Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.7 No.2, 2018, pp.133-138

Optimization of Microblog Representation Using Deep Learning Approach

K. H. Walse
Professor& Head, Department of Computer Science and Engineering,
Anuradha Engineering College, Chikhli, Maharashtra, India
Email: kwalse1234@gmail.com

Abstract - **Microblogging, as a new form of online communication in which users talk about their daily lives, to gather real-time news, opinion about people or share information by short posts, has become one of the most popular social networking services today, e.g. to stay in touch with friends . Finding proper representations of microblog texts is a challenging issue. The overview of microblog included how to extract information from microblog, also discuss about Microblogging and Twitter. This paper included existing research on Optimization of Microblog Representation by using existing techniques and methods as well as microblogging services.Current approaches, for microblog reduction are single indexing. In this work, we have proposed double indexing method for microblog reduction. We have used our own generated dataset for microblog reduction. We have collected tweets from real time twitter for both single and double indexing technique. Also, we have use LDA with Semantic Similarity algorithm. We compared our results with single indexing. Experimental results shows that double indexing gives better performance than single indexing. This may happen because we will show both original and reduced tweets, so it will be known to user which tweet are removed and relevant part shows as output. We have also compared our results with the state-of-art. The proposed double indexing gives better performance than existing 1-ROCA of improved online SVM technique.**

Keywords: **Microblogs, Twitter, Text representation, Sentiment analysis, Topic detection, Clustering, SVM**

I. INTRODUCTION

Microblogging services have great popularity among the Internet users. From last few years, Twitter and SinaWeibo have grown most attractiveness in Internet users which are microblogging services. Users post the short text on social media, like twitter have microblog texts limit Of 140 character [1].Twitter has tons of row data and information, with user unable to classify it into right category. Microblog becomes a fast and wide information broadcasting channel [2]. Microblog short length helps to detect user's interested topic from microblog text rapidly [3]. Currently, the Internet is not used just for communication, but also as a platform for users to express their thoughts, ideas and feeling [4]. Motivated by deep learning approach, deep network produce low dimensional text, we apply the deep learning methodologies to perform optimization of tweet.There are some issues and challenges that motivate the development of new deep learning approach to improve dimensionality reduction of microblog. The main challenging issue is finding proper representation of microblog text. Microblog texts have length of 140 characters to each tweet [1]. The microblog poses serious challenges of traditional sentiment analysis and classification methods, just because of its inherent characteristics. The previous assumptions are broken by the inherent characteristics of microblog content, which call for SA approach which tolerant to high levels of noise such as: Sparsity, Non-standard vocabulary and Noise [3].Microblog Dimensionality Reduction has some limitations and we limit our scope to daily uses. Such as, we can represent the tweet of any user. Means user from any field, we can fetch its tweet only by knowing its twitter id. The research problems are formally stated here. Our research aims to effectively represent low dimensional representation of microblog text. This research is suitable for lower sized data, but as size increases, the data representation becomes more complicated.

To improve the result of dimensionality reduction, we take advantage of the semantic similarity derived from two types of microblog-specific information, namely the retweet relationship and hashtag [1]. In this dissertation we are proposing a new method called as Double Indexing, which is based on single indexing method. This method will provide proper representation of microblog text. Proposed method based on dimensional reduction method, which provide enhance search. It makes proper representation of tweet.We calculate the enhance performance of tweet. Compare the both performance using the indexing method and use the high performance indexing method. This paper discusses about techniques used by microblog and helps to understands the existing challenges and issues in this research field and also explain overview of propose system.

II. LITERATURE REVIEW

Microblogging is a form of blogging in which entries typically consist of short content such as phrases, quick comments. Notable services include Twitter, Tumblr and Google Buzz. Recently, as microblogging services have gained wide popularity, users can use it for novel purpose such as real time events [5].

Among the various microblogging services twitter is popular service. Twitter is a s ocial networking and microblogging services that allow user to send and read 140 characters short messages known as tweets, also user can share and discover topics of interest in real time [5].Microblog is an information publication, spread and huge platform based on relationships between the users. We briefly review some microblog reduction studies that are related to existing technique are shown in following Table I.

TABLE I EXISTING TECHNIQUE SUMMARY

S. No.	Paper Title	Year	Author	Technique	Outcome	Drawback
1.	Microblog Dimensionality Reduction—A Deep Learning Approach	2016	Lei Xu, Chunxiao Jiang, Yong Ren and Hsiao-Hwa Chen	Deep learning methodology	Dimensionality reduction of tweet	Sparse Data problem and HDP-based models are not applicable to obtaining representations of tweets
2	Chinese Microblog Topic Detection Based on the Latent Semantic Analysis and Structural Property	2013	Xia Yan and Hua Zhao	Microblog topic detection method	Solving data sparseness problem	Different user will have different wording style, so the traditional topic detection cannot be applied to microblog topic detection directly.
3.	Enhancing Accessibility of Microblogging Messages Using Semantic Knowledge	2011	Xia Hu, Lei Tang and Huan Liu	Feature extraction	To extract different features from the entities	The accessibility of messages has been very limited
4.	Sentiment Analysis of Chinese Micro-blog Using Vector Space Mode	2014	Zhi-Qiang Xian, Y. X. Zou and Xin Wang	The proposed SACM system	This paper shows classification accuracy upto 80.86%	The length of words less than 140 character
5.	Content vs.Context for Sentiment Analysis: a Comparative Analysis over Microblogs	2012	Fotis Aisopos, George Papadakis, Konstantinos Tserpes and Theodora Varvarigou,	Sentiment analysis features	Content-based features captured by n-gram graphs and context-based captured by polarity ratio	There is no standard tokenization method for multilingual documents
6.	Efficient Deep Learning Approach for Dimensionality Reduction using Micro blogs from Big data	2017	Mr. M. Vengateshwaran, Mrs. C. Ramyapriyadarsini and Ms. N. Valarmathi	Feature extension for short text	Information broadcasting	Multi-class short-text has lower performance as well as incorrectness
7.	Enriching short text representation in microblog for clustering	2012	Jiliang TANG, Xufei WANG, Huiji GAO, Xia HU, Huan LIU	Enriching short text representation in microblog for Clustering	Multilanguage knowledge integration and feature reduction concurrently through matrix factorization techniques	Their limited text length, general shortenings and the problems of synonymy and polysemy
8.	Research on Microblog Filtering Technology Based on Improved Online Support Vector Machine Model	2016	HuiNing, Song Li, FanhuZeng and Li Xu	Microblog Filtering Technique based on improved Online SVM	Decrease classification errors and improve the classification of data	Topic unrelated microblog

A. Deep Learning Methodology: Inspired by the success of deep learning on traditional texts, in this paper apply, deep learning methodologies to perform dimensionality reduction on tweets. Used retweet and hashtag information used for modify the training set. For training the deep networks specific semantic knowledge is utilized [1].

B. Microblog Topic Detection Method: Traditional topic detection method cannot be applied on microblog topic detection directly, because it was short and grass-root text. Microblog topic detection method is based on the latent semantic analysis and the structural property. As consideration of the dialogic property of the micro-blog, our scheduled method first creates semantic space based on the replies to the thread, with the goal of solving data sparseness problem. Second, create the micro-blog model based on the latent semantic analysis. Lastly, combined the semantic computation with time information [6].

C. Feature Extraction: A methodology has been developed that includes preparation of semantic database and then employ it to extract the necessary classification features from the database. The Feature extraction process for short and grass-root text is complex. It is more complex in world of social blogging, also user try to use synonyms. To extract different features from the entities and then use these extracted features to train the machine [7].

D. The Proposed SACM System: SACM investigates using a vector space model. Microblogs characteristically consists of one or two sentences with the length of words less than 140 character. So the granularity used in this study of SACM system is at the sentence, word or phrase level. The SACM system consists of data preprocessing, feature extraction, Feature dimension reduction, Training Classifier and Evaluation. The proposed SACM system reaches 80.86% classification accuracy [2].

*E. Sentiment Analysis Features:*Microblog content poses serious challenges to the applicability of traditional sentiment analysis and classification methods, due to its inherent characteristics. To overcome this problem introduces a new method that relies on two orthogonal, but corresponding sources of proof: content-based features captured by n-gram graphs and context-based ones captured by polarity ratio [3].

F. Feature Extension for Short Text: The continuous growth of immediate messaging technology and the spread of information processing technology stimulated the growing of short-text information processing technology, like mobile phone SMS, QQ chat , BBS, instant messaging software has become a main channel for information broadcasting [8][9].

G. Enriching Short Text Representation in Microblogfor Clustering: Social media websites allow users to interchange short texts such as Tweets in microblogging, user status in friendship networks. These short text pose has challenges to traditional text mining task, e.g. clustering [8][10][11].

*H. Microblog Filtering Technique Based on Improved Online SVM:*Microblog filtering technology is applied to the micro-blog service. By using this, reduce the running time of the filter by reducing the training set, reducing the number of training times and reducing the number of iterations. Micro-blogs are divided into two categories by the filtration system: topic related micro-blog and topic unrelated microblog [12][13][14].

III. OVERVIEW OF PROPOSED WORK

We discuss the proposed methodology with its design and implementation. We start our section with Microblog dimensionality reduction process. Microblog dimensionality reduction is a sub-domain of Microblogging. Microblog dimensionality reduction process is consisting of real time data, blogging and immediate messaging.

A.Microblog Dimensionality Reduction

We begin microblog dimensionality reduction with a description of the real time data process and the blogging considered for this work. We then describe the Microblog fetching agent and finally we discuss the Semantic search algorithm and Single character indexing technique that were used for performing the double indexing task.Our main aim is to compare single indexing techniques with Double indexing techniques, to check whose performance gives better accuracy over another. The dimensionality reduction processes for both indexing is same and it performed individually.

We proposed a framework for double indexing technique, which is based on single indexing technique. This technique will provide proper representation of microblog text. These steps include data collection, data stored in indexing; evaluate performance and comparison of result. By this approach, real time tweet can be fetched and stored into single and double index database.

Plan of work divide into following modules

1. Development of microblog fetching agent.
2. Use single indexing to store results from the blog fetcher.
3. Evaluate the performance of single indexing blog fetcher.
4. Use double indexing to store results from the blog fetcher
5. Valuate the performance of double indexing blog fetcher
6. Comparisons of result.

B. LDA with Semantic Similarity

In this Latent Dirichlet Allocation the graphical model representation of LDA in these boxes is plate representing replicates. The Outer plates represent the documents and the inner plates represent the repeated inner choice of topics within documents.

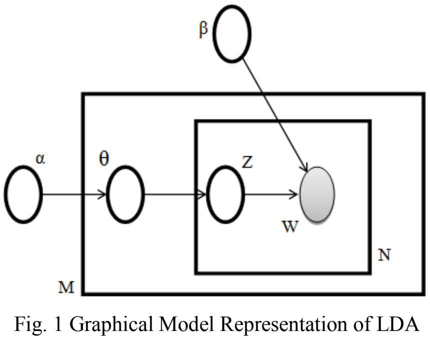

Fig. 1 Graphical Model Representation of LDA

This model is based on f ollowing notations and terminology.

1. A word is the basic unit of discrete data, defined to an item to from a vocabulary indexed by{1...v}.We represents words using unit-basis vectors that have a

single component equal to one and all other components equal to zero.

2. A document is a sequence of N words denoted by a w={w1,w2,w3...wN}
A corpus is a collection of M documents denoted by
D= {w1, w2…wM}

```
Input: a collection of positive training documents D;
minimum support sj as threshold for topic Zj; number
of topics V
Output: U_E = {E(Z1),…..E(ZV)}
  1: Generate topic representation f and word-topic
     assignment zd; i by applying LDA to D
2: U_E := □
3: for each topic Z_j _ [Z_1, Z_v] do
4: Construct transactional dataset +_j based on □
      and zd, i
5: Construct user interest model X_zj for topic Z_j using
     a pattern mining technique so that for each pattern
     X in X_zj,
     supp(X) >□ j
6: Construct equivalence class EðZjÞ from XZj
7: U_E :=U_E □ {E(Z_j)}
8: end for
```

Algorithm 1.LDA with Semantic Similarity

IV. IMPLIMENTATION

Implementations of our proposed work are as follows

A. Fetching Agent

For collecting real time tweets, we can develop fetching agent. It can fetch the real time tweets of any user, whose twitter id we know. Fetched tweets are stored in database. Database is based on single index and double index technique. Tweets are stored in both indexes. Following figure shows the implementation of microblog fetcher with performance index and comparison.

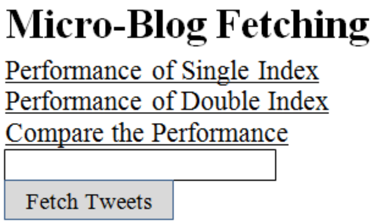

Fig. 2 Performance and Comparison links of Single and Double Index

B. Flow Chart

A Flow chart is type of diagram that represents workflow or process of our work. Following flowchart represents different steps involved in execution of Microblog Reduction.

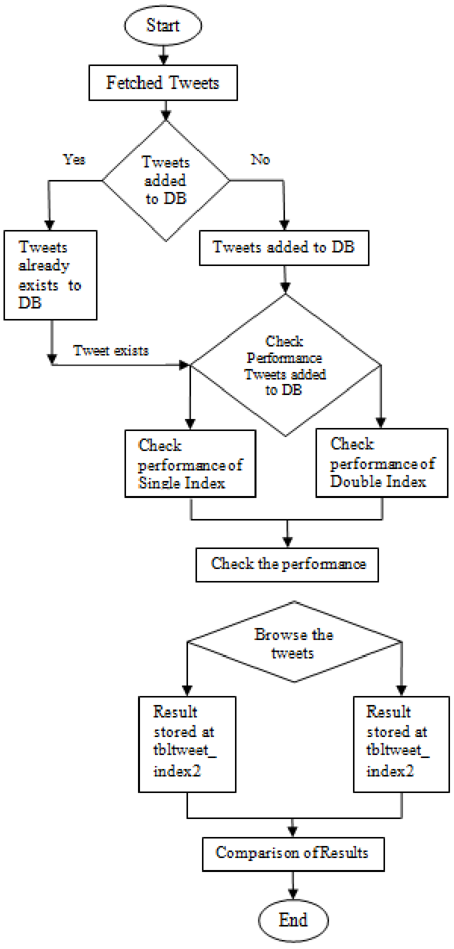
Start
Fetched Tweets
Yes
Tweets added to DB
No
Tweets already exists to DB
Tweets added to DB
Tweet exists
Check Performance Tweets added to DB
Check performance of Single Index
Check performance of Double Index
Check the performance
Browse the tweets
Result stored at tbltweet_ index2
Result stored at tbltweet_ index2
Comparison of Results
End

Fig.3 Flow Chart of Microblog Reduction in Own dataset

V. EXPERIMENTAL RESULT ANALYSIS

We conducted experiment on the Twitter dataset to acknowledge future users with some results. For this purpose, we proposed Double Indexing technique for dimensionality reduction. To evaluate whether the proposed approaches can help to get better low dimensional of tweets, for conducted experiment. In this section, we first describe the preparation of datasets. Then we show the results, after that comparison of results.

A. Dataset

To evaluate our model, we conduct experiments on dataset of Twitter (fetch real time tweets). This tweet is obtained with the help of fetching agent. We sample the data by knowing twitter id of any person.

B. Calculate the Mean delay

Calculate the mean delay of double index and single index.
Mean delay = Delay / Number of tweets fetched

C. Comparision on Own Dataset

For comparing the efficient performance of single index and double index, we fetch the same tweets in both indexing.

For example, twitter-id @ZEEBUSINESS fetches the 9 tweets, whereas @my-iplclub fetches 13 tweets, @Sushama fetches 3 tweets and @thehill fetches 3 tweets, and then stored in single indexing.

TABLE II NUMBER OF TWEET FETCHED IN SINGLE INDEX AND DOUBLE INDEX

Twitter-Id	Number of tweets fetched in Single Index	Number of tweets fetched in Single Index
@ZEEBUSINESS	9	9
@my-iplclub	13	13
@Sushama	3	3
@thehill	3	3

D. Comparison of Mean Delay on Own Dateset

Table III shows the comparison of mean delay of single index and double index. Whose indexing having low time delay, it performance good. Hence the result of double indexing is more effective than single indexing for all twitter ids. We have also explained this in graphical format.

TABLE III COMPARISON OF MEAN DELAY ON THE BASIS OF FETCHING TIME

Twitter-id	Mean Delay of Single Index	Mean Delay of Double Index
@ZEEBUSINEE	0.000075	0.000046
@my_iplclub	0.000035	0.000027
@Sushama	0.000523	0.000075
@thehill	0.000337	0.000075

Following Fig. 3 Graphical representation of Mean Delay in Single Index. We take on x-axis Number of searches and on y-axis Mean Delay.

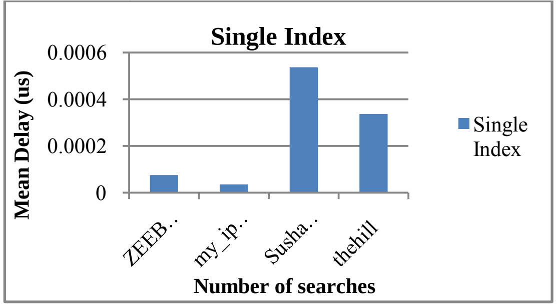

Fig. 3 Graphical representation of Mean Delay in Single Index

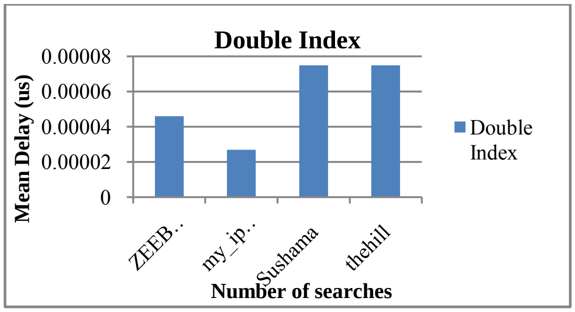

Fig. 4 Graphical representation of Mean Delay in Double Index

Following graph represents the comparison of Mean Delay of Single Index and Double Index. We take on x-axis Number of Searches and on y-axis Mean Delay.

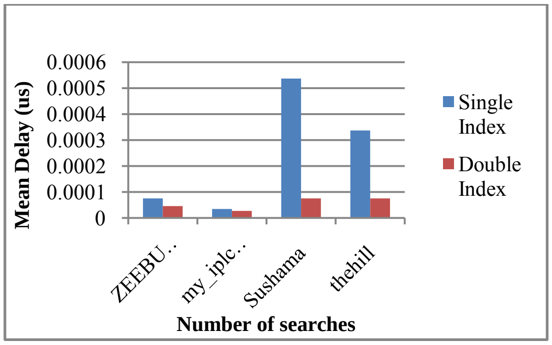

Fig. 5 Comparison graph of Mean Delay of Single Index and Double Index

E. Existing Microblog Filteration Technique

Microblogs divided into two categories by filtration system: Topic related to microblog and Topic unrelated to microblog.Existing 1-ROCA of improved online SVM method used parameter such as, Microblog Topic m and Mean Delay. Our proposed Double Indexing method use parameter such as, Number of searches and Mean Delay. Compare the time values of both methods, irrespective of number of searches, because the delay is the mean delay for different searches. Double indexing gives very low mean delay over 1-ROCA of improved online SVM technique. So, double indexing gives better performance.

TABLE IV COMPARISON GRAPH OF MEAN DELAY OF 1-ROCA OF IMPROVED ONLINE SVM AND MEAN DELAY OF DOUBLE INDEX

Topic	Mean Delay of 1-ROCA of improved online SVM	Mean Delay of Double Index
1	0.2051	0.000046
2	0.0006	0.000027
3	0.1037	0.000075
4	0.0001	0.000075

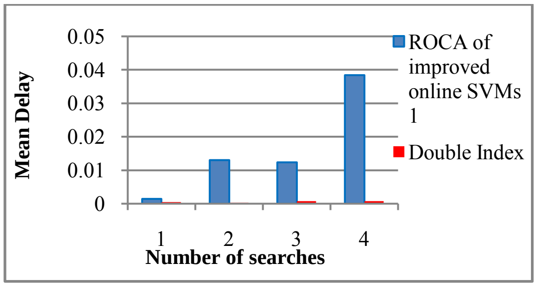

Fig. 6 Comparison graph of Mean Delay of 1-ROCA of improved online SVM and Double Index

From above comparison graph it is clear that, Mean delay of Double Indexing take very less time over 1-ROCA of improved online SVM. From that we can say Double Indexing take low timing, it gives better performance.The smaller this ratio the more convenient is the user experience, so we have the better experience in case of Double Indexing.

VI. DISCUSSION

We recently reviewed the most prominent works on microblog reduction. Among them, we discuss the deep learning approach of microblog reduction. The main goal of this work is to implement double indexing technique. For that, we spent our time to obtaining dataset. We were finally able to obtain the dataset using tweet fetcher agent. It can fetch real- time tweets and stored into single and double index database. After that, we compared the fetching time of both indexing. As compared to single indexing, double indexing has very low time delay. So, it gives better performance over single indexing. From that we can say, our method is more successful.

VII. CONCLUSION

In this work, we have used own generated dataset of tweets for microblog reduction. We specifically focus on how to apply deep networks to perform dimensionality reduction on microblog texts. In this review, we study various existing techniques and methods. Along this report a concise study of the previous work on the microblog reduction is presented. We present a method to efficiently identify reduction of tweets and also used LDA with Semantic Similarity algorithm. From the graphical comparison of time we concluded that, double indexing gives better performance and have low time delay than single indexing. Although the results gives better performance for double indexing. Also, our double indexing gives better performance than existing 1-ROCA of improved online SVM technique. Double indexing has very low time delay over 1-ROCA of improved online SVM. There is more scope for research in this area.

REFERENCES

[1] Lei Xu, Chunxiao Jiang, Senior Member, IEEE, Yong Ren, Member, IEEE, and Hsiao-Hwa Chen, Fellow, IEEE, “Microblog Dimensionality Reduction—A Deep Learning Approach”, In*Knowledge and Data Engineering, 2016, IEEE Transaction*, Vol. 28, No.7,pp. 1779-1789, 2016.

[2] Zhi-Qiang Xian, Y. X. Zou and Xin Wang,“Sentiment Analysis of Chinese Micro-blog Using Vector Space Model”, 2014.

[3] Fotis Aisopos, George Papadakis, Konstantinos Tserpes and Theodora Varvarigou, “Content vs. Context for Sentiment Analysis:a Comparative Analysis over Microblogs”, In *Proceedings of the 6th ACM*, p. 12, pp. 187-196, ACM 2012.

[4] Xiaoqian Lui and Tingshao Zhu, “Deep learning for constructing microblog behavior representation to identify social media user’s personality”, In *PeerJ Comput. Sci. 81, 2016,* pp. 1-15, 2016.

[5] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi , “Want to be Retweeted? Large Scale Analytics onFactors Impacting Retweet in Twitter Network.” *IEEE International Conference on Social Computing, 2010*, pp. 177 – 184, 2013.

[6] Xia Yan and Hua Zhao, “ Chinese Microblog Topic Detection Based on the Latent Semantic Analysis and Structural Property.” In *JOURNAL OF NETWORKS,* Vol. 8, No. 4, pp. 917-923, 2013.

[7] Xia Hu, Lei Tang and Huan Liu, “Enhancing Accessibility of Microblogging Messages Using Semantic Knowledge.” pp. 2465-2468, 2011.

[8] Mr. M. Vengateshwaran, Mrs. C. Ramyapriyadarsini and Ms. N. Valarmathi, “Efficient Deep Learning Approach for Dimensionality Reduction using Micro blogs from Big data”, *IJRASET,* Vol. 5, Issue 2, pp. 5-10,2017.

[9] Ms. Payal R. Rathi and Dr. K. H. Walse, “Survey on optimization of microblog Representation”, In Proceedings *International Conference-EECCMC 2018*, pp. 1-8, Jan 28 and Jan 29, 2018.

[10] Jiliang TANG, Xufei WANG, Huiji GAO, Xia HU, Huan LIU, Computer Science & Engineering, Arizona State University, Tempe, AZ 85281, USA“Enriching short text representation in microblog for clustering”, In *Front. Comput. Sci.,* Vol. 6, No. 1, pp. 1-13, 2012.

[11] Kishor H. Walse, Rajiv V. Dharaskar and Vilas M. Thakare “A Study on the Effect of Adaptive Boosting on Performance of Classifiers for Human Activity Recognition”, In*Proceedings of the International Conferenceon Data Engineering and Communication Technology*, Advances in Intelligent Systems and Computing, Vol. 469, pp. 419-429, 2017.

[12] Shrunkhala Satish Wankhede, Ms. S. A. Chhabria and Dr. R. V. Dharaskar “Contolling Mouse Cursor Using Eye Movement”, In *IJAIEM,* pp. 1-7, 2013.

[13] K. H. Waslse and D. R. Dhotre “Wireless Network: Performance Analysis of TCP”, *Information Technology Journal*, Vol. 6, No. 3, pp. 363-369, 2007.

[14] Hui Ning, Song Li, Fanhu Zeng and Li Xu “Research on Microblog Filtering Technology Based on Improved Online Support Vector Machine Model0”, *IEEEInternational Conference on Mechatronics and Automation,* pp. 2326- 2332, 2016.