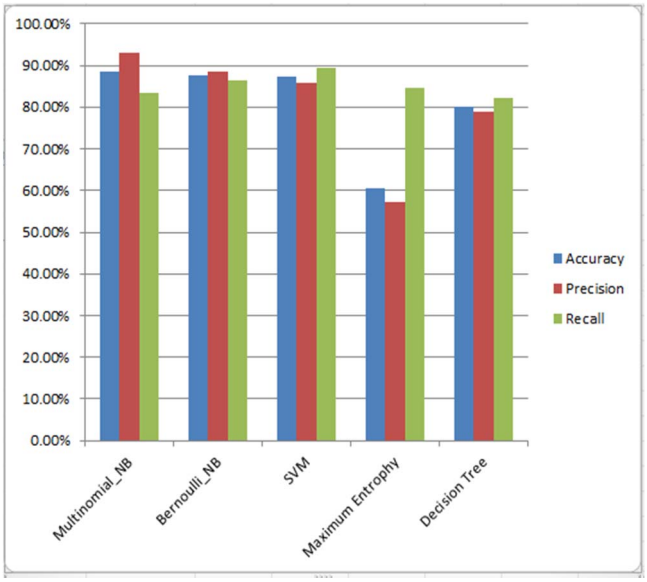

Fig. 2: Performance of difference classifiers

Every classifier is sensitive to parameter optimization. Although the result shows that Multinomial Naïve Bayes classifier is better than SVM, this is only true for selected parameters because multinomial NB show the worse results when training dataset is small.

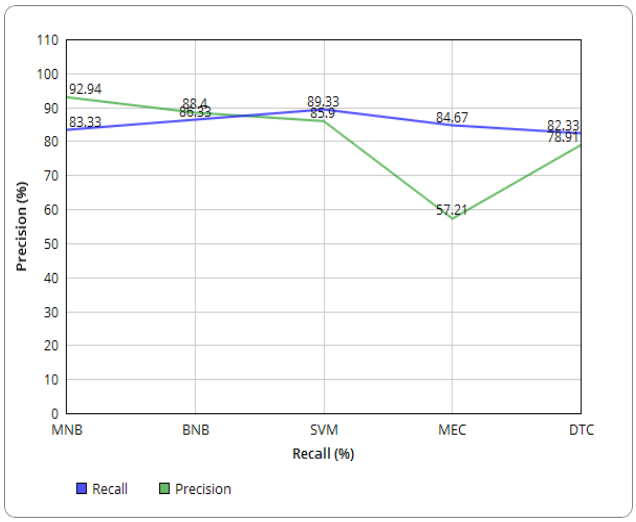

Fig.3: Classifiers for precision vs recall

VI. Conclusion and Future Plan

Sentiment analysis is very essential to understand the expression of feelings about anything like product, social media etc. It can be done by lexicon (LN) and machine learning (ML) approaches. LN can fail to calculate the score of expression if a word not found in the dictionary. While, ML is easier and more efficient but it requires labeled data. In this paper, we use ML approach for polarity classification on movie review data. This approach divides the dataset into two sets, i.e., train and test set. First of all a data set is collected from the movie review site. Next, we perform pre-processing on data by using NLP tool. Then, after creating features vector the data set is trained using ML classifiers, namely, Multinomial NB, Bernoulli NB, SVM, Maximum Entropy and Decision Tree classifiers which are tested using test dataset. Finally, we show our experimental results which present that the accuracy (88.5%) of Multinomial NB is better than others.

In near future, we would like to extend this work using deep learning approaches.

References

[1] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level sentiment analysis incorporating social network," In Proc. of ICKDDM, IEEE, pp. 1397-1405, 2011.

[2] X. Fan, X. Li, F. Du, Xin Li, Mian Wei, "Apply word vectors for sentiment analysis of APP reviews," In Proc. of ICSI, IEEE, 2016.

[3] T. Mullen, N. Collier, "Sentiment analysis using support vector machines with diverse information sources", In Proc. of ICEMNLP, pp. 412-418, 2004.

[4] S. Naz, A. Sharan, N. Malik, "Sentiment classification on twitter data using support vector machine," In Proc. of ICWI, 2018.

[5] M. Hu and B. Liu, "Mining and summarizing customer remarks," In Pro. of ICKDDM, IEEE, pp. 168–177, 2004.

[6] M. Gammon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," In Proc. of ICCL, pp. 841-847, 2004.

[7] B. Pang, L. Lee, and S. Vaithyanathan "Thumbs up?: sentiment classification using machine learning techniques," In Proc. of ICEMNLP, pp. 79-86, 2002.

[8] Y. Woldemariam, "Sentiment analysis in a cross-media analysis framework," In Proc. of ICBDA, IEEE, 2016.

[9] J. Ding, H. Sun, X. Wang, X. Liu, "Entity-level sentiment analysis of issue comments," In Proc. of IWEASE, IEEE, 2018.

[10] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha, M. Alkeshr, "Effective method for sentiment lexical dictionary enrichment based on word2vec for sentiment analysis," In Proc. of ICIRKM, Malaysia, 2018.

[11] S. Vanaja, M. Belwal, "Aspect-level sentiment analysis on e-commerce data," In Proc. of ICIRCA, 2018.

[12] P. Porntrakoon, C. Moemeng, " Thai sentiment analysis for consumer's review in multiple dimension using sentiment compensation technique.," In Proc. of ICEECTIT, 2018.

[13] B. Seref, E. Bostanci, "Sentiment analysis using naïve bayes and complement naïve bayes classifier algorithms on handoop framework," Int. Symp. on Multisciplinary Studies and Innovative Technologies, 2018.

[14] W. Medhat, A. Hassan, "Sentiment analysis algorithms and applications: Asurvey," Shams Engineering, vol. 5, pp. 1093–1113, 2014.

[15] F. Xianghua, L. Guo, G. Yanyan, and W. Zhiqiang, "Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon," Knowledge-Based Sys., vol. 37, pp. 186-195, 2013

[16] Y. Wang, "Advanced naïve bayes algorithm design with part-of-speech tagger on sentiment analysis," In Proc. of Int. Conf. on Computer System, Electronics and control, 2017.

[17] H. Cho, S. Kim, J. Lee, and J.-S. Lee, "Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews," Knowledge-Based Sys., vol. 71, pp. 61–71, 2014.

[18] T. Ali, D. Schramm, M. Sokolova, and D. Inkpen, "Can i hear you? sentiment analysis on medical forums," In Proc. of ICNLP, Asian Federation of Natural Language Processing, Nagoya, Japan, pp. 667–673, 2013.

[19] S. A. Mahtab, N. Islam, M. Rahman, "Sentiment analysis on bangladesh cricket with support vector machine," In Proc. of ICBSLP, 2018.

[20] T. P. Sahu, Sanjeev Ahuja, "Sentiment analysis of movie review: A study on feature selection & classification algorithm," In Proc. of ICMCC, 2016.

[21] A. S. Zharmagambetov, A. A. Pak, "Sentiment analysis pf a document using deep learning approch and decision trees," In Proc. of ICECC, 2015.

[22] N. El-Fishawy, A. Hamouda, G. M. Attiya, and M. Atef, "Arabic summarization in twitter social network," Ain Shams Eng., vol. 5, no. 2, pp. 411–420, 2014.

[23] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," Cognitive behavioural sys., Springer, pp. 144–157, 2012.

[24] A. Hogenboom, F. Boon, and F. Frasincar, "A statistical approach to star rating classification of sentiment," Management Int. Sys.. Springer, pp. 251–260, 2012.

[25] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexiconbased methods for sentiment analysis," Comp. linguistics, vol. 37, no. 2, pp. 267–307, 2011.

[26] Movie Review 2000 dataset, https://github.com/riyadatik/Sentiment-Analysis-on-Movie-Review-Data/blob/master/Data%20set.xlsx, 2019.

[27] V. Narayanan, I. Arora, A. Bhatia, "Fast and accurate sentiment classification using an enhanced Naïve Bayes model," Int. Data Eng. and Aut. Learning, Lec. Notes. in Com. Sci., vol. 8206, pp 194-201, 2013.