International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 08 | Aug 2018 www.irjet.net p-ISSN: 2395-0072

Classifying Twitter Data in Multiple Classes Based On Sentiment Class Labels

Richa Jain[1], Namrata Sharma[2]

[1]M.Tech Scholar, Department of CSE, Sushila Devi Bansal College of Engineering, Indore (M.P.), India
[2]Assistant Professor, Department of CSE, Sushila Devi Bansal College of Engineering, Indore (M.P.), India

---

Abstract–*The applications of sentiment analysis of text data is increases day by day. Now in these days it is used for conducting product reviews, as helping function of stock market data analysis, product reviews and feedback analysis of e-commerce. Therefore need of accurate classification of text data in their sentiment classes become more crucial. In this presented work the data mining techniques are employed for evaluating text data in multiple sub-classes of positive and negative data. In this context a new model is proposed for investigation and design. This model usages twitter dataset as experimental dataset. The initial data is preprocessed first for eliminating the stop words and special characters from text data. In next process the negation filter is applied to group the data into positive and negative classes. In this context the negation operator in given text sentences are considered. This grouped data is used with the NLP parser for finding the POS (part of speech) information from text data. This information is works here as the text features and can be used classifying the entire data samples. Based on the NLP attributes the data is transformed into unstructured to structured data source. After conversion of data SVM classifier is used for performing training and testing of trained data model. Based on different experiments the performance of implemented technique is measured and found efficient and accurate for micro-blog data analysis in terms of multiple classes.*

Keywords: NLP, Sentiment analysis, text classification, SVM classification, performance evaluation.

1.INTRODUCTION

The scope of text data mining is increases day by day because the text data is generated in a huge amount and the analysis and pattern extraction from such kind of huge and unstructured data format is a complicated task. Therefore the different data mining techniques and algorithms are applied on text data for recovering the valuable patterns. The text mining approaches are little bit different from the traditional data mining techniques because of the nature and format of data. Therefore we need a technique that process the initial information and convert it in such format by which the data is become acceptable for the learning and classification algorithms. In this presented work the text data mining is the key area of interest. Therefore the different approaches and techniques of text mining are evaluated. In addition of that the sentiment based classification for text data is proposed for design and development.

The sentiment analysis of text data is performed in order to find the author's emotions in the given text. Thus this technique is suitable for evaluation of user's product reviews, social text analysis, terror or unsolicited message analysis and spam filtering in micro-blogs. The most of such techniques basically classify the data in two major classes namely positive and negative. There are very few techniques available that works on sub-classes classification of text data. Additionally the available multiclass text classification techniques are not much accurate. Therefore the proposed technique is wishes to enhance the performance of sentiment based multi-class classification technique in terms of accuracy. Using the classification techniques first the text data is learned by some classification or data mining algorithm and then classify when the similar text patterns are appeared to classify.

2.PROPOSED WORK

This chapter provides the understanding about the proposed methodology for improved sentiment based text classification. Therefore the proposed methodology and the proposed algorithm are discussed in this chapter.

A. System Overview

The data mining and their techniques are employed on different formats of data for process the data and obtain the required information from raw data. In this context different algorithms and processes are involved for refining the information, evaluation of information and group then similar kinds of patterns. For performing these tasks the data mining techniques support the supervised and unsupervised techniques of learning. The supervised learning techniques are efficient and accurate as compared to the unsupervised learning techniques. The supervised learning techniques basically learn on the predefined patterns where the output is known to correct and adjust the errors during the learning

and in the techniques of unsupervised learning the algorithms directly employed on data to cluster the information into groups. In this work the text data is tried to classify using the supervised learning technique in addition of that the sentiment based analysis is also performed on data for finding the authors emotions in the given text blocks.

The proposed work is aimed to find an effective combination of techniques that can improve the performance of multi-class classification of text data according to their hidden sentiment patterns. In this context a new model is proposed for design and implementation. That text classification model implements the SVM (support vector machine) as the classification algorithm. Additionally the NLP (natural language processing) parser which is used to find the part of speech tagging from the text sentences are used for feature extraction. In addition of that the negation based grouping of data is also performed in initial phases for improving the classifiers performance. This section the basic overview of the proposed text classification technique is provided and in next section the detailed modeling of the proposed system is described.

B. Methodology

The proposed methodology of the proposed system is explained using figure 2.1. This diagram contains the different intermediate processes that help to process the information and extract the desired information from raw data.

Input dataset: the supervised learning techniques required the training samples by which algorithm becomes trained. Therefore some initial input samples are required for initiating the learning process. In this presented work the sentiment or emotional class labels are required to be predicting therefore the twitter micro blog data is provided as input to the system. The twitter dataset contains the user's information and the twitted data by the users.

Data preprocessing: after input of the initial training samples the data preprocessing is performed. The preprocessing of data is essential step of data mining technique. Using this filtering of information is performed by which unwanted data is removed and the target information is recovered from the input data. In addition of that the aim of data is to manipulate is such a way by which the target algorithm can learn with the processed data. Therefore transformation of data, mapping of data, and other techniques are applied. In this presented work the text data is accepted as input therefore two basic operations are involved. First is used for removing the stop words and second is used for removing the special characters from data. In order to achieve the required function the find and replace function is implemented with the list of stop words and special characters. And using the find and replace function the target data is removed from the input dataset.

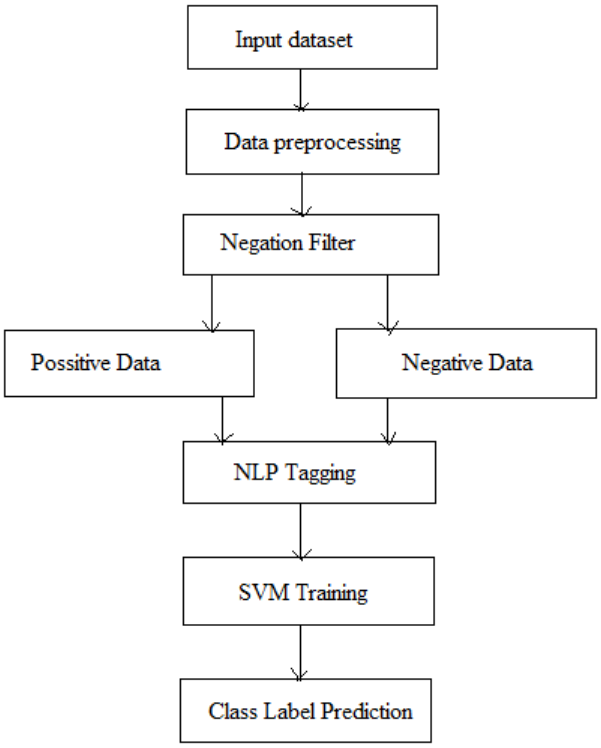

Figure 2.1 Proposed system architecture

Negation filter: after preprocessing of data the data is cleaned and now can be used for classification task. In this context a negation filter is implemented to group entire text data into two groups i.e. negative and positive data groups. The filtering of data is performed in such manner by which the sentences which consumes the negation such no, not are targeted to cluster the data. After this filter the data is grouped in two classes namely the positive data and negative data.

Positive data: the set of positive data contains those sentences that are not utilizing negations.

Negative data: the text data with the negation is clustered in this group. Both the data samples are now used with the next process for finding the part of speech information.

NLP tagging: the input data is basically the unstructured source of information and the length and other properties of each data instances are different from each other. Therefore

to make the data suitable the text feature in terms of POS (part of speech) is used for classifiers training and testing. The Stanford NLP (natural language processing) parser is employed in this phase of data processing. The NLP parser first maps all the data into the part of speech information. And the implemented function counts the POS information. Using the counted information from the input data is transformed in a two dimension vector. The example of parsed data is given using table 2.1.

Noun	Pronoun	Verb	Adverb ….
1	2	1	0

Table 2.1 example of POS tagging

SVM training: the SVM (support vector machine) is a supervised learning algorithm. This algorithm is basically used for classifying two classes. This classifier basically works on distance based technique, in this context the entire data is first scattered into a N dimensional space where the dimensions of space are equivalent to the number of attributes in the given dataset. Additionally the dataset instances are treated as a single point in space. In order to classify the data instances the hyper plan is plotted in optimal distance by which the data become separable. In this presented work the primarily positive and negative data groups are classified in again sub classes therefore the SMO (Sequential minimal optimization) technique is utilized for classifying the sub-classes of data.

Class label prediction: after training of the SVM classifier the test data samples are produced to the classifier. The test set is prepared on the basis of input training sample and 30% randomly selected data instances are selected for preparing the test dataset.

C. Proposed Algorithm

This section summarizes the processes involved in the methodology section in form of algorithm steps. Table 2.2 contains the processes involved in the proposed methodology.

Input: training dataset D

Output : class labels of text data C

Processes:

1. $R_n = readDataset(D)$
2. $for(i = 1; i \leq n; i++)$
   a. $P_i = removeStopWords(R_i)$
   b. $P_i = removeSpecialCharacters(P_i)$
3. $End\ for$
4. $[Positive_p, Negative_q] = negationFiletr(P_n)$
5. $POS_n = NLP.TagData([Positive_p, Negative_q])$
6. $T_{model} = SVM.Train(POS_n)$
7. $C = T_{model}.classify(TestData)$
8. Return C

Table 2.2 proposed algorithm

3.RESULT ANALYSIS

This chapter provides the results analysis in the form of measured performance parameters. The different computed parameters and relevant outcomes are reported in this chapter.

A. Accuracy

Accuracy is the measurement of the data mining system performance in terms of how accurately a data mining algorithm performs the classification. That is computed by the ratio of total correctly classified data and total samples to be classified. The following formula can be used for computing classifier's accuracy.

$$accuracy = \frac{total\ correctly\ classified\ data}{total\ samples\ to\ classify} X100$$

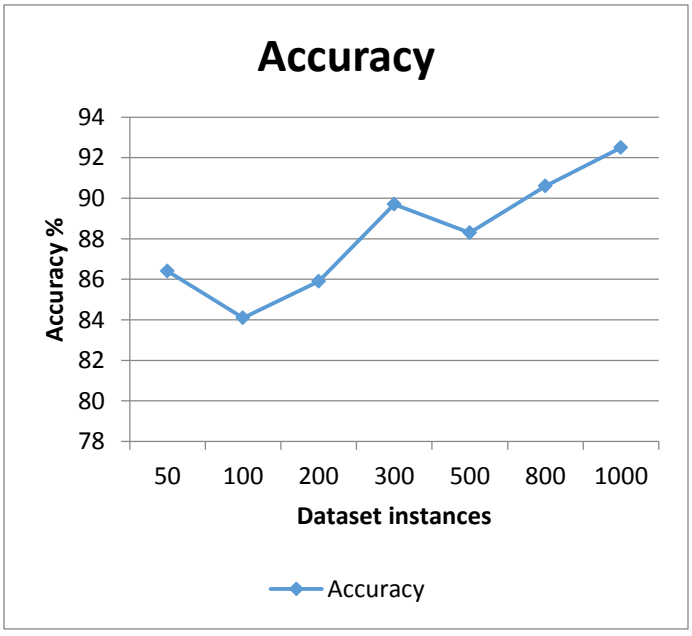

Figure 3.1 Accuracy in percentage

Dataset Size	Accuracy
50	86.4
100	84.1
200	85.9
300	89.7
500	88.3
800	90.6
1000	92.5

Table 3.1 accuracy in percentage

Accuracy of the proposed sentiment based text classification technique is evaluated and reported using table 3.1 and figure 3.1. In this diagram the X axis contains the size of dataset in terms of number of instances additionally the Y axis contains the corresponding accuracy of the proposed approach in terms of percentage. According to the obtained performance of classification system the proposed technique improves their accuracy as the amount of data instances for training and testing is increases. Therefore the proposed system is acceptable for real world application usages.

B. Error Rate

Error rate is also an essential parameter of classifier performance using this parameter it is measured how much amount a classifier misclassify the data. That is the ratio of data misclassified and total samples provided for classification to algorithm. The error rate of a classifier is evaluated using the following formula:

$$error\ rate = \frac{misclassified\ data}{total\ data\ to\ classify} X100$$

Or

$$error\ rate\ \% = 100 - accuracy$$

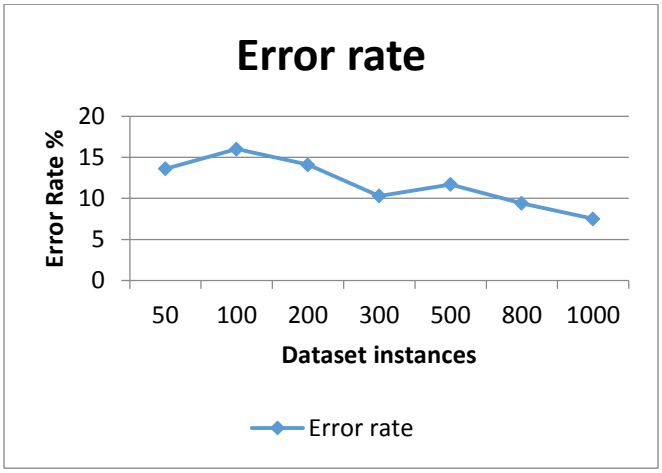

Figure 3.2 error rate percentage

Dataset Size	Error rate
50	13.6
100	16
200	14.1
300	10.3
500	11.7
800	9.4
1000	7.5

Table 3.2 error rate percentage

The computed error rate of the proposed text classification data model is demonstrated using table 3.2 and figure 3.2. In this diagram the X axis shows the amount of data supplied for classification in terms of number of instances of data objects and the Y axis of the diagram shows the error rate percentage. According to the obtained performance the proposed technique minimizes the amount of error percentage as the amount of data size is increases. Therefore the proposed approach is acceptable for real world application usages.

C. Memory Usages

Memory usages of a process are also known as the space complexity of algorithm. A data mining algorithm for classification how much amount of main memory consume is termed here as the memory usages of the process. The java based process how much memory usages can be estimate using the following formula:

$$memory\ usages = total\ memory - free\ memory$$

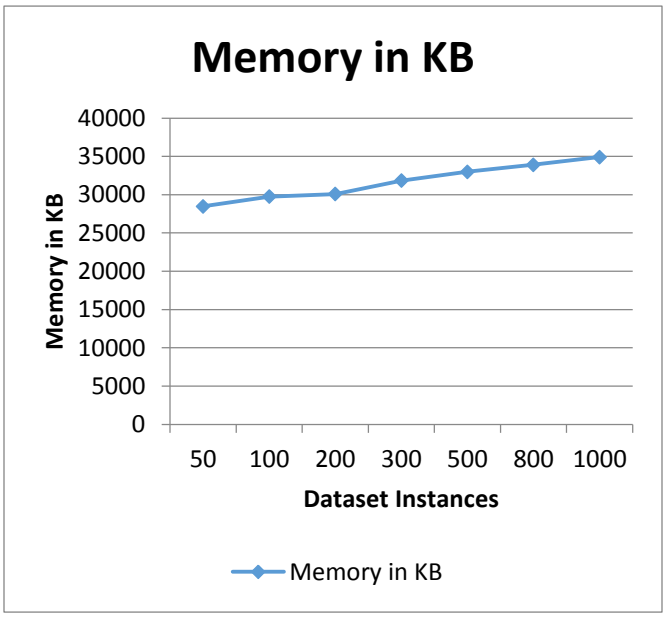

Figure 3.3 memory consumption in KB

The memory consumption of the proposed classification technique of text is demonstrated using figure 3.3 and table 3.3. According to the given graph the X axis contains the amount of data set instances to be classified and the Y axis represents the computed memory usages of the given process. The memory consumption of the target algorithm is computed here in terms of KB (kilobytes). According to the obtained results the memory consumption of the system is increases as the amount of data instances for classification is increases. But the increment in memory is not much higher therefore it is acceptable in cost of their classification accuracy. Therefore the memory consumption is directly depends upon the size of data to be classify.

Dataset Size	Memory in KB
50	28474
100	29747
200	30084
300	31837
500	32994
800	33913
1000	34927

Table 3.3 memory consumption in KB

D. Time Consumption

Time consumption is the measurement of time required to classify the input data. In this given system the time consumption is computed using the following formula:

$$time\ consumption = end\ time - start\ time$$

Dataset Size	Time in MS
50	36
100	43
200	57
300	74
500	95
800	120
1000	146

Table 3.4 time consumption in MS

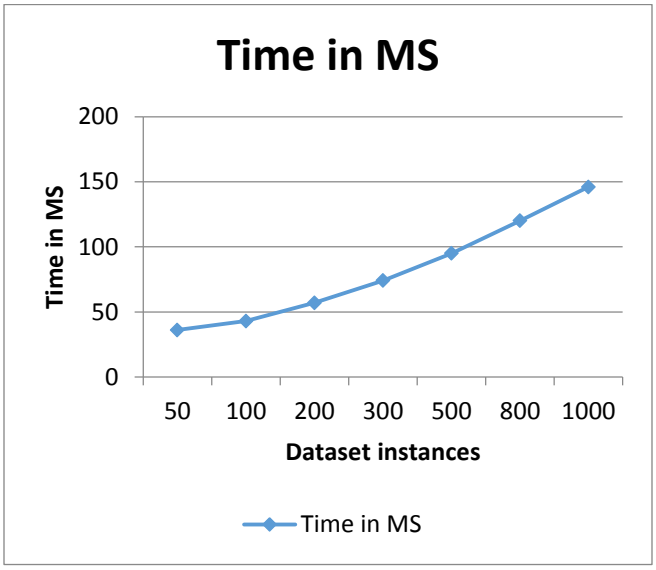

Figure 3.4 time consumption in MS

The time consumption of the proposed classification technique is demonstrated using figure 3.4 and table 3.4. The time is measured here in terms of milliseconds (MS). The diagram shows in X axis the amount of dataset size input to classification and the Y axis shows the time consumed for performing the classification. According to the given graph the time consumption of the system is increases as the amount of data size is increases. Additionally that is also observed the amount of time is depends on the amount of data supplied for classification.

4 .CONCLUSION

This chapter provides the conclusion of the proposed sentiment analysis based text classification. The experimentations and observations are used for providing the conclusion of the proposed work in addition of that the future extension of the proposed work is also discussed in this chapter.

A. Conclusion

The trend of text classification approaches are increases day by day. As the amount of text data requirements is increases we need some efficient and accurate data mining techniques that help for data analysis. In addition of that the applications of text classification techniques are also increases in different purpose sometimes it is used for making groups of text data and sometimes it is required to find the emotions from data. The classification of text data is complicated as compared to structured data source because the data in text format is not regular in size and amount. Therefore to find suitable attributes and features from data additional efforts are required to implement.

The proposed work is tried to design and develop an enhanced method by which the text is classified in sentiment based classes. Therefore first the data is preprocessed and unwanted data is removed from the initial data. In next the data is filtered on the basis of negation operators available in text sentences. After applying the negation filter the data is clustered into positive and negative data groups. In further the data is transformed into the structured information in this context NLP features of text data is used. In order to identify the NLP features the POS tagging on data is performed and based on part of speech information the data is transformed into two dimensional vector. This 2D vector is used with the SVM classifier for training and testing of implemented method.

The implementation of the proposed approach is performed using JAVA technology and NetBean IDE. After implementation of the proposed technique the performance of the proposed classification system is computed and reported. The table 4.1 contains the summary of the conducted experiment based results analysis.

S. No.	Parameters	Remark
1	Accuracy	Accuracy of the proposed approach is also improved with size of training data additionally after sometimes it remains consistent between 90-95% of classification rate
2	Error rate	The error rate is reduces with the amount of training samples increases but sometimes it fluctuating because of the noise composition of data
3	Memory usages	The memory usages of system are not varying with the amount of data but still it depends on the size of input data.
4	Time complexity	The time complexity of the proposed approach is acceptable and it is depends on the size or number of instances of data produced for classification

Table 4.1 performance summary

According to the obtained performance of the proposed multi-class sentiment analysis technique the proposed technique found the optimal performance and efficient for computing the class labels of text data. Thus the proposed technique is acceptable for the data classification and utilization with the real world applications.

B. Future Work

The main aim of the proposed work is to find an optimum classification technique to improve the performance of multiclass sentiment classification. In this context a new data model is proposed for implementation and the performance is improved. In near future the following work is extended using the following concepts.

1. The proposed technique implemented with the traditional SVM classification technique the performance can more optimized with the help of ensemble based learn approach in near future the ensemble learning is proposed for implementation
2. The proposed technique just consider the negation feature for filtering the data in two groups in near future more characteristics are recovered for optimizing the negation filter

REFERENCES

[1] MondherBouazizi and TomoakiOhtsuki, “A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter”, 2169-3536, 2017 IEEE, VOLUME 5, 2017

[2] Han, Jiawei, Jian Pei, and MichelineKamber, "Data mining: concepts and techniques", Elsevier, 2011.

[3] Bharati M. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, Volume 1 Number 4, pp. 301-305.

[4] Dunham, M.H. Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, New Delhi, 2003. Print. ISBN: 81-7758-785-4, 2006.

[5] Ian H. Witten; Eibe Frank; Mark A. Hall, —Data Mining: Practical Machine Learning Tools and Techniques (3rd Ed.), Elsevier, 30 January 2011.

[6] Kantardzic and Mehmed, "Preparing the Data." Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition (2003): 26-52.

[7] Sebastiani, F., "Machine learning in automated text categorization." ACM Computing Surveys, Vol. 34, No. 1, pp. 1–47, 2002.

[8] Sonali Vijay Gaikwad and ArchanaChaugule "Text Mining Methods and Techniques", International Journal of Computer Applications (IJCA), Volume 85 – No 17, January 2014.

[9] Ian H. Witten, "Text mining", 2004, available online at: https://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

[10] Miss LatikaKaushik, "Text Mining - Scope and Applications", Journal of Computer Science and Applications, Volume 5, Number 2 (2013), pp. 51-55

[11] RamzanTalib and Muhammad KashifHanif, "Text Mining: Techniques, Applications and Issues", International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016

[12] A.H. Tan, Text Mining: The State of the Art and the Challenges, in PAKDD99 Whorkshop on Knowledge Discovery from advanced Databases, Beijing, China, and April 1999.c.