I.J. Information Engineering and Electronic Business, 2017, 4, 31-37
Published Online July 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2017.04.05

An Improved Information Retrieval Approach to Short Text Classification

Indrajit Mukherjee
Department of Computer Science & Engg. Birla Institute of Technology Mesra, India
Emai: imukherjee@bitmesra.ac.in

Sudip Sahana
Department of Computer Science & Engg. Birla Institute of Technology Mesra, India
Emai: sudipsahana@bitmesra.ac.in

P.K. Mahanti
Department of Computer Science University of New Burnswick Saint John, Canada
Emai: pmahanti@unbsj.ca

***Abstract*—**Twitter act as a most important medium of communication and information sharing. As tweets do not provide sufficient word occurrences i.e. of 140 characters limits, classification methods that use traditional approaches like "Bag-Of-Words" have limitations. The proposed system used an intuitive approach to determine the class labels with the set of features. The System can able to classify incoming tweets mainly into three generic categories: News, Movies and Sports. Since these categories are diverse and cover most of the topics that people usually tweet about .Experimental results using the proposed technique outperform the existing models in terms of accuracy.

***Index Terms*—**Twitter, topic modeling, Word-Sense Disambiguation .

I. Introduction

Recently online social media has emerged as a medium of communication and information sharing. Status updates, blogging, video sharing and social networking are some of the ways in which people try to achieve this. Popular online social media sites like Facebook, Orkut or Twitter allow users to post short message to their homepage. These are often referred as micro-blogging sites and the message is called a status update. Status updates from Twitter are more commonly called as tweets.

Tweets are often related to some event based on topic of interest like music, dance or personal thoughts and opinions. A tweet can contain text, emotion, link or their combination.

Tweets have recently gained a lot of importance due to their ability to disseminate information rapidly. Twitter defines a low level information news flashes portal (Bharath Shriram, 2010).

Obviously, even if this system cannot represent a serious alternative to the authoritative information media, considering the number of its authors and the impressive response time of their contributions, Twitter can provide a real-time system that can also predate the best newspapers in informing the web community about the emerging topics. Popular search engines like Google and Bing have started including feeds from Twitter in their search results. Researchers are actively involved in analyzing these micro-blogging systems. Some research areas include understanding usage and communities (Java et al. 2007) discovering user characteristics (Rao et al. 2010), detecting spam (Zheng et al. 2015) and so on.

In this paper, an unsupervised classifier are created and evaluated, which will be able to automatically classify tweets. The classifier does not require any external information and analyze the tweets using an unsupervised machine-learning approach. Once the classifier has been built, is evaluated by compared with other classifiers and it show better results.

This paper is organized as follows: In section 2, overview of the existing works is presented, Section 3, contains the system design and the implementation details. Section 4, contains the experimental results, comparison and its explanation. Finally in Section 5, draws conclusion into current limitations and future research scope.

II. Related Work

More recent works do away with web searches and instead utilize data repositories. One of the richest data sources of information is the Wikipedia. By integrating knowledge available within the Wikipedia, tweets can be enhanced with more semantic knowledge. On the other hand, online querying of Wikipedia and parsing concepts are not suitable for real-time applications because of the time constraints. Although it eliminates the problem of data sparseness, it is very time consuming and there is a need to understand what concepts of Wikipedia are useful to extract. Also, there is a need to analyze the final

features extracted and eliminate redundant and useless features from classification. This may again require querying the web to select only the necessary features. Use of Wikipedia concepts to determine closeness between texts was explained in (Gabrilovich & Markovitch 2007). Concept-Based Information Retrieval Using Explicit Semantic Analysiswas described in (Egozi O. et al. 2011). Adding semantic to microblog posts was introduced in (Meij E. et al. 20012). A topic vector based space model for short text classification was introduced in (Chen M. & Jin X. 2011).

Dimensionality reduction (DR) of text datasets for feature selection has been widely adopted in the past. Yiming Yang et al. (Yang and Pedersen, 1997) performed a comparative study of some of these methods including, document frequency, information gain, mutual information, $\chi 2$ -test(CHI) and term strength(TS). They concluded that IG and CHI are the most effective in dimensionality reduction. Selection of the best feature subset by increasing relevancy of feature with target class and reducing redundancy between chosen features by using mRMR technique (Peng et al., 2005).

Topic models have been applied to a number of tasks that are relevant to our goal of classifying Twitter status messages. There has been some work with regards to using topic models for information discovery. (Allen J. 2002) presents a framework to build classifiers using both a set of labeled training data and hidden topics discovered from large scale data collections. An unsupervised algorithm described in (Cataldi M. 2010) extracts both the topics expressed in large text collection and models how the authors of the documents use those topics. Such author-topic models can be used to discover topic trends, finding authors who most likely tend to write on certain topics and so on.

III. System Design and Implementation

System design provides an architectural overview of the proposed system.

A. Normalization of Tweets

Tweets are written extremely colloquially, containing an unusually high amount of repetition, novel words and interjections. Normalizing tweets would make work in this area easier as well as in any other area that involved analyzing tweets. The goal is to remove as much noise as possible from tweets, so any elements that were not absolutely necessary to form a grammatical English sentence were removed. For example, the tweet "@user213 hw r u?? i'm gud:" would have been translated as "How are you? I am good". Data preprocessing is done to eliminate the incomplete, noisy and inconsistent data such as a*bbreviation* (shortform), abbreviation(acronym), typing error/misspelling, punctuation omission /error, wordplay, removing URLs, removal of retweets. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process.

B. Topic modeling

Datasets are created based on 1.47 million tweets collected from the Twitter streaming API from Jan to April 2015, and pre-process the tweets and bring the tweets into a normalized English form. Efficient techniques have been employed to clean tweets before classification. These normalized tweets are then manually labeled as belonging into one of the three classes namely news, movies and sports. In case the tweet exhibits flavors of multiple classes, the best possible class is chosen as the label. After labeling the tweets, Topic modeling is applied on the three classes of tweets to extract features for the classes.

Topic models do not make any assumptions about the ordering of words. It disregards grammars as well. This is particularly suitable to handle language and grammar irregularities in Twitter messages.

MALLET (Machine Learning for Language Toolkit)is used to generate different topic model and an open source software toolkit which provides a Java-based package to do various machine learning tasks.

A topic modeler consists of three stages.

*Input stage***:** This stage involves converting the training corpus into an acceptable format. MALLET is used to build the system provides a special input command for converting the training data into MALLET's special internal format also remove certain stop words from the tweets during this stage.

Training stage: Once the data is available in MALLET's internal format, train the input data using topic modeling based on MALLET's 'train-topics' command using the number of topics as 200.After training it stands out that a value in the range 200 to 400 gives reasonably fine-grained results.

Output stage: The output from a topic modeler is typically an inference file and a file containing top 'k' words associated with each topic. The inference file is used to infer topics from the data set. These topics stand out to be the features for classifying the tweets. In proposed system, 2 values of 'k' are considered.

'k' = all the topics
'k' = top 10 topics ,i.e., the default value in MALLET.

C. Feature reduction

Selecting a subset of relevant features for building robust learning models is another research problem. Hence MALLET is used to select the feature set, which generally follows the definitions of classes.

When consider ('k' = all the topics), then such a large feature set leads to a problem of Curse of Dimensionality. As the feature set becomes too large, tweet becomes difficult to visualize and the basis for classification is lost. Increase in number of features also results in higher model building time and makes the classification slower. Hence, there is a need to effectively analyze and prune features and reduce the feature size to an optimal value. Also, there might be several overzealous, unimportant features that degrade the performance of a classifier.

The feature set is reduced in order to efficiently classify the tweets ('k' = the top 10 topics). Experimental results show that with only a small set of features, the classifier achieves a significant improvement in accuracy when compared to the previous feature set. But, this feature set also contains multiple features which are common among the classes. A feature should correspond to only one class. There is a need for presence of discriminating features per class. The quality of the feature set is very critical to the performance of the system. The presence of these multiples common features bring down the quality of the accuracy. They lead to more number of misclassification. Thus, these common features are also removed from the feature set with the help of WSD and removing these features further increases the accuracy of the system.

D. Classification

Classification is a supervised data mining technique that involves assigning a label to a set of unlabeled input objects. The various classifier algorithms applied are:

- Naïve Bayes classifier
- Support Vector Machine(SVM)
- Classification Tree
- k Nearest neighbor

The proposed method is an unsupervised classifier which has no such information available still experimental results show that the proposed method outperform some of the supervised classifiers and give better result in comparison to them.

Proposed Algorithm:

STEP 1: The features extracted for the three distinct classes are stored in files.

STEP 2: The new tweet which has to be correctly classified and the feature sets are fed into the system.

STEP 3: The tweet is then disambiguated. Disambiguation involves tokenizing the tweet, making the tokens Case-less, removing stop words, lemmatizing the tokens using WordNet, stemming the tokens and finally the stemmed tokens are Part of Speech tagged.

STEP 4: A loop executes on each word in the tweet. A POS tagged word is selected and all senses of that word are learned.

STEP 5: If the learned sense is not a Verb or Noun then it is ignored and skip to the next sense.

STEP 6: Loop on all other words in the same tweet and find their senses.

STEP7: Then the definition of all the senses are extracted from WordNet.

STEP 8: The senses of a particular word are then compared with the senses of the remaining words. An overall score is evaluated and the maximum score is then considered.

STEP 9: The senses which give these maximum scores are then returned.

STEP 10: The steps from 4 to 9 are also executed on the feature sets.

STEP 11: Wu & Palmer (WUP) Similarity between the senses of the feature sets and the words of the tweet are then evaluated.

STEP 12: The feature set which gives the maximum similarity with the tweet is considered the correct feature set. The class of the feature set is then extracted and the tweet is classified to that class.

The proposed method, basically classifies a tweet based on the semantic similarity between the tweet and the three feature sets. It classifies a tweet based on the WUP-Similarity.

This proposed method also outperforms Latent Semantic Analysis (LSA) which is an unsupervised classifier.

IV. Experimental Results

In the first set of experiments, we came up with approximately 4800 tweets taken randomly from different labeled as belonging into one of the three classes namely News, Sports and Movies. The distribution of tweets per class is shown below:

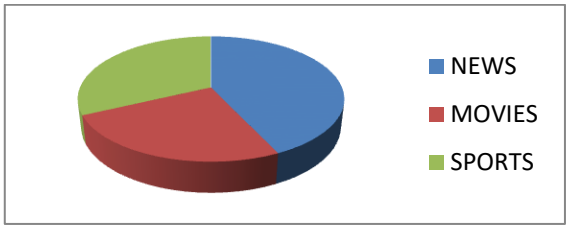

Fig.1. Distribution of tweets per class

The classification algorithms, namely Naïve Bayes, Classification Tree, k Nearest neighbor and Support Vector Machine were used on the training data.

The result that SVM classifier gives (without Topic Modeling) is as follows:

Table 1. Confusion Matrix of SVM without Topic Modeling

	NEWS	SPORTS	MOVIES	
NEWS	1552	0	48	1600
SPORTS	1552	0	248	1800
MOVIES	1274	0	126	1400
	4378	0	422	4800

This result in Table 1 is given by SVM when the feature set is obtained by Term Frequency method. This method does not give good result. So in order to improve the accuracy Topic Modeling is used to extract the features. Topic modeling is then used on the normalized tweets to find the features for the different classes. MALLET is used as the topic modeling tool.

After extracting the features three types of feature sets are used. The *first feature set* which contain all the features extracted from MALLET, are used to classify a tweet correctly to the respective class. This feature set doesn't give good accuracy. Such a large feature set results in higher model building time and makes the classification slower. Thus, the feature set is reduced.

The *second feature set* contains those top 10 features which have the highest weightage. This feature set gives very good accuracy in comparison to the previous set. The *third feature set* is a further reduction of the second set. The second feature set is further reduced by removing the common features from the different classes and also removing the words having semantic similarity using *Word-Sense Disambiguation (WSD)* (Chen P. el al. 2009). WSD is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings.

These feature sets are then fed to the above mentioned classifiers. The classification results have now been analyzed using a confusion matrix. A confusion matrix is a visualization method used in machine learning, typically when the number of categories exceeds 2. Given an amount of possible categories, the rows display the predicted category, while the columns display the actual categories. Thus, it is possible to check if the classifier is confusing two categories hence the name 'confusion matrix'.

Table 2. Confusion Matrix of Naive Bayes (all features)

	MOVIES	NEWS	SPORTS	
MOVIES	840	278	282	1400
NEWS	378	840	382	1600
SPORTS	288	288	1224	1800
	1506	1406	1888	4800

Table 3. Confusion Matrix of SVM (all features)

	MOVIES	NEWS	SPORTS	
MOVIES	1288	54	58	1400
NEWS	95	1400	105	1600
SPORTS	813	810	177	1800
	2196	2264	340	4800

Table 2 and 3 shows the confusion matrices obtained when all feature set are taken and is applied to the classifiers Na ïve Bayes and SVM respectively.

Due to the large size of the feature set, the accuracy obtained is not good. So, the feature set is further reduced. When the top 10 features set is fed to these classifiers, the accuracy improves. The confusion matrices obtained are as shown below.

Table 4. Confusion Matrix of Naive Bayes (10 top features)

	MOVIES	NEWS	SPORTS	
MOVIES	1120	135	145	1400
NEWS	57	1480	63	1600
SPORTS	155	153	1492	1800
	1332	1768	1700	4800

Table 5. Confusion Matrix of SVM (10 top features)

	MOVIES	NEWS	SPORTS	
MOVIES	1232	168	0	1400
NEWS	0	1600	0	1600
SPORTS	0	527	1273	1800
	1232	2295	1273	4800

Table 4 and 5 shows the improved result after feature reduction. The result gets further improved when the third feature set is applied to the classifiers. The confusion matrices obtained are as shown in Table 6 and 7.

Table 6. Confusion Matrix of Naive Bayes (reduced features)

	MOVIES	NEWS	SPORTS	
MOVIES	1120	138	142	1400
NEWS	0	1518	82	1600
SPORTS	21	199	1580	1800
	1232	2295	1273	4800

Table 7. Confusion Matrix of SVM (reduced features)

	MOVIES	NEWS	SPORTS	
MOVIES	1176	224	0	1400
NEWS	0	1559	41	1600
SPORTS	24	370	1406	1800
	1200	2153	1447	4800

These results show that the use of Topic Modeling increases the efficiency of the classifiers.

The proposed algorithm (unsupervised algorithm) on the other hand, gives better result in comparison to some of the supervised classifiers. The Proposed algorithm considers the feature sets obtained from MALLET and then uses these feature sets to classify a tweet. This algorithm does not use any of the available classifiers and still gives much better result. The confusion matrix obtained when the third feature set is applied to the proposed method is as shown in Table 8.

Table 8. Proposed algorithm Confusion Matrix

	MOVIES	NEWS	SPORTS	
MOVIES	1300	0	0	1400
NEWS	571	1029	0	1600
SPORTS	0	0	1800	1800
	1871	1029	1800	4800

In fact, the proposed algorithm gives also much better result when compared to the method used by the unsupervised Latent Semantic Analysis (LSA) to classify a tweet. The confusion Matrix of table 9 shows the accuracy of LSA.

Table 9. Confusion Matrix of LSA

	MOVIES	NEWS	SPORTS	
MOVIES	0	700	700	1400
NEWS	0	1600	0	1600
SPORTS	0	0	1800	1800
	0	2300	2500	4800

As seen in the above Table 9, the proposed unsupervised algorithm gives better result in comparison to the supervised algorithms SVM and Naives Bayes. But, as the feature set 1 is too large so it leads to higher model building time and also makes the classification slower. So, the features set are further reduced to top 10 features and its accuracy is improved. This feature set also contains certain some common features which lead to

misclassification. Thus, the further reduced features into third category gives much better accuracy.

The proposed algorithm (PA) also gives better accuracy with respect to well known supervised classifier and the unsupervised classifier LSA.

Table 10. Shows the accuracy obtained by the proposed algorithm, SVM, Na ïve Bayes and LSA for different categories of feature sets. In table 1., proposed algorithm (PA) shows better result with respect to SVM, Na ïve Bayes and LSA. The PA achieve better accuracy for the case of reduce features and also reduce the execution time.

Table. 10. Accuracy of Proposed Algorithm (PA) with different classifiers (different categories of features Sets)

Algorithms	Accuracy (%)
PA (all features sets)	70.00
SVM (all features sets)	63.00
Na ïve Bayes (all features)	60.00
PA (top 10 features sets)	81.93
SVM (top 10 features sets)	86.00
Na ïve Bayes (top 10 features)	85.13
PA (reduced top10 features)	88.06
SVM (reduced top10 features)	86.51
Na ïve Bayes (reduced top 10 features)	87.57
LSA	67.00

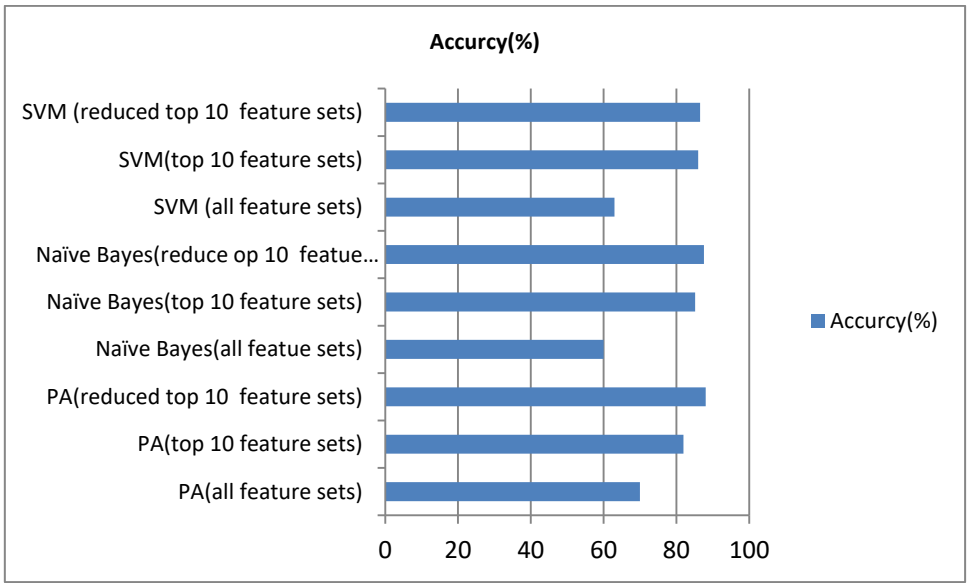

Fig.2. Accuracy of Proposed Algorithm (PA) with different classifiers (with 3 different categories of features Sets)

Fig. 2 shows the accuracy of the proposed classifier in comparison with different classifiers (different categories of features Sets). Reduced features set give much better result than other classifiers.

V. Conclusion

The work described in this paper is a step towards efficient classification of tweets using topic modeling. Tweets are harder to classify than larger corpus of text. This is primarily because there are few word occurrences and hence it is difficult to capture the semantics of such messages. Hence, traditional approaches when applied to classify tweets do not perform as well as expected.

Existing works on classification of tweets integrate messages with meta-information from other information sources such as Wikipedia and WordNet. Automatic text classification and hidden topic extraction approaches perform well when there is meta-information or when the context of the tweet is extended with knowledge extracted using large collections. But these approaches require online querying which is time-consuming and unfit for real time applications. These approaches eliminate the problem of data sparseness but enhance the feature set.

A framework is proposed to classify Twitter messages which have 140 character limits. In this framework, three different categories of feature sets have been used to classify incoming tweets into three generic categories – News, Sports and Movies. Here, the method used to classify a tweet is an unsupervised method as it does not require any source of data or labeling the tweets. The proposed method gives much better accuracy than the existing supervised classifiers such a SVM and Naïve Bayes. It even gives better accuracy over the unsupervised classifier LSA.

A "perfect classifier" does not exist. It is always a compromise between several factors that are application dependent. However, the underlying goals of all classifiers are the same, higher accuracy and better speed. In this paper, we have tried to achieve both the goals but there is a scope for a lot of improvements.

The experiment is done with Twitter messages only. The feature set is tailored towards various characteristics of tweets like presence of @, shortening of words etc. We hope to come up with a generic framework that can perform consistently well on different types of microblogs and multi-lingual tweets.

REFERENCES

[1] Bharath Shriram, short text classification, Ohio university, 2010.

[2] Shu Zhang et al., Semi-supervised Classification of Twitter Messages for Organization Name Disambiguation, International Joint Conference on Natural Language Processing, 869–873, Nagoya, Japan, 14-18 October 2013.

[3] S Zhang et al., Semi-supervised Classification of Twitter Messages for Organization Name Disambiguation, International Joint Conference microblogs. In Conference on Social Media, 31, 2010.

[4] I. Hemalatha et al., Pre-processing the Informal Text for efficient Sentiment Analysis, IJETTCS,1,2, 2012.

[5] A. Java X. Song, T. Finin, and B. Tseng, Why we twitter: understanding microblogging usage and communities. In Procs WebKDD/SNA-KDD '07 (San Jose, California), 56-65, 2007.

[6] J.Allan, editor. Topic detection and tracking: event-based information organization, Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[7] Kriti Puniyani, Jacob Eisenstein, Shay Cohen, and Eric P. Xing, Social links from latent topics in microblogs. In Conference on Social Media, page 31, 2010.

[8] Richard Beaufort, Sophie Roekhaut, Louise-Aḿelie Cougnon, and C ́edrick Fairon, A hybridrule/model-based finite-state framework for normalizing SMS messages. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 770–779, Uppsala, Sweden, 2010.

[9] J.Allan, editor, Topic detection and tracking: event-based information organization. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[10] Mario Cataldi, Luigi Di Caro and Claudio Schifanella, Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation, MDMKDD,2010

[11] Rao, D.; D., Y., Shreevats, A., and Gupta, M., Classifying Latent User Attributes in Twitter. In Proceedings of SMUC-10, 710–718, 2010.

[12] Edgar Meij, Wouter Weerkamp, and Maarten de Rijke, Adding Semantic to Microblog Posts, In proceedings of 5th ACM Web Search and Data Mining, pages 563-572, 2012.

[13] Offer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich, Concept-Based Information Retrieval Using Explicit Semantic Analysis, ACM Transactions on Information Systems, Vol. 29, No. 2, Article 8, 2011

[14] Quan, X.J., Liu, G., Lu, Z., Ni, X.L., Liu, W.Y., Short text similarity based on probabilistic topics. Knowl. Inf. Syst., 25(3):473-491, 2010.

[15] Chenglong Ma, Weiqun Xu, Peijia Li, Yonghong Yan, Distributional Representations of Words for Short Text Classification, Proceedings of NAACL-HLT '15, Denver, Colorado, 33–38, 2015.

[16] Stephane Clinchant and Florent Perronnin, Aggregating continuous word embeddings for information retrieval, In Proceedings of the Workshop on CVSM and their Compositionality, ACL,Sofia, Bulgaria, 100–109, 2013.

[17] Aixin Sun, Short text classification using very few words, In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, 1145–1146, 2012.

[18] X. Zheng et al., Detecting spammers on social networks, The Journal of Neurocomputing,159, 27–34,2015.

[19] Mengen Chen, Xiaoming Jin, and Dou Shen. Short text classification improved by learning multigranularity topics. In Proc. of IJCAI, 1776–1781, 2011.

[20] Gabrilovich,E. and Markovitch, S., Computing semantic relatedness using wikipedia-based explicit semantic analysis, In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07). Morgan Kaufmann Publishers, 1606–1611, 2007.

[21] I. Mukherjee, V. Bhattacharya, P.K. Mahanti, Samudra Banerjee, Text Classification using Document-Document Semantic Similarity, International Journal of Web Science(1757-8795), 2, 1-2, 2013.

[22] Toriumi, Fujio, and Seigo Baba. Real-time Tweet Classification In Disaster Situation, Proceedings of the 25th International Conference Companion on World Wide Web and International World Wide Web Conferences Steering Committee, 2016.

[23] Hussain, Muhammad IrshadAlam, Evaluation of graph centrality measures for tweet classification, Proceedings of the International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), 2016.

[24] Joao Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy and Abdelhamid Bouchahia, A Survey on Concept Drift Adaptation, ACM Computing Surveys, Vol. 1, No. 1, Article 1, January 2013.

[25] Castillo, C. ; Mendoza, M. , and Poblete, B., Information credibility on twitter, In Proceedings of the 20th International Conference on World Wide Web, WWW '11, pages 675–684, New York, NY, USA, ACM, 2011.

[26] Kairam, S. R., Morris, M. R., Teevan, J. Liebling, D. , and Dumais, S., Towards supporting search over trending events with social media, In Proceedings of ICWSM 2013, the 7th International AAAI Conference on Weblogs and Social Media, 2013.

[27] I. Mukherjee, Jasni M Zain, P.K. Mahanti, An Automated Real-Time System for Opinion Mining using a Hybrid Approach, I.J. Intelligent Systems and Applications, 7, 55-64, 2016.

[28] Celik, Koray, and TungaGungor, A comprehensive analysis of using semantic information in text categorization, Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on IEEE, 2013.

[29] A. Zubiaga, D. Spina, V. Fresno, R. Mart ńez, Real-Time Classification of Twitter Trends, Journal of the American Society for Information Science and Technology (JASIST), 2014.
[30] Bing-kun WANG, Yong-feng HUANG, Wan-xia YANG, Xing LI, Short text classification based on strong feature thesaurus, J Zhejiang Univ-Sci C (Comput & Electron) 2012 13(9):649-659, 2012.
[31] Ping Chen, Wei Ding, Chris Bowes, David Brown, A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge, The 2009 Annual Conference of the North American Chapter of the ACL, pages 28–36, Boulder, Colorado, June 2009.
[32] Lili Yanga, , Chunping Lia , Qiang Dingb, Li Lib, Combining Lexical and Semantic Features for Short Text Classification, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, 2013
[33] Xiaojun Quan , Gang Liu , Zhi Lu , Xingliang Ni , Liu Wenyin. Short text similarity based on probabilistic topics. Knowledge and Information Systems, v.25 n.3, p.473-491, 2010.
[34] Paolo Ferragina , Ugo Scaiella. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). Proceedings of the 19th ACM international conference on Information and knowledge management, 2010.
[35] Jinhee Park,Sungwoo Lee, Hye-Wuk Jung and Jee-Hyong Lee. Topic word selection for blogs by topic richness using web search result clustering. Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication, 2012.
[36] Gu Bin, S. Sheng Victor, Keng Yeow Tay, Walter Romano, Shuo Li, Incremental Support Vector Learning for Ordinal Regression, IEEE Transactions on Neural Networks and Learning Systems, vol. 67, no. 2015.
[37] Anuj Mahajan, Sharmistha, and Shourya Roy, Feature Selection for Short Text Classification using Wavelet Packet Transform, Proceedings of the 19th Conference on Computational Language Learning, pages 321–326, Beijing, China, July 30-31, 2015.

Authors' Profiles

Indrajit Mukherjee received M.Sc., Electronics degree from the University of Ranchi, India in 1995, MCA degree from BIT Mesra, Ranchi, India in 2001, PhD(Computer Science) from BIT Mesra, Ranchi, India in 2013. Currently, he is an Assistant Professor in the Department of Computer Science & Engineering, BIT Mesra Ranchi, India. His research interests include Web-Based learning, Data Mining, Big Data Handling, Web Service Applications, and Soft Computing. He has more than 20 research papers to his credit. He is a lifetime member of Indian Society for Technical Education (ISTE), India.

Sudip Kumar Sahana was born in Purulia West Bengal, India on 8th October, 1976. He received the B.E degree in Computer Technology from Nagpur University, India in 2001, and the M.Tech. degree in Computer Science in 2006 from the B.I.T (Mesra), Ranchi, India where he has done his Ph.D.(Engineering) in 2013. His major field of study is in Computer Science. He is currently working as Asst. Prof. in the Department of Computer Science and Engineering, B.I.T(Mesra), Ranchi, India. His research and teaching interests include soft computing, grid computing, network traffic management and artificial intelligence. He is author and reviewer of number of research papers in the field of Computer Science. He also serve as editorial member of various international journal of repute. He is a lifetime member of Indian Society for Technical Education (ISTE), India.

Prabhat Kumar Mahanti, male, is Professor of Dept of Computer Science (CS), University of New Brunswick Canada. He obtained his M.Sc. from IIT-Kharagpur, India, and Ph.D. from IIT-Bombay India. His research interests include Software engineering, software metrics, reliability modelling, modelling and simulation, numerical algorithms, finite elements, mobile and soft computing, verification of embedded software, neural computing, data mining, and multi-agent systems. He has more than 100 research papers, technical reports to his credit.

How to cite this paper: Indrajit Mukherjee, Sudip Sahana, P.K. Mahanti,"An Improved Information Retrieval Approach to Short Text Classification", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.9, No.4, pp.31-37, 2017. DOI: 10.5815/ijieeb.2017.04.05