DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4277

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

IMPROVED INFORMATION FILTERING FOR TWITTER BASED ON THE SEMANTIC KNOWLEDGE

Kuldeep Singh
Dept. of Computer Science and Engineering
J.B Institute of Technology
Dehradun, India

Himanshu Suyal
Dept. of Computer Science and Engineering
Govind balabh Pant Engineering College
Ghurdhuri, Pauri Garhwal, India

Sunil Bisht
Dept. of Computer Science and Engineering
Govind balabh Pant Engineering College
Ghurdhuri, Pauri Garhwal, India

Abstract: As the outgrowth of the internet as well as the social networks like twitter, Facebook user may get flooded out of raw information. Twitter message is short and may not contain enough contextual information, so traditional clustering method which utilizes the traditional method like "Bag-of-words" accept some restriction. To overwhelm with this trouble, we offered an automatic text classification process. Some social networking sites have imposed limits on the no of character for the users like twitter imposed limit of only 140 characters to the users to post any tweet. To classify these kinds of tweets is a tedious task or even impossible to classify. Short text is hard to sort out due to the lack of semantic information, therefore in this research paper, a novel approach is presented that incorporate the semantic database and utilized it to elicit the necessary features to separate the short text. Experimental results indicate that the proposed approach effectively classify the incoming tweets into predefined categories such as 'News', 'Events', 'Personal message' and 'Deals'.

Keywords: short text, twitter, clustering, classification, semantic, feature extraction.

I. INTRODUCTION

Text classification is widely played very important roles in many application fields. With the enhancement of the web applications or social networks, a number of short texts are increasing. Close to popular social networking websites, twitter, restrict the users to only 140 characters and this led to the users to express their expressions or thoughts in less number of rows. To automatically classify these short texts into predefine category is a tedious job because a short text contains less word co-occurrences and less contextual information which is not sufficient to separate. So traditional clustering algorithm cannot be enforced on the short text because it has very poor classification [1].

Classification is an operation of putting data into one or more predefined category of division. Classification of the short text is very complex and it became more complex in the field of social networking because people often use slang and synonyms to express their views and emotions. To surmount this trouble, we suggested a novel approach which is grounded on the formulation of the semantic data- set to classify the short text and experimental results indicate that the suggested methodology is making more accurate results as compared to the previous methodology.

The remainder of the paper is organized as follows. In Section 2 focus on the literature review. In Section 3 describes the detailed discussion on our proposed methodology. In Section 4 discusses about the experimental results. We have discussion about the conclusion and future work on Section 5.

II. LITERATURE SURVEY

One of the primary challenges of text classification is that it is hard to separate the text data due to its sharpness and high dimensionality. It became more difficult to distinguish the short text due to lack of word co-occurrence and information because of the limited duration of short text, so some traditional clustering algorithm causes the very poor result of classification [1]. To surmount this trouble with existing work, several methods of the classification for short text came into the delineation. Sankaranarayanan et al. [2] proposed a novel approach to classify the tweet into two classes: News and Non-News. Hong et al. [3] proposed an approach in which different schema is compared to classify the tweets. Sri-ram et al. [4] compared some features to master the problem of short text. Close to existing work on the short text classification is one room to integrate some extra knowledge on the short text by using some background knowledge like Wikipedia, WordNet, etc. [5-6]. By incorporating background knowledge, short text can contain more word co-occurrence. Banerjee, et.al [5] directed a query on the database in which Lucene index is built which used the snapshots of Wikipedia. The main disadvantage of this approach is that it is not feasible for the highly volatile data like news feed and it will not capture up-to-date information. On the other hand, online query on Wikipedia is not practicable for the real time system. Hu, et.al [6] used Wikipedia and WordNet for enhancing the existing features. Wikipedia was used to the concept and WordNet for the key words. Also, they proposed a hierarchical approach in which short text is categorized into segment, phrases and words.

Another means to convert the short text into longer text by applying the external repository or search engine like

Google, Bing, etc., for extracting the more information about the short text [7-8]. For every pair of short text, they squared up the similarity by retrieving statistics of the search engine consequences. Although these types of techniques need more emphasis on the addition of disambiguation approaches, for example 'jaguar" and "car" are concerned, but when web search is done many hits came for the "animal" rather than the "car". Hence on that point is some need of explicit feedback from the users to direct searching whether they want "car" or "animal". It will not do well on the semantic similarity because it is time consuming and not feasible for the real time systems. Although, these techniques deliver an advantage that they can place pre-dominant terms between messages so there is no need to compute word co-occurrence because they are likely to belong to the same context. The other advantage of web search is that it doesn't require pre-existing taxonomy. And then such methods can be used in several applications.

In the recent work probabilistic topic modeling also widely practiced in the area of text mining [9-11]. The basic thought behind the topic modeling is that to extract the topic from the domain related database because the subject is relatively minor as compared to the long text, so dimensions of each text are minor and the vector space model is no longer sparse. D. Blei, et.al [9] and Yue Lu, et.al [10] reduce the dimensionality of text by reducing the number of the document into a number of topics by using some topic distribution parameter, and then applying some traditional clustering algorithm to classify the text. Q. Diao, et.al [11] examines the short text by taking that short text belongs to certain subjects. Gui-Rng Xue et.al [12] proposed an approach to separate the text data which covers the traditional PLSA algorithm to incorporate both types of data; labeled or unlabeled into a unified probabilistic model. D. Ramage et.al [13] designed an algorithm for multi-labeled classification by doing one to one mapping between topics and labels. Jinhee Park et.al [14] proposed a methodology to extract the important topic words from blog by measuring richness of the content in the blog. Paolo Ferragina et.al [15] used a powerful toll TAGME to extract the significance of full phrases for adding some brief explanation to the textbook to help with understanding of short text. Xiaojun Quan et.al [16] suggested a model which is founded on the similarity measurement method for selection of feature words based on the both lexical weight and relation of subjects to which words belongs.

Yang et.al [17] proposed an approach to classify the short text which is based on topic modeling approach and feature selection was done by the expected cross entropy. Chu, Zi, et.al [18] suggested a classification scheme which automatically detects any twitter account either human, but or cyborg, in which it uses three stage first entropy based component, second a spam detection component and the third account properties component. Zhang, et.al [19] proposed topic summarization approach to manage the different nature like short, dissimilar, noisy, etc. In this round robin algorithm used for generating template based summaries. Garber, et.al [20] proposed an approach for prediction of the crime using twitter, in which author analyzes the twitter specific linguistic analysis and statistical topic modeling approach employed to automatically identifying discussion. Guo, et.al [21] proposed an approach to mine hot topics from twitter stream, in which author used the frequent pattern stream mining algorithm (i.e. FP) to discover the hot topics from the twitter stream. Lkeda, et.al [22] analyze market using twitter analysis, in which author used a hybrid approach of text based and community based method for eliminating of the demographic of a twitter user.

One of the primary challenges of the above related work is that raising the feature by the search engine or Wikipedia and WordNet, doing this causes feature set to be too big which creeps the new problem of Curse of Dimensionality [23]. In this problem when data set became too heavy, it is very hard to treat as well as analyze, so there needed an efficient method which utilizes the optimal feature size. Too, subsequently applying the external knowledge, there are unimportant external features extracted which degrade the operation.

In conclusion, related work on the short text mainly focuses on the removal of data sparseness by incorporating the more knowledge along the short text using the Wikipedia or WordNet or background knowledge using the search engine like Google, Yahoo, Bing, etc. Although these techniques improves the accuracy somehow but lead to a novel problem of Curse of Dimensionality [23]. Which causes the algorithm slower and difficult to break down the information.

To overcome the above problem, In this paper, we proposed a novel approach to classify the incoming tweets into predefine category "News", "Deals" "Event", "Personal Message". In which semantic database is developed, the reason behind the provision of semantic data is that it will increase the semantic knowledge.

Experimental results indicate the improvement on the traditional method "BOW (Bag of Words)".

III. PROPOSED METHODOLOGY

Proposed methodology can be distinguished in four forms, in a first phase our feature selection algorithm is discussed and hybrid approach with the traditional algorithm. In phase 2 searching for the feature. In phase 3 feature reduction is done using logical operation and at the third and last phase classification is done for measuring the accuracy of the proposed methodology.

The complete process of our proposed methodology as follows:

1. Creating the semantic database for enhancing the semantic knowledge.
2. Applying N-feature algorithm for feature selections.
3. After Selecting feature selections reduce the features using logical operations
4. Train the machine using labeled data.

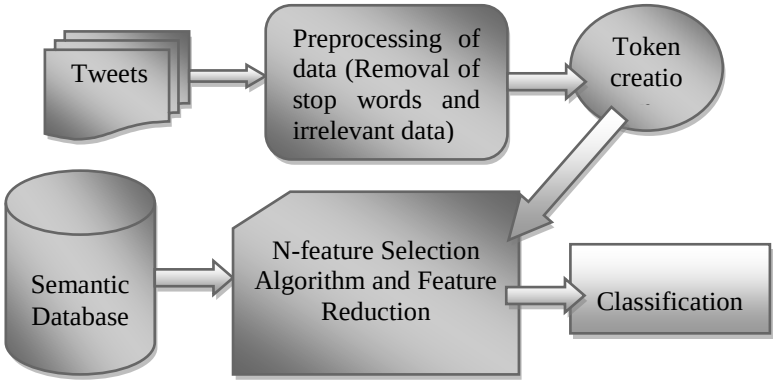

Figure 1 Classification process of feature selection algorithm

Phase 1: N-feature selection algorithm

In our approach we adopted a new feature selection algorithm which is simple and yet effective approach to extract the features from the entities. The N-feature algorithm is so efficient due to the binary feature extraction which make easier to learn using any machine learning language. For feature selection algorithm we built up a semantic database. The Reason behind the development of the semantic database is that it will increase the classification accuracy. We select the nine features and selection of these nine features is based on the detailed study of the social networking site Twitter. To classify the upcoming into predefine category, we select the four categories, namely 'News', 'Events', 'Personal message' and 'Deals'. The grounds behind the selecting these four category peoples has always invested their time for searching the right news and cases. Reason behind the selecting the personal message category because peoples always convey their emotions and feelings to the author by posting 'personal messages'. 'Deals' is selected because it contains best offer.

To classify the incoming tweet into some predefine category, we select nine features which further reduce to the seven features.

1. *Opinion describing words:* A database was created which contain the information about the opinion words.
2. *Abbreviation present in the sentences:* - People always used abbreviation to present their emotions so a database was created which contain the abbreviation.
3. *Time-Related Information and Event describing words:* - The database was so prepared that only describes the events. The words present in the database related to this feature was contextual only to the category event, thus almost eliminating the ambiguity of meaning of words present in the micro blogs.
4. *Presence of Personal Reference:* - A message meant for an individual often contains '@username'. Username is the user id of that particular individual and is forever unique and is not to be befuddled with the on screen name of that blogger.
5. *Presence of the Currency, deal related information:* - A semantic database that refers only to deals and offers was prepared.
6. *Presence of Emphasis:* - Users often repeat certain character of a word to show their aggression or opinion towards a particular incident or question. For example "Very Good" often used for appreciation.
7. *Presence of date and time information:* - An algorithm was developed to detect presence of time and date related information, the presence of time related info presence of Am or Pm information etc.
8. Aside from these features, two more features were pulled up, the features described the presence of timely information, i.e. before noon or after midday, date information present in DD/mm/yy or DD-mm-yy format. The features represent event related information in the text.

Phase 2: Searching for the features

The feature selection procedure commences with the searching algorithm, but before searching the features preprocessing of data is needed in which removal of stop words and irrelevant information from the entities necessary because stop words are those words which will not take part in the classification and the removal of those words causes the consumption of considerable time for sorting. The proposed algorithm is given below.

Input- i where i is the number of tweets
Output- feature_matrix(7, i)
7 - Number of features we adopted
i - Number of tweets
Step 1. Preprocessing of data (removal of stop words)
Step 2. Searching for the @ symbol
if @ symbol present
feature_matrix (1,i) = 1;
else
feature_matrix (1,i) = 0;
Step 3.Searching for event word present in tweets by searching for event semantic database we prepared
if event words present
feature_matrix (2,i) = 1;
else
feature_matrix (2,i) = 0;
Step 4. Searching for opening words present in tweets by searching for an opinion, semantic database we developed.
if opinion words present
feature_matrix (3,i) = 1;
else
feature_matrix (3,i) = 0;

Step 5.Searching for deal words present in tweets by searching for deal semantic database we prepared.
if deal words present
feature_matrix (4,i) = 1;
else
feature_matrix (4,i) = 0;
Step 6. Searching for opinion word present in tweets by searching for abbreviation semantic database we prepared.
if opinion words present
feature_matrix (5,i) = 1;
else
feature_matrix (5,i) = 0;
Step 7. Searching for emphasis word present in tweets by searching for emphasis semantic database we prepared.
if emphasis words present
feature_matrix (6,i) = 1;
else
feature_matrix (6,i) = 0;
Step 8. Searching for time and date information and feature reduction
if (fwd_s==1||das_date==1||am==1||pm==1)
word_features(7,i)=1;
else

word_features(7,i)=0;

In our approach we also used a hybrid of the BOW and N-feature selection algorithm. Our N-feature selection algorithm producing the binary results. Binarization helps in reducing the error caused by redundancy present in BOW feature. The experimental result showing our approach also increased the accuracy over the conventional BOW. The proposed algorithm is given below.

Input- i, where i is the number of tweets
Output- feature_matrix(7, i)
7 - Number of features we adopted
I - number of tweets
Step 1. Preprocessing of data (removal of stop words)
Step 2. Searching for the @ symbol
if @ symbol present
feature_matrix (1,i) = 1
else
feature_matrix (1,i) = feature_matrix (1,i);
Step 3. Searching for event word present in tweets by searching for event semantic database we prepared
% word_count- no of words present
if event words present
feature_matrix (2,i) = 1* word_count;
else
feature_matrix (2,i) = 0;
Step 4. Searching for opening words present in tweets by searching for opinion, semantic database we prepared.
if opinion words present
feature_matrix (3,i) = 1* word_count;
else
feature_matrix (3,i) = 0;
Step 5. Searching for deal words present in tweets by searching for deal semantic database we prepared.
if deal words present
feature_matrix (4,i) = 1* word_count;
else
feature_matrix (4,i) = 0;
Step 6. Searching for opinion word present in tweets by searching for abbreviation semantic database we prepared.
if opinion words present
feature_matrix (5,i) = 1* word_count;
else
feature_matrix (5,i) = 0;
Step 7. Searching for emphasis word present in tweets by searching for emphasis semantic database we prepared.
if emphasis words present
feature_matrix (6,i) = 1* word_count;
else
feature_matrix (6,i) = 0;
Step 8. Searching for time and date information and feature reduction
if (fwd_s==1||das_date==1||am==1||pm==1)
word_features(7,i)=1;
else
word_features(7,i)=0;

Phase 3: Reducing the features

There was total nine features extract from every incoming which was reduced to the seven features. Reduction of features causes reduction in time consumption for extracting the features. A logical operation OR is taken for the reduction of the features.
Bringing down the features always reduced the calculation time, complexity as well as the less storage, which is the real time problem with the current real time scenario. Features are cut because they post the same information about the events. Image 2 shows the performance of feature reduction.

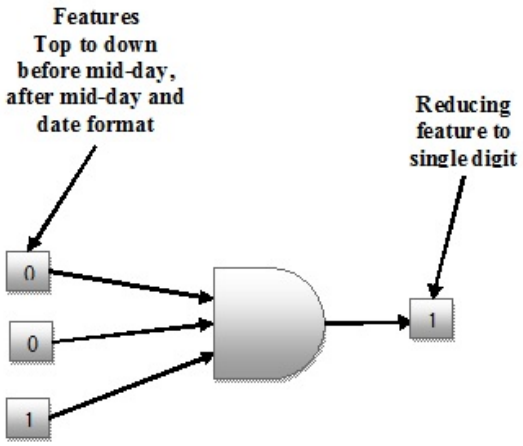

Figure 2: Logical operation for feature reduction

Phase 4: Classification

After the extraction of feature we get the feature matrix which contains the binary results so we can easily train the machine using labeled data. We use Support vector machine (SVM) for classification of binary class and Neural Network (N N) for classification of multi class

IV. EXPERIMENTAL RESULTS

Experimental setup

We developed and classified the recent tweets by human cognizance from the different users and eliminated those tweets which is not English and followed a more of a distorted sort of English often referred as Hinglish a distorted kind of English. After taking out the non-English tweets finally we remain with 5302 tweets. After taking out stop words and stemming process, we continue with the 7405 unique words.

All Experimental is done in core i3 processor with 4 GB ram in Matlab software.

Performance measurement and evaluation

Figure 3 represents the performance graph of neural network for proposed algorithm, which carries a minimum error of. 066282 when taking only 32 epoch or instances to fit the best classification, then the time complexity of the proposed methodology is also dejected. Even later on the 32nd epoch machine went on for 6 more epochs to check the strength of any outlet on the formulation.

Figure 4 represents the performance graph of the proposed hybrid algorithm, which contain the minimum error 0.067689 which is taking 60 epochs or instances.

These numbers do not show any major problems with the preparation. The validation and test curves are very similar. If the test curve had increased significantly before the

validation curve increased, then it is possible that some over fitting might have occurred.

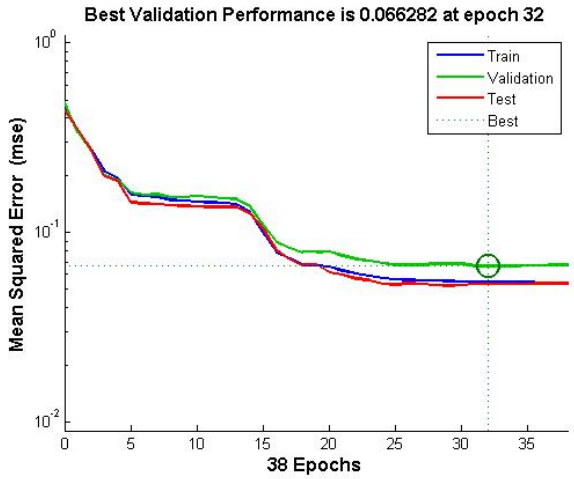

Figure 3: Performance graph of proposed algorithm

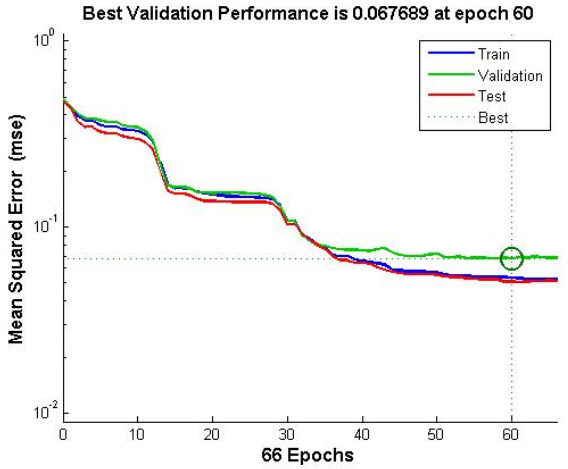

Figure 4: Performance graph of proposed hybrid algorithm

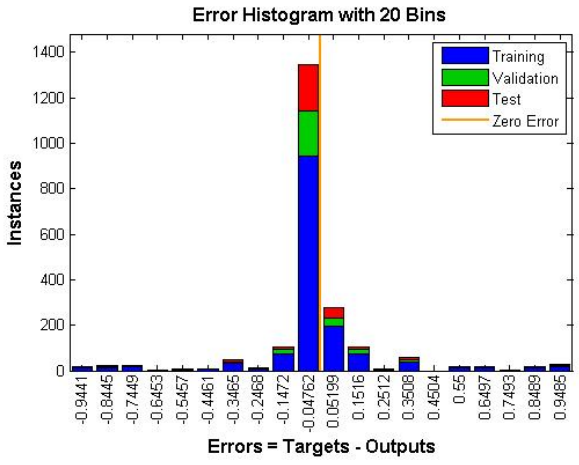

Figure 5: Error histogram of proposed algorithm

Figure 5 represents the error histogram of the proposed hybrid algorithm of neural network, training data represented by the blue bars, validation data represented by the green bars, and the red bars represent testing data. Indication of outliers can be recognized by the histogram. In this, most of the error fall close to the zero which makes proposed algorithm less affected by the outlaws. Most of the error fall between -0.35 to 0.35 and training, testing laying on the outer 0.9485.

Figure 6 represents the error histogram of the hybrid algorithm of neural network. In this, most of the error fall close to the zero which makes hybrid algorithm less affected by the outlaws. Most of the error fall between -0.25 to 0.25 and training, testing laying on the outlier 0.9496.

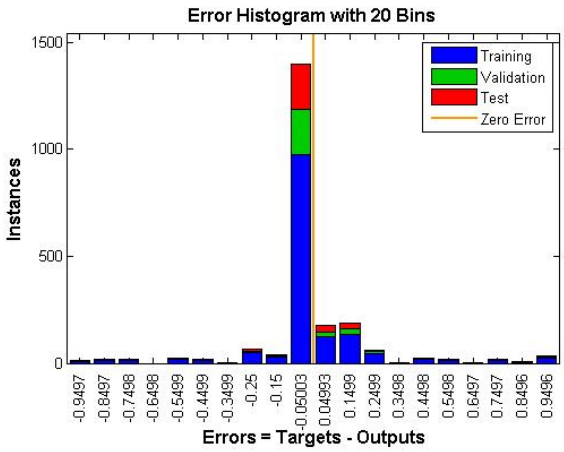

Figure 6: Error histogram of proposed hybrid algorithm

Figure 7 and figure 8 represent the confusion matrix of the proposed algorithm and proposed hybrid algorithm respectively, it shows that there are four confusion matrix, first one for the training, second one for the validation, third one for the test and last for the overall accuracy.

Figure 7: Confusion matrix of proposed algorithm

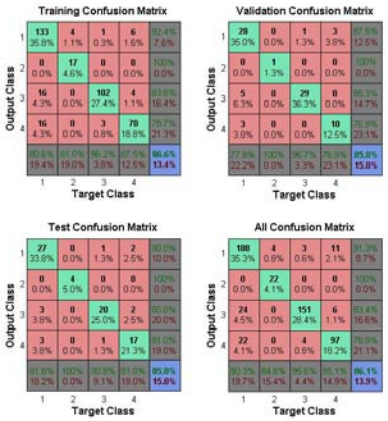

Figure 8: Confusion matrix of proposed hybrid algorithm

Figure 9 indicates that the algorithm achieves the accuracy of 14.5% over BOW and the proposed hybrid algorithm achieved the 14.9 % accuracy over BOW using Neural Network (NN). It indicates that the algorithm achieves the accuracy of 14.5% over BOW and the proposed hybrid algorithm achieved the 14.9 % accuracy over BOW using Neural Network.

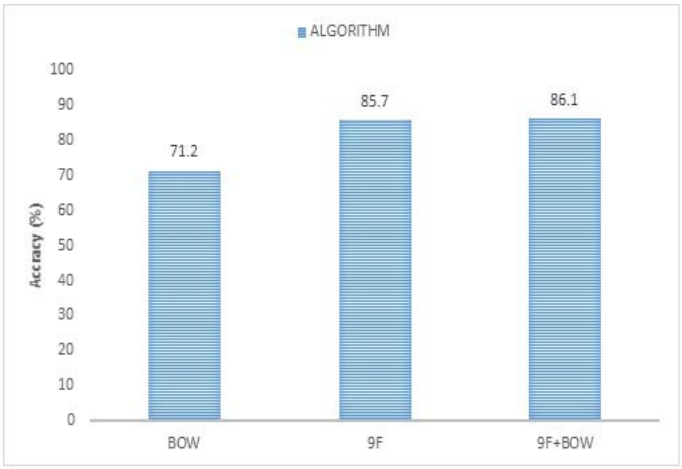

Figure 9: Accuracy of proposed algorithm using Neural Network (pop out in each bar)

Figure 10 represents the individual accuracy of all classes. It indicates that proposed algorithm have 16 % more accuracy for 'NEWS' class, 11.5% more accuracy for 'PERSONAL MESSAGE' class, 11.7 % more accuracy for 'DEAL' class and 8.1 % accuracy for class 'EVENT' over the BOW and proposed hybrid algorithm achieve 26.3 % more accuracy for 'NEWS' class, 21% more accuracy for 'PERSONAL MESSAGE' class, 6.9 % more accuracy for 'DEAL' class and 12.1% accuracy for class 'EVENT' over the BOW.

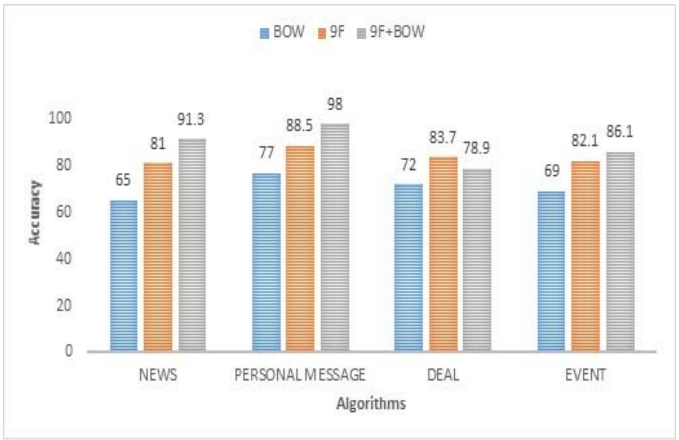

Figure 10: Individual accuracy of all classes using Neural Network (pop out in each bar)

Figure 11 represents the overall accuracy of the proposed algorithm 91 % using SVM, It shows that the algorithm achieves the accuracy of 21% over BOW and the proposed hybrid algorithm achieved the 24% accuracy over BOW using SVM.

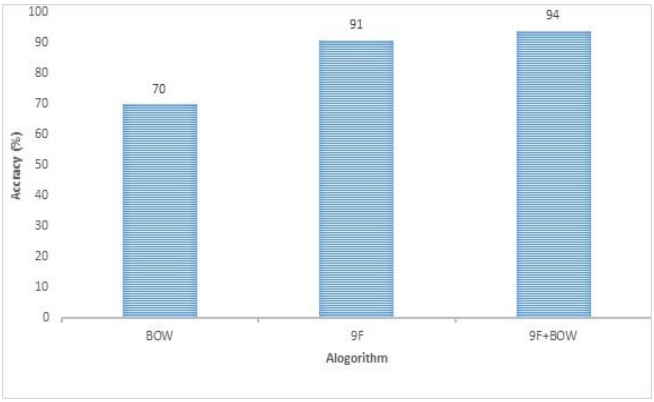

Figure 11: Accuracy of proposed algorithm using Support Vector Machine (pop out in each bar)

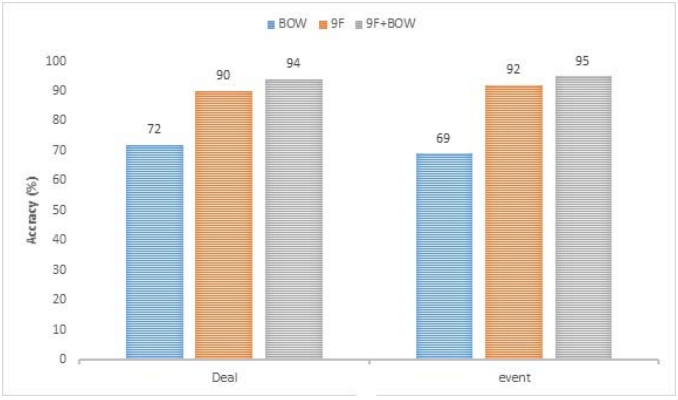

Figure 12: Individual accuracy of all classes using SVM (pop out in each bar)

Figure 12 represents the individual accuracy of all classes using SVM, here only two classes we used because SVM can better classify two class problem. It indicates that proposed algorithm have 18 % more accuracy for 'DEAL' class and 23% accuracy for class 'EVENT' over the BOW and proposed hybrid have 20% more accuracy for 'DEAL' class and 26% accuracy for class 'EVENT' over the BOW.

We believe that with the increment of the training data, we can increase the accuracy by consuming the initial time cost for the training.

V. CONCLUSION AND FUTURE WORK

We have proposed a novel approach to classify the incoming tweets into predefine class by using the semantic knowledge. By employing this kind of system user can categorize the data according to their interest which increase the information filtering it will really helpful for handheld devices.

In future we are planning to incorporate more semantic knowledge. By causing this we believe that accuracy will be incremental. Twitter is widely used for handheld devices so accuracy is the primary worry. Development of a toolbox of text classification presently not available in Matlab, we can work as a third party text classification plugin. GUI based operation, server based operation etc.

VI. REFERENCES

[1] X. -H. Phan, L. -M. Nguyen, and S. Horiguchi. “Learning to classify short and sparse text & web with hidden topics from large-scale data collections” In proceeding of the 17th international conference on World Wide Web, WWW '08, pp. 91-100. ACM, 2008.

[2] Sankaranarayanan, Jagan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. "Twitterstand: news in tweets." In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 42-51. ACM, 2009.

[3] Hong, Liangjie, and Brian D. Davison. "Empirical study of topic modeling in twitter." In Proceedings of the First Workshop on Social Media Analytics, pp. 80-88. ACM, 2010.

[4] Sriram, Bharath, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. "Short text classification in twitter to improve information filtering." In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 841-842. ACM, 2010.

[5] Banerjee, Somnath, Krishnan Ramanathan, and Ajay Gupta. "Clustering short texts using Wikipedia." In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 787-788. ACM, 2007.

[6] Hu, Xia, Nan Sun, Chao Zhang, and Tat-Seng Chua. "Exploiting internal and external semantics for the clustering of short texts using world knowledge." In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 919-928. ACM, 2009.

[7] Jin, Ou, Nathan N. Liu, Kai Zhao, Yong Yu, and Qiang Yang. "Transferring topical knowledge from auxiliary long texts for short text clustering." In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 775-784. ACM, 2011.

[8] Mengen Chen, Xiaoming Jin, and Dou Shen. “Short text classification improved by learning multi-granularity topics.” In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, pp. 1776-1781, 2011.

[9] D. Blei, A. Ng, M. Jordan and J. Lafferty. “Latent Dirichlet allocation”, The Journal of Machine Learning Research, Vol-3, pp. 993-1022, 2003.

[10] Yue Lu, Qiaozhu Mei, Chengxiang, and Zhai. “Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA”, Information Retrieval, vol-14, pp.178-203, 2011.

[11] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. “Finding bursty topics from microblogs”, In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Vol-1, pp. 536-544, 2012.

[12] Gui-Rong Xue, Wenyuan Dai, Qiang Yang, and Yong Yu. “Topic-bridged PLSA for cross-domain text classification”. In proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008.

[13] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. “Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 248-256, 2009.

[14] Park, Jinhee, Sungwoo Lee, Hye-Wuk Jung, and Jee-Hyong Lee. "Topic word selection for blogs by topic richness using web search result clustering." In Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication, pp. 80, ACM, 2012.

[15] Ferragina, Paolo, and Ugo Scaiella. "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)." In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1625-1628. ACM, 2010.

[16] Xiaojun Quan, Gang Liu, Zhi Lu, Xingliang Ni, and Liu Wenyin. “Short text similarity based on probabilistic topics” Knowledge and Information Systems, vol-25, pp. 473-491, 2010.

[17] Yang, Lili, Chunping Li, Qiang Ding, and Li Li. "Combining Lexical and Semantic Features for Short Text Classification." Procedia Computer Science, vol-22, pp. 78-86, 2013.

[18] Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?." Dependable and Secure Computing, IEEE Transactions, vol-09, pp. 811-824, 2012.

[19] Zhang, Renxian, Wenjie Li, Dehong Gao, and You Ouyang. "Automatic twitter topic summarization with speech acts." Audio, Speech, and Language Processing, IEEE Transactions, vol-21, pp. 649-658, 2013.

[20] Gerber, Matthew S. "Predicting crime using Twitter and kernel density estimation." Decision Support Systems, vol-06, pp. 115-125, 2014.

[21] Guo, Jing, Peng Zhang, and Li Guo. "Mining hot topics from twitter streams." Procedia Computer Science, vol-9, pp. 2008-2011, 2012.

[22] Ikeda, Kazushi, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. "Twitter user profiling based on text and community mining for market analysis." Knowledge-Based Systems, vol-51, pp. 35-47, 2013.