International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

# Using SVM and Stopword removal method in Microblogging Classroom

**Vidya Dhuttargaon**[1], **Amit R. Sarkar**[2]

[1]ME-II, SVERI College of Engineering, Pandharpur, Solapur University, India

[2]Professor, Department of C.S.E, SVERI College of Engineering, Pandharpur, Solapur University, India

**Abstract:** ***In recent year, Microblogging is a popular technology in social networking applications. All users publish online short text messages with the help of micro blogging which size less than 200 characters. It can be use via web and instant messaging clients etc.It can be an effective tool in the classroom and education community.Microblogging can be categorization for two types such as relevant and irrelevant questions. In relevant questions consistof questions that the teacher wants toaddress in the class. In a question text, empirical results and analysis show that personalizationbetween each other. With the help of question text Microblogging leads to better categorization accuracy.It is also important to utilize the correlation between questions, lecture materials, correlation between questions asked in a lecture. The experimental results likewise demonstrate that the end of stop words which prompts better relationship estimation in the middle of inquiries and arrangement exactness. Then again, fusing understudies' votes on the inquiries that don't enhance classification exactness, despite the fact that a comparable highlight has been indicated to be successful in group inquiry noting situations for evaluating question quality.***

**Keywords:** Data set, Support VectorMachines (SVM), Removal and Cosine Similaritywith Tf-Idf and Okapi techniques.

## 1. Introduction

Microblogging is a web 2.0 technology which is a type of blogging that lets the users post short text messagesto their community in real time such as e-mails, web, messengers, and mobile devices. The size of text message should be less than 200 characters. Twitter [11] is a microblogging tools which can be effectively used for successfully utilized for expert systems. Recently, the fundamental utilization microblogging instruments in classroom situations as a specialized instrument between an understudy and the teacher [9], [27]. There are many advantages and disadvantages of microblogging.The classrooms are that the quantity of inquiries remarks a teacher gets from the understudies. To the best of our insight, there is exceptionally restricted former chip away at the classification of applicable and unessential microblogging messages or inquiries in classroom situations [4]. Earlier deal with instructor specialists incorporates an inquiry positioning capacity that scores every inquiry and tries to choose the best inquiries to react to [24]. They use the inquiry message and in addition a customized methodology (i.e., questions from an understudy who has been asking great inquiries are favored) to separate between the pertinent and unimportant inquiries. As of late, Cetintas et al. likewise used the relationship between inquiries and accessible address materials in addresses alongside personalization and inquiry message, and demonstrated that it further enhances the order exactness [4]. Other comparable former work in group inquiry noting situations (where clients post an inquiry and have their inquiries addressed by others, for example, Yahoo! Answers [32],Wiki Answers [31], and Baidu Zhidao [2] use users' votes to figure out which questions should to rank high for question search [25],[26]. In spite of the fact that these methodologies are truly essential to enhance the adequacy of the questions arrangement, those methodologies don't consider 1) the correlation between questions asked in the same lecture, 2) the impact of removing stop words or not over the classifiers' performance. This paper proposes a content categorization approach that can naturally distinguish relevant and irrelevant questions asked in an lecture.It is demonstrated that when address materials are definitely not accessible, the methodology of using the connections between inquiries asked in an address turns into a decent distinct option for the methodology of using the connection between inquiries

At long last, it is demonstrated that disposal of stop words prompts better relationship estimation in the middle of inquiries, and prompts better classification precision. Yet, end of prevent words from the highlight space of classifiers or while figuring the relationship in the middle of inquiries and accessible address materials do not have a noteworthy effect.

## 2. Related Work

In Related Work we discussed, Inmicroblogging identification of relevant and irrelevant questions supported classroom.Microblogging has as of late been utilized as a specialized device between a student and the instructor [9], and additionally with different students. The principle center is on how microblogging can be utilized academically [9], and on the best way to examine microblogging in the setting of learning [3], [7]. The important issue in microblogging upheld classrooms and separation classrooms. In this Scenario is the quantity of instructoran educator Receives from the students. It can be bigger than what can be answered in a constrained time. Subsequently, there is the need to choose the best questions to react to [4].Soh et al. endeavor to comprehend this issue in their work on instructor specialists with an inquiry positioning ability that scores every inquiry and tries to choose the best inquiries for the educator to react to [24]. Question content and personalization (i.e., favoring inquiry originating from an understudy who has been asking great inquiries) is utilized to separate between the applicable and superfluous inquiries. In spite of the fact that question content and personalization are very vital for recognizing pertinent and unimportant inquiries, their work disregards 1) using the correlation between questions and available lecture materials and 2) using the correlation among questions asked in a lecture.

In a Recent work shows that, using thecorrelation between questions asked in a lecture and theavailable lecture materials,it is help to better identify relevant and irrelevant questions [4]. They don't consider utilizing the relationship between questions themselves asked in an lecture. Besides, they don't consider the impact of stop words evacuation on the classifier performance, when they are not uprooted for the bag-of-words and when the connections among questions themselves and the relationship among questions and accessible lecture materials are computed.

## 3. Data

In this point data is collected from personal finance class. The microblogging tool, HotSeat, is particular for utilization in classroom, so it can be utilized for any course and encourages correspondence between an student with the instructor and other student which is possessed by Purdue University and has been created by the third and the fourth creators. The teacher and student of a class are individuals from the microblogging class. In every lecture, a different page is made so student can post their short messages related with that specific address. The framework likewise has a double voting system such that each questions. The framework additionally has a double voting instrument such that each inquiry address has a normal of 26.9 important inquiries with a standard deviation of 9.4 and 10.8 insignificant inquiries with a standard deviation of 8.5. The biggest aggregate number of inquiries saw in anlecture is 69 and there is exceptionally set number of rehashed lectures due to the voting framework. Samples of significant and unessential inquiries can be seen in Table 1. We utilize two human annotators (the first creator and a specialist in fund) and ask them to expound every inquiry as either being significant orunessential. The annotators achieve a Kappa of 0.868 on 162 inquiries (i.e., inquiries of the initial four lectures) and thusly whatever remains of the information was expounded by the first annotator just. Each lecture has an openly accessible presentation record significant to the lecture and they are utilized as theaccessible pertinent lecture materials with no adjustment. The course has a syllabus record that examines about course policies, exams, projects, quizzes, insights

## 4. Methods

In this section, we describe the techniques of Support Vector Machines, Stop words Removal and Cosine Measure with Tf-Idf and Okapi.

## 5. SVM

In Microblogging questions/messages are in textual format. Thus, identifying their sorts can be dealt with as a content categorization issue. Bolster Vector Machines have been indicated to be a standout amongst the most precise and broadly utilized content categorization strategies [15], In this work, SVM i.e., with a straight portion was utilized as text categorization classifier that can be figured as an answer for an optimization problem as underneath:

$$\{w, b\} = \min ½ \, || \, \psi \, ||2 + C = \textstyle\sum_{i=0}^{N} £i \qquad (1)$$

Subject to yi (w.d + b) – 1 + £i
Where di represents is the ith document,
Yi= {-1, +1} represent binary classification of di.

Note that, the positive class (i.e., y = +1) speaks to the significant questions, and the (i.e., y = -1) speaks to the negative class irrelevant questions. w! , b is the parameters of the SVM model. C has the control over the exchange off between classification precision and margin, which is tuned observationally. The categorization of every SVM classifier is found out by two fold cross approval in the preparation phase.

### Cosine Similarity with Tf-Idf and Okapi:

Cosine Similarity Tf-Idf and Okapi is a measure of likeness between two vectors by figuring the cosine of the angle between them. It is in text mining to analyze content archives. In this work, the similitude scores between questions are figured by the basic Cosine measure [1] as below:

$$\mathrm{Sim}\,(Qi, Qj) = (\cos(Qi, Qj)^{n} = \frac{Qi * Qj}{||Qi||\,||Qj||} \qquad (2)$$

Where Qi is the ith question and Qj is the jth question in an address, Qi and Qj are sentences spoken to as the bag-of-words vectors.
"*" signifies the speck result of these vectors.

The connection between questions and lecture materials are additionally ascertained in the same way.. We can use two common weighting schemes, viz. Tf-Idf [1] and Okapi [21]. Tf-Idf used for term frequency and inverse document frequencyand Okapi utilized for furthermore considers report sizes by favoring shorter yet pertinent reports.

## 6. Models

### Modeling with Terms and Personalization

The main goal of Modeling with Terms and Personalization is that to select the best questions to respond. It is instinctive to utilize personalization along side the terms of questions since diverse understudies have distinctive questions asking propensities and a few understudies have a tendency to ask more significant or unessential questions amid the addresses than different understudies. Thusly, usage of personalization highlights empowers the grouping model to adjust to understudies' diverse inquiry asking propensities by utilizing the data from their previousquestions.In this work, we utilize two highlights for personalization, for example, (i) rate of applicable questions asked by an student, and (ii) rate of immaterial questions asked by an student. SVM classifier that uses personalization along side the terms. There are two gauge classifiers, for example, SVM_Terms and SVM_Terms Pers, resp.

**Modeling with Terms, Personalization, and Students' Votes**

In this modeling approach, another binary-valued highlight, whether an question gets more than 5 percent of all votes in an address or not, is joined not with standing the terms and personalization offers that has been said some time before. This approach also same as SVM_Terms Pers Votes.

**The Effect of Eliminating Stopwords**

In this point the fourth arrangement of investigations was led to assess the adequacy of including or barring stop words. In the other word that constitutes the highlight space of the SVM models and constitutes the highlight space to be utilized by the Cosine Similarity. Which is utilized to gauge the likeness between inquiries in an address and the comparability in the middle of inquiries and accessible address materials? It can be seen in Table 3.that including or barring stopwords does not prompt steady execution upgrades (that are huge with p-esteem not exactly 0.01) over one another. Yet, including stopwords is by all accounts more vigorous and is reliable with earlier work [24]; subsequently, it is decided to be the default for whatever is left of the tests (i.e., Table 4). In the same way, evacuating stopwords while ascertaining the connections between inquiries and accessible address materials has been found to be not altogether unique in relation to including them. In further experimentation, evacuating stopwords from the highlight space of all SVM classifiers has been discovered to be not altogether unique in relation to the

## 7. Conclusions, Discussions, Future Work

In this paper proposes a novel utilization of content classification to recognize applicable and immaterial microblogging questions in a classroom. Numerous demonstrating and weighting designs are examined for this application through experiments. It demonstrated to be gainful to use the relationship among inquiries and accessible address materials, relationships between inquiries asked in an address. It is discovered to be altogether more compelling to uproot stop words when computing the relationships among inquiries themselves. Toward the end understudies' votes on inquiries is discovered not to be powerful, in spite of the fact that it has been demonstrated to be helpful in Group inquiry noting situations for inquiry quality appraisal.Accordingly, this work does not unequivocally manage incorrect spellings, contractions, and so forth, and utilize the inquiries as they are. In a comparative chip away at inquiry subjectivity examination over

Client produced inquiries, it is noticed that the addition procured by utilizing just the inquiry content is practically identical to the addition gained by using NLP, which produces more mind boggling highlights. In this way, utilizing NLP is not justified regardless of the expanded time and space unpredictability [18]. Yet, it is advantageous to investigate the impact of using NLP on this application 1) to improve the highlight space and 2) to manage the poorly designed inquiries in a different work completely. Second, stand out course is utilized for experimentation as a part of this work. More courses can be utilized to evaluate the heartiness of the proposed calculations. Future work will be directed mostly in that direction.

## 8. Acknowledgments

We thankful to Teacher, Friends and Department of Computer Engineering for their constant guidelines and support. We also thank there searchers as well as publishers for making their resources available and teachers for their guidance. We also thankful college authorities for providing the required infrastructure and support.

## References

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, pp. 75-82. Addison Wesley, 1999.

[2] Z. Baidu, http://zhidao.baidu.com, Dec. 2010.

[3] K. Borau, C. Ullrich, J. Feng, and R. Shen, "Microblogging forLanguage Learning: Using Twitter to Train Communicative and Cultural Competence," Proc. Eighth Int'l Conf. Web Based Learning, pp. 78-87,

[4] S. Cetintas, L. Si, Xin, S. Chakravarty, H. Aagard, and K. Bowen,"Learning to Identify Students' Relevant and Irrelevant Questionsin a Micro-Blogging Supported Classroom," Proc. 10th IntelligentTutoring Systems (ITS '10) Conf., pp. 281-284, 2010.

[5] S. Cetintas, L. Si, Y.P. Xin, D. Zhang, and J.Y. Park, "AutomaticText Categorization of Mathematical Word Problems," Proc. 22$^{nd}$Int'l FLAIRS Conf.,

[6] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi, "Short andTweet: Experiments on Recommending Content from InformationStreams," Proc. 28th ACM Int'l Conf. Human Factors in ComputingSystems, pp. 1185-1194, 2010.

[7] C. Costa, G. Beham, W. Reinhardt, and M. Sillaots, "Micro- Blogging in Technology Enhanced Learning: A Use-Case Inspectionof PPE Summer School 2008," Proc. Workshop Social InformationRetrieval for Technology Enhanced ccosta_microblogging.pdf, Dec. 2008.

[8] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, "Searching Questions byIdentifying Questions Topic and Question Focus," Proc. 46th ACLConf., pp. 156-164, 2008.

[9] G. Grosseck and C. Holotescu, "Can We Use Twitter forEducational Activities?" Proc. Fourth Int'l Scientific Conf., eLearningand Software for

[10] P. Han, R. Shen, F. Yang, and Q. Yang, "The Application of CaseBased Reasoning on Q&A System," Proc. Australian Joint Conf.Artificial Intelligence,

[11] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter:Understanding Micro-Blogging Usage and Communities," Proc.Ninth WEBKDD Conf., pp. 56-

[12] J. Jeon, B. Croft, and J. Lee, "Finding Semantically SimilarQuestions Based on Their Answers," Proc. 28th ACM SIGIR Conf.,pp. 617-618, 2005.

[13] J. Jeon, B. Croft, J. Lee, and S. Park, "A Framework to Predict the Quality of Answers with Non-Textual Features," Proc. 29th ACMSIGIR Conf. Research and Development in Information Retrieval,pp. 228-235, 2006.

[14] V. Jikoun and M. de Rijke, “Retrieving Answers from FrequentlyAsked Questions Pages on the Web,” Information and Knowledge mgnt pp. 76-83, 2005.

[15] T. Joachims, “Text Categorization with Support Vector Machines:Learning with Many Relevant Features,” Proc. 10th European Conf. pp. 137-142, 1998.

[16] J.Keefer,“HowtoUseTwitterinHigherEducation,”http://silenceandvoice.com/archives/2008/03/31/how-to-use-twittering-Higher education, Mar. 2008.

[17] Lemur IR Toolkit, http://www.lemurproject.org, Dec. 2010.

[18] B. Li, Y. Liu, A. Ram, and E. Agichtein, “ExploringQuestion Subjectivity Prediction in Community QA,” Proc. 31stACM SIGIR Conf. pp. 735-736, 2008.

[19] M. Michelson and S. Macskassy, “Discovering Users’ Topics of Interest on Twitter: A First Look,” Proc. Fourth Workshop AnalyticsFor Noisy Unstructured Data in Conjunction with the 19th ACM CIKMConf., 2010.