IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Bayesian Classification for Social Media Text

Amit Kumar Mittal[*1], Shivangi Mittal[2], Digendra Singh Rathore[3]

[1.] Department of Computer Engineering Institute Of Engineering & Technology, Devi Ahilya University, Indore, India.
[3]Department of Electrical Engineering Govt. Polytechnic College, Dewas, India
[3.] Department of Computer Engineering Institute Of Engineering & Technology, Devi Ahilya University, Indore, India

Corresponding Author: amittal@ietdavv.edu.in,+91 9993315549

Available online at: www.ijcseonline.org

Accepted: 13/Jul/2018, Published: 31/July/2018

Abstract— The data mining is a technique by which the computational algorithms are trained for finding the similar patterns from the huge or raw set of data. The training of the algorithms is performed on the patterns which are required to extract from the data. The training of the algorithms can be supervised or unsupervised. The main advantage of the supervised learning algorithms, these are efficient, accurate and effective as compared to the unsupervised learning approaches. In this presented work the text classification is the key area of study. The text classification techniques are used to classify according to their categories or the domain specific knowledge. Thus the text classification has rich applications. Among a number of applications of the text classification the social media based text classification and the sentiment analysis of the user's text is comparatively new work in the text mining. In this presented work the social media based text is mined for discovering the user sentiments or moods which are expressed using the twitter based text communication. Therefore big data analytics are used to performing the text classification. First the twitter data is hosted on the HDFS directory and then the features are computed using the Map-reduce technique. The collected features are then labelled using the NLP tool which is used to discover the part of speech composition of the text sentences. After parsing the text using NLP tool the Bayesian classifier is implemented for classification of the social media text. The implementation of the proposed technique is performed using the JAVA technology. After implementation the performance of the proposed system is evaluated in terms of accuracy and the complexity. Both the performance parameters show the proposed sentiment analysis technique is effective and accurate for classifying the social media text for orientation discovery of user text.

Keywords— classification; sentiment analysis; supervised learning text orientation; text mining.

I. INTRODUCTION

The sentiment analysis and emotion based text classification is becomes more and more popular in the data mining domain. The key reason behind this the number of business applications need to know about the orientations of the users and the user's mood, interest and other personalized view of user. Therefore in different recommendation engines the user behaviour and the user orientation and interest are evaluated. Apart from these applications the key issue in sentiment based text analysis is to understand the hidden feeling in the text documents of the users. Therefore in various stress management systems the social media data is used to analyse the emotions of the user. This analysis need to implement some computer based algorithms which understand the patterns of the text data and classify according to their emotional patterns.

The classification of data is a kind of supervised learning which needs training from the similar patterns of data. Additionally after training these algorithms are able to distinguish the similar patterns on which the algorithms are takes training. In this presented work the twitter dataset is used for analysis and sentiments based text analysis. Now in these days that is much popular application among the youngsters. The students, politicians, business owners and others frequently usages the twitter and react on the current trending patterns. These reactions on the social media represent the nature, mood and the emotions of the user for the specific post or article.

In this presented work the twitter dataset is used for analysis and sentiments based text analysis. Now in these days that is much popular application among the youngsters. The students, politicians, business owners and others frequently usages the twitter and react on the current trending patterns. These reactions on the social media represent the nature, mood and the emotions of the user for the specific post or article.

Section I contains the introduction of Bayesian classification for social media text, Section II contain the proposed system for the sentiment text analysis, Section III contain the results and their detailed understanding, and Section IV summary of entire work and future directions.

II. PROPOSED WORK

The proposed system for the sentiment text analysis and their accurate evaluation a new system is prepared using the traditionally available techniques. The organization of methodologies for obtaining sentiment based text analysis is given using figure 1.

Twitter dataset: The machine learning techniques required to phase of processes involved in the classification techniques first the learning through the previous examples and then utilizing the previous examples for classifying the data of the similar pattern. The given twitter data is used for the learning in the proposed technique of classification.

HDFS: That is the big data repository which is used to store the data for learning and classification. Thus that is just storage architecture of data for processing in big data environment.

MAPREDUCE: That is a data processing tool used with the Hadoop infrastructure to process the data and reduce the amount of data from the actual amount of data using the mapping techniques. The training data placed in Hadoop directory is processed using the map-reduce and produced in next phase for storing them in to column arrangement.

HIVE: Hive provides the data storage structure in the column manner thus the different number of attributes which are participating in classification is organized for the text data is used for further processing in the table manner.

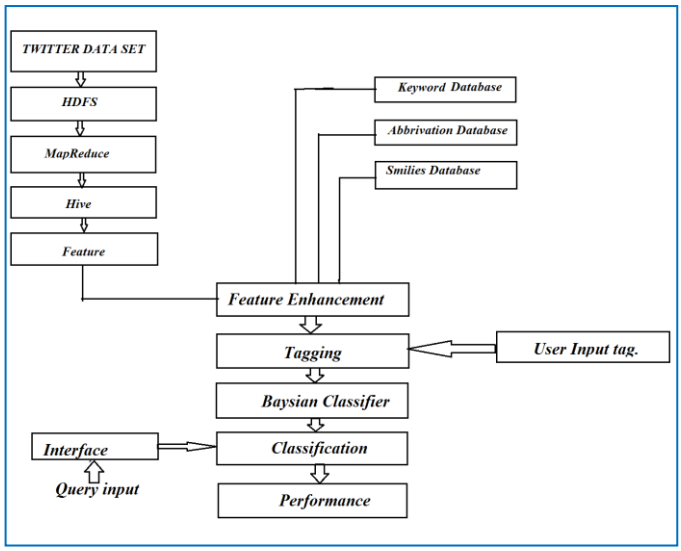

Figure 1 proposed methodology

Features: In this phase the column based data is processed to find the word occurrence frequency for the given or specified classes in the sentiments. Therefore the term frequency is provided by the following formula:

$$word\ frequency = \frac{number\ of\ time\ occured\ a\ word}{total\ amount\ of\ word}$$

Feature enhancement: That is a feature enhancement technique or the data quality improvement technique by which the incomplete words and the different text symbols are recovered from three different data bases. These three different data are as follows:

- **Keyword database:** Keyword database contains the significant amount frequent words that are frequently occurred in any kind of text sentence organization such as, is, am, are, this, that, to and others. Using the feature enhancement technique the words are compared to the database and removed from the available features.
- **Abbreviation database:** In most of social network sites the number of abbreviations is used during the text communication such as for take care the people usage the terms TC. Thus a database is prepared with the Abbreviation and their full forms to reform again the entire aspect of the text data.
- **Similes database:** During text communication for expressing the moods and the emotions sometimes user are also usages the graphics or special characters. These special characters and graphics in the social network are known as the similes. Thus a database with the similes and their meaning is also prepared to enhance the features by completing the sentences.

Tagging: In this place the user interaction with the feature list is required to supply the initial tags these tags are in terms of noun, pro-noun and other semantics of the text data.

Bays classifier: In this phase the tagged data from the previous phase is utilized to develop the statistical classifier. Which first prepare the training from the training data set and then the test set is used to provide the classification of the data according to the hidden sentiments.

Interface: The provision is made in order to simulate the training and test of the proposed sentiments based text classification model. That is a basic user interface design which is used to submit the training and testing dataset for classification and performance measurement.

Query input: That is a part of testing phase which is used to produce the unknown text for finding the orientation of the text communication without any class labels and that the responsibility of the data model to analyse the text data and provide the class labels for the unknown text according to the tagged data.

Performance: After the classification task the performance of the classification system is evaluated in terms of their accuracy, error rate and the other efficiency parameters.

III. RESULTS ANALYSIS

After implementation of the proposed text classification system are performed. This chapter provides the detailed discussion about the preformed experiments and their computed results in terms of different performance parameters. The obtained results and their detailed understanding are given in this chapter.

A. Accuracy

In a data mining based classification system the amount of correctly recognized patterns are known as the classification accuracy. The accuracy of the system in terms of percentage can be computed using the following formula.

$$accuracy = \frac{accurately\ classified\ patterns}{total\ input\ patterns} X100$$

The accuracy of the implemented algorithm is represented using table 1 and figure 2. The given graph figure 2 contains the accuracy of the implemented algorithm. The X axis of the diagram contains the amount of data during the training and testing and Y axis contains the obtained performance in terms of accuracy percentage.

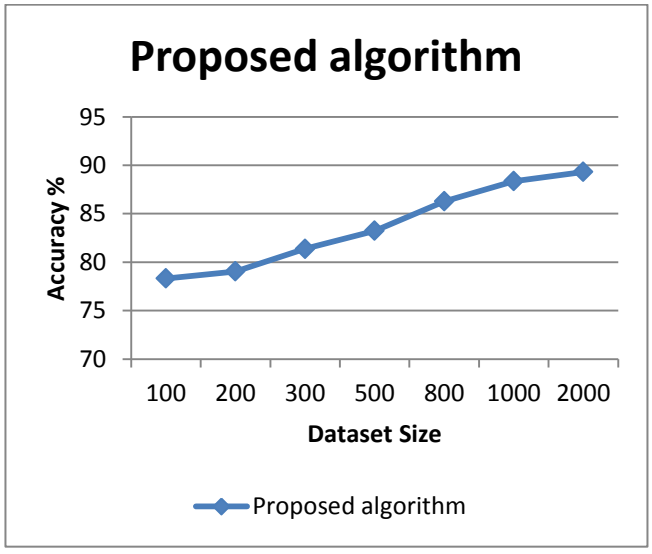

Figure 2 accuracy

According to the obtained results the performance of the proposed classification technique provides more accurate results. Additionally the accuracy of the learning model is increases as the amount of instances for the learning of algorithm is increases.

Data size	Proposed algorithm
100	78.32
200	79.05
300	81.37
500	83.24
800	86.28
1000	88.35
2000	89.32

Table 1 accuracy

B. Error rate

The amount of data misclassified samples during classification of algorithms is known as error rate of the system. That can also be computed using the following formula.

$$error\ rate\ \% = \frac{total\ misclassified\ patterns}{total\ input\ patterns} X100$$

Or

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

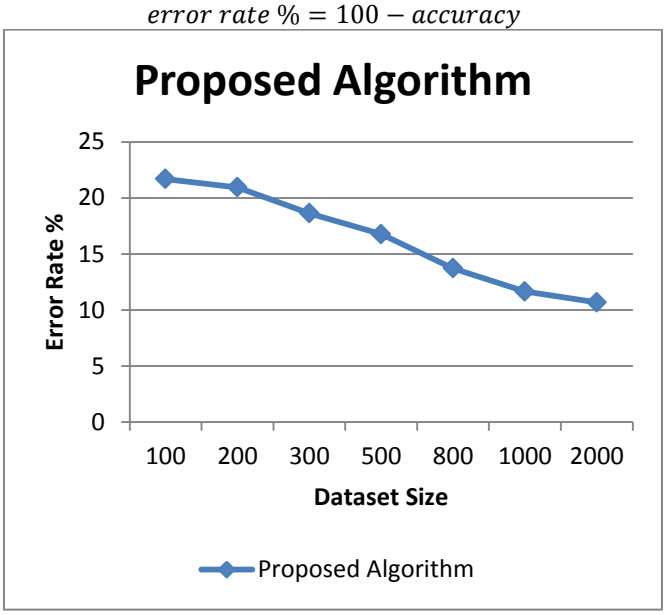

Figure 3 error rate

Dataset size	Proposed algorithm
100	21.68
200	20.95
300	18.63
500	16.76
800	13.72
1000	11.65
2000	10.68

Table 2 error rate

The figure 3 and table 2 shows the comparative error rate of implemented algorithm. In order to show the performance of the system the X axis contains the amount of data used for training and the Y axis shows the performance in terms of error rate percentage. The performance of the proposed classification is effective and efficient during different experimentations and reducing with the amount of data increases. Thus the presented classifier is more efficient and accurate than the traditional approaches of text classification.

C. Memory usage

Memory consumption of the system also termed as the space complexity in terms of algorithm performance. That can be calculated using the following formula:

$$memory\ consumption = total\ memory - free\ memory$$

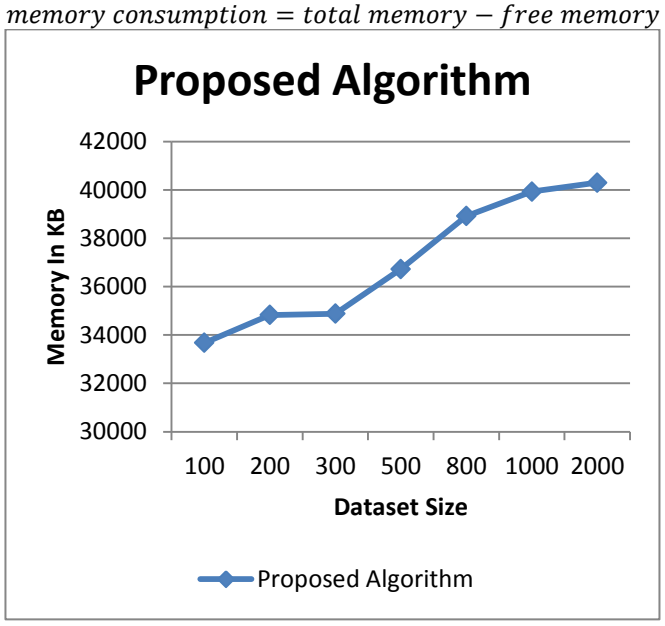

Figure 4 memory consumption

The amount of memory consumption depends on the amount of data reside in the main memory, therefore that affect the computational cost of an algorithm execution. The performance of the implemented classifier for sentiment classification is given using figure 4 and table 3. For reporting the performance the X axis of figure contains the amount of data required to execute using the algorithms and the Y axis shows the respective memory consumption during experimentations in terms of kilobytes (KB). According to the obtained results the performance of algorithm demonstrates similar behaviour with increasing size of data, but the amount of memory consumption is increases with the amount of data.

Dataset size	Proposed algorithm
100	33677
200	34827
300	34882
500	36728
800	38918
1000	39928
2000	40299

Table 3 memory consumption

D. Time consumption

The amount of time required to classify the entire test data is known as the time consumption. That can be computed using the following formula:

$$time\ consumed = end\ time - start\ time$$

The time consumption of the proposed algorithm is given using figure 5 and table 4. In this diagram the X axis contains the size of dataset and the Y axis contains time consumed in terms of milliseconds. According to the comparative results analysis the performance of the proposed technique shows the less time consumption. But the amount of time is increases in similar manner as the amount of data for analysis is increases.

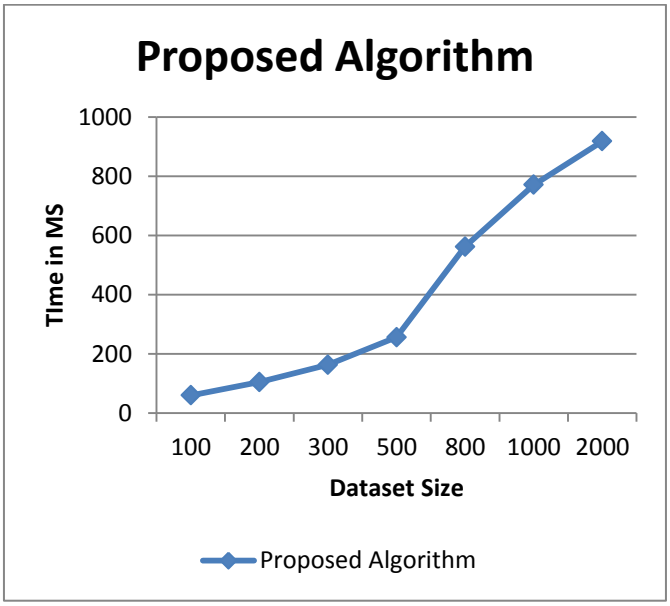

Figure 5 time consumption

Dataset size	Proposed algorithm
100	60
200	105
300	163
500	256
800	562
1000	771
2000	918

Table 4 time consumption

IV. CONCLUSIONS

The proposed work is intended to find the data mining approach which is used to classify the text data according to their sentiments. Therefore the proposed study is focused on analysing the text mining techniques and classification algorithms. The chapter provides summary of entire work performed additionally future extensions are also suggested.

A. Conclusion

Data mining technique supports both the kinds of learning methods supervised and unsupervised to analyse data. The learning algorithms are help to classify or categorize the data in groups on which the patterns are belongs (known as class labels). The algorithms are computer based programs that used to analyse data without human efforts by using the past experience of the similar pattern data. Now in these days the popularity of social media applications is growing. In these platform the significant amount of users are expressing their emotions using the text. Sometimes this data is in quantity are very large thus the big data can be used for proper evaluation and processing. Most of the social media networks are consuming the services of big data.

The proposed work analyse social network based text more specifically the twitter based text for sentiments analysis and the text orientation discovery. Therefore to analyse the data in more effective manner the proposed work involves the big data environment for data storage, and pre-processing. For storage the HDFS file system is used and to pre-process and the feature extraction the MapReduce is used. The pre-processed selected features form text is then stored in the hive data structure and their tagging using the NLP tool is performed. The NLP tagged data is now used to make training of the Bayesian classifier. During the training of the classifier the probability distribution for both the classes are computed and then the trained model is used for data classification. To classify the data the real world twits can be used with the similar tagging concepts.

The implementation of proposed big data based sentiment classification system is performed using JAVA technology and their performance analysis is performed using the experimental data analysis basis. The performance of system is estimated for finding system accuracy and error rate in prediction. Additionally for performance in terms of time and space complexity is also evaluated to provide the efficiency. The performance summary is given using table 5.

S. No.	Parameters	Remark
1	Accuracy	The accuracy of the learning model is depends on the quality and quantity of data. As the learning data is increases for both orientation the accuracy of the system is increases in similar manner
2	Error rate	The proposed system is adoptive and preserves the past learning experience in the data storage. Thus increasing learning patterns reduces the error rate of the system
3	Memory	Memory consumption is proportional to the amount of data to be processed but the memory consumption is demonstrate the similar behaviour not much fluctuating
4	Time	Less time consumption of the classification system is observed. That is not much fluctuating with the respective amount of data

Table 5 performance summary

According to the obtained results the system is efficient and accurate for classifying the text according to sentiments orientations. Thus proposed model for text classification is adoptable and efficient.

B. Future work

The proposed work is accurate and efficient for classifying the patterns of text data. The technique is successfully able to distinguish the sentiments hidden in text. In near future that is promising approach for providing accurate classification. The following extension may be feasible for future work.

a. Adopt more literature for improve the technique for dynamic or stream based big data environment.
b. Need to improve the real time data opinion of the user directly implementable with any social media web application.
c. The approach is also extendable for multiclass classification of the emotions with different other kinds of applications such as stock market price prediction and other.

REFERENCES

[1] Yue Gao, Fanglin Wang, Huanbo Luan, Tat-Seng Chua, "Brand Data Gathering From Live Social Media Streams", ICMR'14, April 01-04, 2014, Glasgow, United Kingdom. Copyright 2014 ACM 978-1-4503-2782-4/14/04

[2] A Comparison of Several Approaches to Missing Attribute Values in Data Mining, Jerzy W. Grzymala-Busse and Ming Hu, Springer-Verlag Berlin Heidelberg 2001, pp. 378−385,

[3] Ritika, "Research on Data Mining Classification", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014

[4] Abbas Jafari, S.S.Patil, "Use of Data Mining Technique To Design A Driver Assistance System", Proceedings of 7th IRF International Conference, 27th April-2014, Pune, India, ISBN: 978-93-84209-09-4

[5] A.K. Jain, M.N. Murthy, P. J. Flynn, "Data Clustering: A Review", © 2000 ACM 0360-0300/99/0900–0001

[6] Khaled Hammouda ,"A Comparative Study of Data Clustering Techniques", Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

[7] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek, "Density-based Clustering", WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30

[8] B. V. Rama Krishna, B. Sushma, "Novel Approach to Museums Development & Emergence of Text Mining", ISSN 2249-6343, International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 2, Issue 2

[9] H. P. Luhn, "A Business Intelligence System", Volume 2, Number 4, Page 314 (1958), Nontopical Issue, IBM Research Journals
[10] Andreas Hotho, Andreas Nurnberger, Gerhard Paaß, Fraunhofer AiS, "A Brief Survey of Text Mining", Knowledge Discovery Group Sankt Augustin, May 13, 2005s
[11] Hien Nguyen, Eugene Santos, and Jacob Russell, "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, Vol. 41, No. 6, November 2011
[12] Umajancy. S, Dr. Antony Selvadoss Thanamani, "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013
[13] Miloš Radovanović, Mirjana Ivanović, "Text Mining: Approaches And Applications", Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 2008, 227-234
[14] Siva S. Sivatha Sindhu, S. Geetha, A. Kannan, "Decision tree based light weight intrusion detection using a wrapper approach", Expert Systems with Applications, 2011 Elsevier Ltd. All rights reserved.
[15] M. Jayakameswaraiah and S. Ramakrishna, "Implementation of an Improved ID3 Decision Tree Algorithm in Data Mining System", International Journal of Computer Science and Engineering, Volume-2, Issue-3
[16] Biswajeet Pradhan, "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS", Computers & Geosciences, & 2012 Elsevier Ltd. All rights reserved.
[17] Neha Patel, Divakar Singh, "An Algorithm to Construct Decision Tree for Machine Learning based on Similarity Factor", International Journal of Computer Applications (0975 – 8887) Volume 111 – No 10, February 2015
[18] T. Ramani, M. Ramzan Begam, "Survey: A Techniques implemented on Opinion Mining", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 5 No. 10 Oct 2014
[19] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, (2014) 5, 1093–1113
[20] Tirivangani BHT Magadza, Addlight Mukwazvure, K.P Supreethi, "Exploring Sentiment Classification Techniques in News Articles", IJITKM Volume 8 • Number 1 June-Dec 2014 pp. 55-58 (ISSN 0973-4414)
[21] Alya Al Nasseri, Allan Tucker, Sergio de Cesare, "Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms", © 2015 The Authors. Published by Elsevier Ltd.
[22] Rodrigo Moraes, João Francisco Valiati, Wilson P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN", Expert Systems with Applications, 2012 Elsevier Ltd. All rights reserved.
[23] Suge Wang, Deyu Li, Lidong Zhao, Jiahao Zhang, "Sample cutting method for imbalanced text sentiment classification based on BRC", Knowledge-Based Systems, 2012 Elsevier B.V. All rights reserved..

