International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Mood Prediction On Tweets Using Classification Algorithm

Sameeksha Shrivastava[1], Dr. Pramod S. Nair[2]

[1, 2]Department of Computer Science & Engineering, Indore, RGPV

Abstract: ***Data mining is a technique which offers the computer algorithm to compute patterns and find the category of data using classification and clustering. In data mining classification is performed with supervised learning and unsupervised learning. Selection of algorithm depends upon the type and behavior of data. The data can be as structured and unstructured. Structured data is that which reside in fixed field. It is first depends on creating data model. Unstructured data refers to information that does not have a predefined data model or not organized in a predefined manner. In data mining text mining has become an important research area. Text mining is a discovery of new, previously unknown information by automatically extracting information from different resources [5]. The various applications in text mining are information retrieval, machine learning, data mining, and statics and computation semantics. In form of text data most of the information is stored. Now a days in a direction of multiple language support most of the research is progressing. This system is capable to group the similar data from different kinds of language source according to their original semantic and also being able to gain information across language [2]. In the presented work the identified twitter data set is used to perform text analysis. Therefore the entire input data samples are required to classify in two classes namely positive and negative. Therefore a binary classifier namely ID3 decision tree and their improved variant is utilized for analysis and performing the classification task. Before classification of text data there is need to improve the quality of data. Therefore the raw text data is first pre-processed then tagged according to the lexical means. After tagging on the original text data the classification algorithms are trained and make use to classify the text according to their sentiments. The implementation of the improved ID3 text classification technique and their performance is evaluated in terms of their accuracy and the error rate. These parameters show how accurately the text patterns are identified using the data mining technique. Additionally for finding their performance in terms of their efficiency the time and space complexity is also measured that shows the effective classification with less consumption.***

Keywords: **T**ext analysis, classification, text sentiments, tweeter data, micro-blog

1. Introduction

Microblogging are gigantic store of user generated content about world events. The quantity of Microblog posts everyday have become increasing to a level that hampers the viable recovery of relevant message, and amount of information is conveyed from microblog is increasing rapidly. These websites have evolved to become a source of varied kind of information .This is due to nature of Microblog on which people post real time message about their opinions on variety of topics, discussing current issues complain and express their sentiments on daily events.

Data mining is a technique of mining information from the raw data. Here the information is a term that is relevant to the data which is required by a data miner or application. Text mining deals with the computational analysis of text for knowledge discovery and data pattern analysis. These techniques provides ease in information extraction, natural language processing, and information retrieval. Additionally, these techniques include domains with algorithms and KDD methodologies [3]. There are some applications of text mining such as Enhancing Web Search , Mining Bibliographic Data, Sentiment Classification [4].In this presented work the text data is mined for semantically analyse the text from raw set of data. Initially text data is found in an unstructured manner and labelling of data is complicated task therefore most of the application are utilizing the cluster analysis techniques for categorizing data. But if the data is well labelled then that can be used with the classification algorithms also. Therefore the microblog data can be used with the classification.

In this age of technology most of the computational algorithms and applications are hosted on remote servers. Users consume the remote data using a global information network known as internet. This network provides services and information 24X7 therefore that becomes a part of new generation life. Use of internet also connects us with the imaginary social world such as twitter, Facebook and others. In these social networking web applications a huge amount of text, image and video data is generated. The manual analysis of huge amount of data is a challenging task, therefore computational or statistical techniques are applied on these data to find the targeted patterns for this data.

In this work the text data for microblog analysis is used for preparing the classification algorithm. Basically the microblogs are frequently used now a days. But frequent use of this communication increases the amount of data for manual analysis. The presented model is a text analysis technique which provides the outcome in two steps and works on labelled data. In first the data is processed in order to obtain the text features and then the learning on evaluated features are performed. In order to identify the pattern of data of social networking sites more specific semantic analysis technique is required. On the other hand for accurate classification of these data some traditional data mining technique is developed which provide ease in classifying text data. Therefore the proposed work involve the improved classification algorithm for classifying the text. This classification helps us to find the moods of a user during communication over the social networking web applications.

2. Previous Work

Xia hu [7] published a paper "Exploiting Social Relation for Sentiment Analysis in Microblogging" in 2013, he investigated whether convivial cognation can avail sentiment analysis by proposing a sociological approach to handling noisy and short text for sentiment classification. In particular they present a mathematical optimization formulation that incorporates the sentiments consistence and emotional contagion theories into supervised cognition process and utilize sparse learning to tackle noisy txt in microblogging An empirical study on two authentic world Twitter dataset show the superior performance of given frame work in handling noisy and short tweets.

Ziato Liu [8] suggested a new feature selection method predicate on HowNet and Parts of Speech in his paper "Short Text Feature Selection for Micro-blog Mining". According to the composition of text property they utilize test data set accumulated from sina microblog. The result shows that the short text feature selection method has a substantial amount of information, and good classification result.

Stefan Stieglitz [9] in his paper "Political Communication and influence through Microblogging 6 an Empirical Analysis of Sentiments in Twitter Message and Retweet Behavior", seek to examine whether sentiment occurring in politically germane tweets has an effect on their retweet ability. Predicate on dataset of 64,431 political denoting affective dimension, including positive and negative emotions associated with certain political parties or politicians, in a tweet and its retweet rate. Furthermore, they investigate how political discussion take place in the Twitter network during the periods of political elections. Determinately, authors conclude by discussing the implicative insinuation of results.

Apoorv Agarwal [10] in his paper "Sentiment Analysis of Twitter Data" examine sentiment analysis on Twitter data. Their contribution of this paper are: (1) First introduce POS prior polarity feature. (2) Explore the utilization of tree kernel to obviate the desideratum for tedious feature engineering. The incipient feature and tree kernel perform approximately at same level, both outperforming the state of art of baseline.

Our presented work is similar to these previous work, we propose a method to semantically analyze microblogging data. In our work we have used the twitter data. We have applied the speech tagging on the data set and then classify the data using classification algorithm called Modified ID3 to prepare a predictive model .This model determines the moods of the user when they are using microblogging sites for the tweets.

A growing number of Location-based Social Network services provide time-stamped, geo-located data that opens new opportunities and solutions to a wide range of challenges. To analyze such social media data, the system provides the analysts with an interactive visual spatiotemporal analysis and spatial decision support environment that assists in evacuation planning and disaster management. Junghoon Chae [6] published a paper "Public behavior response analysis in disaster events utilizing visual analytics of microblog data" to demonstrate how to improve investigation by analyzing the extracted public behavior responses from social media before, during and after natural disasters, such as hurricanes and tornadoes.

3. Proposed Data Model

For the sentiment text analysis and their accurate evaluation a new system is prepared using the traditionally available techniques.

A) System Architecture

The organization of traditional methodologies for obtaining sentiment based text analysis is given in figure 1.

*a) **Method***

Step1: Both Test and Training Data set is preprocessed.
Step2: Speech tagging is done on preprocssed dataset. This is a process of marking up a word and tag with corresponding function in parts of speech.
Step3:Every word is replaced with its tag
Step4:Tagged data and associated tag is stored in relational database.Traditional ID3 algorithm and Modified ID3 algorithm is applied on tagged datasets which develops a decision tree model.
Step5:The data model is generated from the training set for classification.
Step6:Test data is applied to perform testing on the model and prediction on dataset is done.

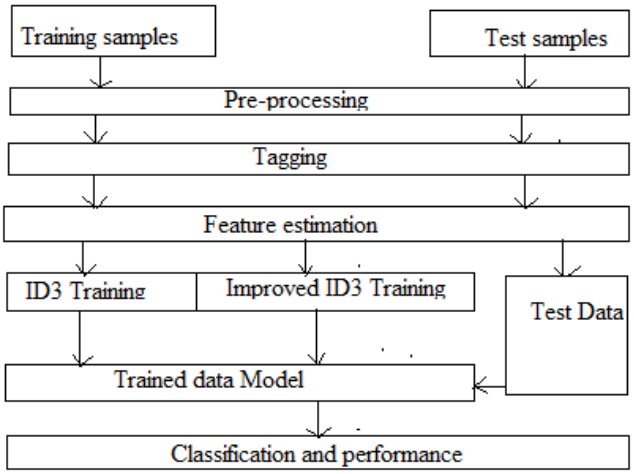

Figure 1: Proposed Model

B) Training Samples

The key aim of the system is to classify the tweets on the social media. Twitter data set is used for the experiment purpose.

C) Test Samples

From the available twitter data set is divided in to training (70% of the existing) and testing (30% 0f existing).

D) Pre-Processing

Both training and test data is pre-processed in this phase, the pre-processing of data involves the removal of punctuations and removal of frequently occurred words.

E) Tagging

It is required to involve feature on data after preprocessing. Therefore the user input tags are applied with the text such as:
Ram is a good boy.
Can be converted into: Noun adjective noun

F) Features Estimation

After tagging the original data is converted into a new encoded format. Therefore the tagged data and the associated tag is stored on a relational data base which contains the encoded attributes and their class labels. The example given using below given table.

Table 1: Feature Data

Noun	Pro-no	Verb	Adv	Adj	Pre	Conj	Good	bad	Classes
2	1	1	0	0	0	0	1	0	1

G) ID3 Training

The given table data is used to learn the traditional ID3 algorithm. Initially a provision is made to select algorithm for training. The system get training from traditional ID3 algorithm if user select it. Entropy and Information Gain is the important factors which is used to select the most useful attribute for classification.

$$\text{Entropy}(S) = \sum_{i=1}^{C} p_i \log_2 p_i$$

Where pi is the probability that any subset of data samples belonging to categories Ci.

Information Gain for attribute A on set S is defined by taking the entropy of S and subtracting from it the summation of the entropy of each subset of S, determined by values of A, multiplied by each subset's proportion of S[1].

$$I_G(S, A) = I_E(S) - \sum^{n} (p\ (\subset_S^{An}) * I_E\ (\subset_S^{An})$$

H) Improved ID3 Training

If user selects the Improved ID3, a decision tree data model from the training sample is developed. AF function correlation is the important factor used to carry out the importance of attribute.AF not only can well overcome the ID3's deficiency of tending to take value with more attributes, but it can also can represent the relations between all elements and their attributes[1].

$$AF = \sum |X_{i1} - X_{i2}| / n$$

Then, the normalization of relation degree function valueis followed

$$V_{(K)} = AF_{(k)} / AF_{(1)} + AF_{(2)} \ldots\ldots\ldots AF_{(m)}$$

$$\textbf{Gain}_{(A)} = \textbf{I}(\textbf{S}_{(1)}, \textbf{S}_{(2)}, \textbf{S}_{(3)} \ldots\ldots\ldots\ldots \textbf{S}_{(m)}) - \textbf{E}_{(A)} * \textbf{V}_{(A)}$$

I) Trained Data Model

Trained model is a finalized decision tree that makes use the input data and converted into a tree structure. This trained data model is prepared by ID3 and improved ID3.

J) Test Data

That is a part of training sample which is used to perform testing of trained data model using the cross validation technique. The cross validation results the accurate amount of data that is correctly recognized using the decision tree.

K) Classification and performance

Finally the performance of the entire system is computed in terms of accuracy, error rate, time consumption, and the memory consumption during training and testing of data.

4. Performance Analysis

After implementation of the proposed system the performance of Improved ID3 classification technique and previously available Traditional ID3 technique is evaluated and compared using their performance graphs. The detailed discussion about the preformed experiments and their results are given below:

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 = (Accurately Classified Patterns /Total Input Patterns) X100

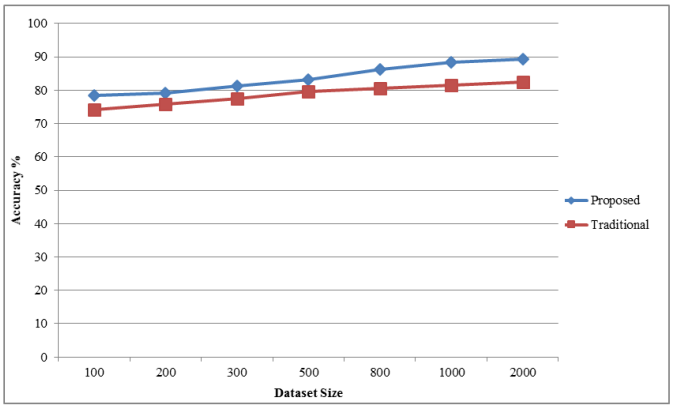

Figure 2: Accuracy of ID3 & Improved ID3

The given graph as specified in figure 2 contains the comparative accuracy of both the algorithms. In this figure blue line shows the Improved ID3 algorithm's performance and the red line shows the performance of the traditional ID3 approach. For demonstrating the performance of the system X axis contains the amount of data during the training and testing and Y axis contains the obtained performance in terms of accuracy. The values of graph is represented using table 2 where the amount of accuracy of the proposed algorithm is given in first column and the second column contains the values of traditional approach namely ID3.According to the obtained result the proposed algorithm shows better performance.

Table 2: Accuracy of Improved ID3 & ID3

Dataset size	*Improved ID3 algorithm*	*ID3 algorithm*
100	78.32	74.25
200	79.05	75.83
300	81.37	77.58
500	83.24	79.53
800	86.28	80.61
1000	88.35	81.47
2000	89.32	82.42

b) Error rate

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

Error rate % = (Total Misclassified Patterns /Total Input Patterns) X 100

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

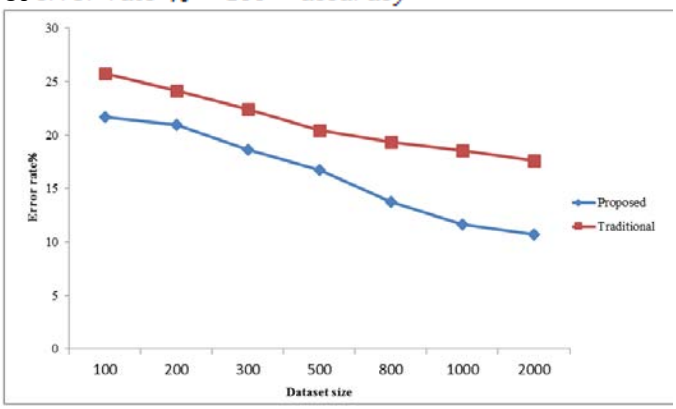

Figure 3: Error Rate ID3 & Improved ID3

The figure 3 and table 3 shows the comparative error rate of both the Improved ID3 and ID3. 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 Improved ID3 classification is effective and efficient during different experimentations and reducing with the amount of data increases. Thus the presented improved ID3 classification technique is more efficient and accurate than the traditional ID3 approach of text classification.

Table 3: Error Rate ID3 & Improved ID3

Dataset size	Proposed algorithm	Traditional algorithm
100	21.68	25.75
200	20.95	24.17
300	18.63	22.42
500	16.76	20.47
800	13.72	19.39
1000	11.65	18.53
2000	10.68	17.58

c) Memory Consumption

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

The amount of memory consumption depends on the amount of data reside in the main memory, therefore that effects the computational cost of an algorithm execution. The performance of both the implemented classifiers for sentiment classification is given using figure 4 and table 4. In this diagram the blue line shows the performance of the Improved ID3 classification scheme and the red line shows the performance of traditional ID3 classification scheme. For reporting the performance the X axis of figure shows the amount of data required to execute using the algorithms and the Y axis shows the respective memory consumption. According to the obtained results the performance of both the algorithm demonstrate similar behavior with increasing size of data, but the Improved ID3 technique consumes additional memory as compared to the traditional ID3 technique.

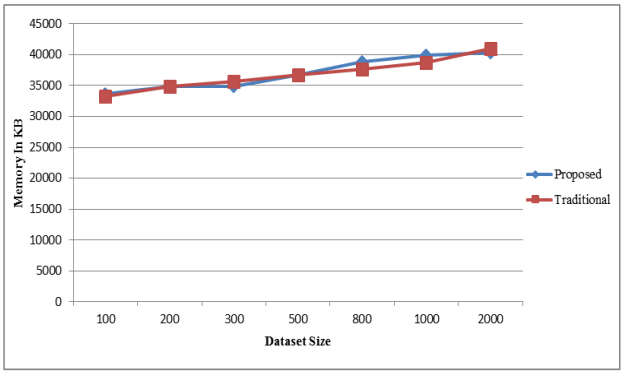

Figure 4: Memory Consumption ID3 & Improved ID3

Table 4: Memory Consumption ID3 & Improved ID3

Dataset size	Improved ID3algorithm	ID3 algorithm
100	33677	33304
200	34827	34881
300	34882	35718
500	36728	36718
800	38918	37661
1000	39928	38716
2000	40299	40928

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 comparative time consumption of the Improved ID3 and traditional ID3 algorithms is given using figure 5 . In this diagram the X axis contains the size of dataset and the Y axis contains time consumed in terms of milliseconds.

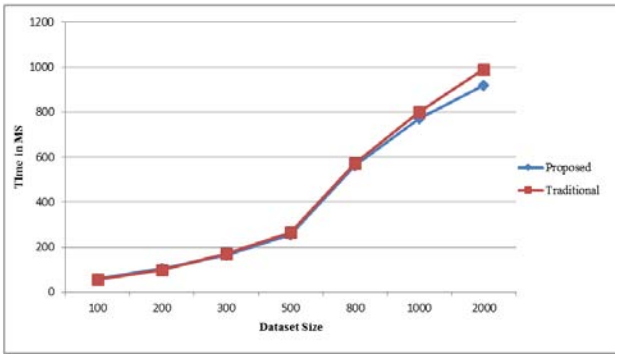

Figure 5: Time Consumption ID3 & Improved ID3

As given in table 5 comparative results analysis the performance of the Improved ID3 technique shows the less time consumption as compared to the traditional ID3 technique.

Table 5: Time Consumption ID3 & Improved ID3

Dataset size	*Improved ID3algorithm*	*ID3 algorithm*
100	60	57
200	105	100
300	163	170
500	256	266
800	562	573
1000	771	802
2000	918	991

5. Conclusion

The proposed work is intended to better analyze the text data according to the sentiments. Therefore the proposed study is focused on analyzing the text sentiments. The classification model for text data is prepared using classification algorithms such as ID3 and Improved ID3. In this work twitter dataset is used for sentiments based text classification. The row data improved by pre-processing, and tagging. The model is trained and tested to recognize the moods of users while using twitter.

According to the obtained results the system is able to classify the data according to their sentiments accurately. The proposed work is adoptable and efficient for classifying the patterns of the text data for bring out the emotions hidden in the text.

The proposed text classification based on the sentiment analysis is very useful for various domains of applications such as:

1) Good for understanding the phycology of the youth.
2) We can understand the thinking process of the youth on the basis of the tweets.

Some future extensions are as

1) Only two moods are predicted which is happy and sad. In future many other moods can also be predicted precisely helps to understand the user intention.
2) This approach is also extendable to classify the type of users.
3) Abbreviation of words in the tweets can be considered as microblogging sites users uses the abbreviation of words.

Limitation of the proposed work is as follows:

1) In this work tagging is done to involve the feature on data but it is time consuming because tagging is done manually token by token not line by line.
2) Can have a faster algorithm to classify the streaming in tweets .The algorithm used here in for classify the tweets have not sufficiently faster enough to deal with the online environment of streaming of data.

References

[1] Chen Jin, Luo De-lin, Mu Fen-xiang, “An Improved ID3 Decision Tree Algorithm”, Proceedings of 2009 4th International Conference on Computer Science & Education, IEEE©2009.
[2] B. V. Rama Krishna, B. Sushma, “Novel Approach to Museums Development & Emergence of Text Mining”, International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 2, No.2.
[3] Andreas Hotho, Andreas Nurnberger, Gerhard Paaß, Fraunhofer AiS, “A Brief Survey of Text Mining”, Knowledge Discovery Group Sankt Augustin, May 13, 2005.
[4] 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, Volume 2, No.8, August 2013.
[5] Vishal Gupta, Gurpreet S. Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, Volume 1, No.1, August 2009.
[6] Junghoon Chae, Dennis Thom, Yun Jang, SungYe Kim, Thomas Ertl, David S. Ebert, “Public behavior response analysis in disaster events utilizing visual analytics of microblog data”, & Elsevier Ltd. 2013.
[7] Xia Hu, Lei Tang, Jiliang Tang, Huan Liu, “Exploiting Social Relations for Sentiment Analysis in Microblogging”, WSDM, Rome, Italy, Copyright 2013 ACM 978-1-4503-18693/13/ 02, February 4–8 2013.
[8] Zitao Liu, Wenchao Yu, Wei Chen, Shuran Wang, Fengyi Wu, “Short Text Feature Selection for Micro-blog Mining”, Conference Paper, IEEE Xplore January 2011.
[9] Stefan Stieglitz, Linh Dang-Xuan, “Political Communication and Influence through Microblogging 6 An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior”, 45th Hawaii International Conference on System Sciences 2012.
[10] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, “Sentiment Analysis of Twitter Data”, Proceedings of the Workshop on Language in Social Media (LSM 2011), p.p 30–38, 23 June 2011