Retrieving Relevant Information From Tweet Posts in Twitter

Dr.Poonam Yadav
Assistant Professor, D.A.V College of Engineering & Technology, Kanina, Haryana 123027, India
poonam.y2002@gmail.com

***Abstract*—Now-a-days, users produce millions of short messages in social networks. Moreover, the retrieval of appropriate information to a particular event from the complete data is not an easy method. Hence, a framework is presented in this paper in order to retrieve the relevant information from Twitter posts for political debates. The main aim of this paper is a set of schemes in the retrieval process for connecting the user. Therefore, the proposed method attains an obviously higher precision using presenting important posts to be labeled. An external information source namely a simulated with an Oracle. Otherwise, domain expert is utilized to provide an accurate retrieval. An active retrieval method requests the labels, which aid enhances the retrieval precise the most, whereas the number of labeling requests is set to least. Finally, the simulation results demonstrate the efficiency of the selection schemes for involving the user in the retrieval process and the benefits of the proposed technique.**

Keywords*—Twitter; Labeled; Hashtags; Debates; Unsupervised method

I. INTRODUCTION

In recent years, tremendous growths of microblogs like Twitter create an enormous publicly accessible information sources. Now-a-days, Twitter has been utilized for political conversation. Moreover, Twitter has made a major impact on political debates and political processes. Besides, the social network has been greatly used by politicians, media, and the public in order to say their view [1] publicly. For example, the social network has been widely employed during the last few elections in the countries such as Canada and United States for live visual media events [2] [22].

The main aim of this paper is to link appropriate tweets to political debates, which have been conversed in the parliament. Generally, the politicians can advantages from this association because they need to follow precise topics and they require differentiating what other persons speak about their topics of interest. Hence, this process can be analysis information retrieval difficulty for microblogs, and the objective is to retrieve the pertinent tweets to every political topic [25] [26]. Nevertheless, the difficulty can also be also viewed as a classification problem. Here, each post we need to assign as a suitable debate. However, there are different problem arise, while working with tweets such as the incorrect spelling, pierced with acronyms, slang, and grammar [3]. Additionally, the main problem is the semantic relatedness of debates. On the other hand, for the most of the reported work on retrieval of tweets regard to distinguished kinds namely technology and economics [4], [5] that is usually easier.

The distribution of tweets over the debates is another challenge because those debates are imbalanced. Additionally, the overall volume of the data is the pertinent tweets to any debate and that are a tiny fraction. Here, the usage of old data to train a model is not relevant and infeasible to label large quantities of data. Moreover, we have assumed as the labeled data is not present to train a model. Hence, the unsupervised method is exploited as well as the idea of query expansion has employed to enhance the superiority of the outcome. Information retrieval is facilitated by means of an advanced compression approaches [22] [23] and modern data transmission technologies [20] [21] and exploited in many intelligence applications [18] [19] .

An automatically increasing the queries might append dissimilar terms to the query that might get worse the retrieval effects [7]. In general, the semi-supervised algorithm is better while comparing with the unsupervised method that needs a petite subset of the data to be labeled. Here, the Active Tweet Retrieval (ATR) is presented in order to enhance the results and to know the optimistic [24] effect of labeled data. It is a set of schemes which aspires an information source in order to choose the least number of instances to the label.

The main aim is to exploit particular features of Twitter to choose the relevant labeling request. Moreover, the main contribution of this paper is to propose a framework for information retrieval in order to relate the tweets. In addition, a set of schemes has been proposed to choose tweets by contemplating the network structure to enhance the precise of the retrieved tweets. Finally, the evaluation of the proposed technique and the effects of the chosen schemes on the results have investigated.

II. RELATED WORK

Numerous studies had focused in order to solve the classification and retrieval difficulties. In [9], to conquer the difficulty of vocabulary mismatch the query expansion methods was exploited to different approaches to improving the terms in the query. Recently, the external knowledge namely WordNet, Freebase and Wikipedia was employed for automatic techniques of query expansion have received good attention [10] [11]. In the topical classification, Chen et al.

[12] applied the semi-supervised technique, and Gupta et al. [13] reported about semi-supervised technique on the basis of the SVM rank for retrieved tweets. In choosing the cases for labeling requests, Hu et al. [15] demonstrate that the structure of micro-blog networks is significant. Imran et al. [16] presented AIDR system which attains labeled data during crowd-sourcing through disasters for identifying informative tweets. The information source is utilized to label, and the Naive Bayes classifier is trained and chosen most uncertain samples [17].

III. Proposed Unsupervised Framework

Fig 1 illustrates the framework for connecting tweets to the equivalent debates. Here, two important parts have been presented: (i) the extracting discriminative features have been present in the unsupervised retrieval (ii) retrieving tweets and the dotted lines represent the active retrieval part.

A. Feature Extraction

In a collection of documents, a user moves towards with a query search for the traditional information retrieval. The initial query is automatically formulated by the enormous number of tweets and the unreliable information requirements. Hence, a set of discriminative terms has been extracted from the transcripts on the basis of the tf-idf values. Moreover, retrieve tweets pertinent to the debates in D by exploiting the terms. Here, the retrieved tweets are employed to expand the list of the discriminative terms. Additionally, the Twitter features namely URLs, hashtags and user mentions are also extracted in the retrieved tweets. These features discriminative powers are not similar. Hence, $\pi_{i,j}$ is a matrix component comprises a non-zero value, and this is set based on the feature type. It denotes the feature i is chosen among the j discriminative features of the debate. Moreover, the hashtags are consistent indicator than the other features.

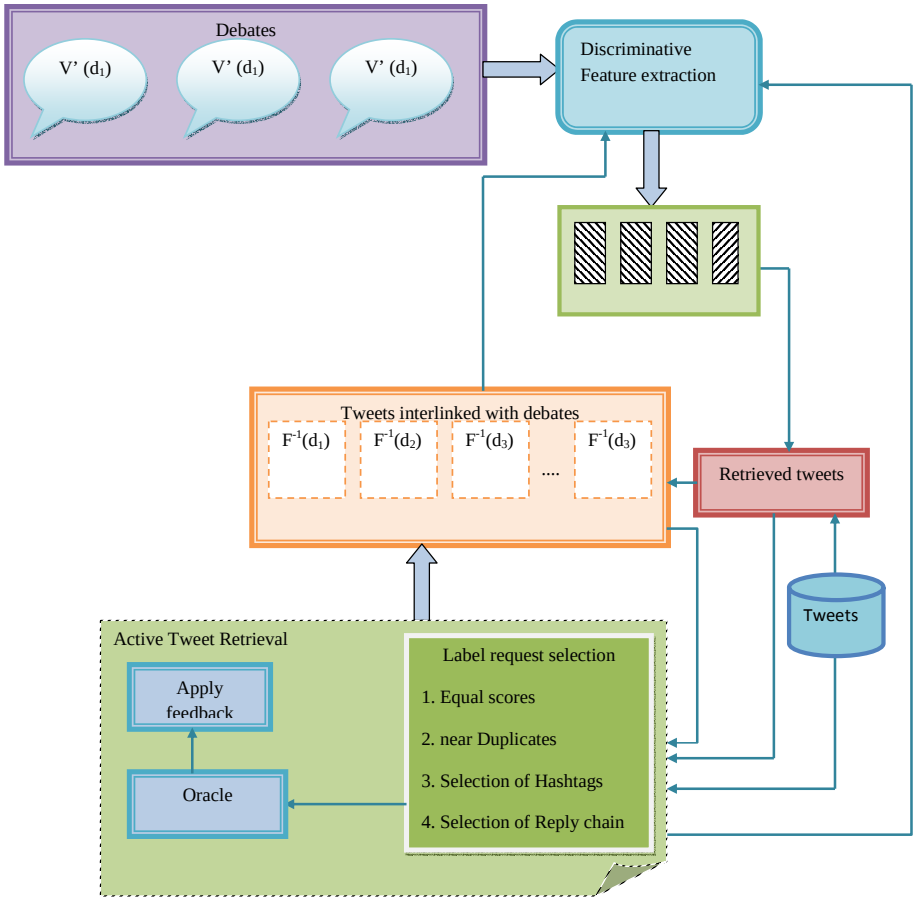

Fig. 1. Block diagram of proposed framework

B. Retrieval of tweets

A similar score for each tweet- debate pair has been calculating which is interlinked with the equivalent debates. Subsequently, the tweet has been assigned as $t_i^{'}$ with the maximum similarity score to the debate. The eq. (1), (2) and (3) presents the estimation score for the tweet $t_i^{'}$ and debate $d_j^{'}$

$$d_j^{'} = \left(\pi_{1j}, \pi_{2j}.....\pi_{sj}\right) \quad (1)$$

$$t_i^{'} = \left(\beta_{1i}, \beta_{2i},....\beta_{si}\right), q\varepsilon[1,....s], \beta_{qi} = \begin{cases} 1, k_q \varepsilon t_i^{'} \\ 0, k_q \notin t_i^{'} \end{cases} \quad (2)$$

$$sim\left(t_i^{'}, d_j^{'}\right) = d_j^{T^{'}} t_i^{'} \quad (3)$$

IV. Retrieval of Active TWEETS

The objective of enhancing the retrieval precise a set of four schemes have been proposed to engross an information source. Here, the labeling request, which has been employed, is a simplest type that is to offer a label of the debate. The number of labeling requests is minimized in order to choose the instances which aid in enhancing the outcomes of the association. The list of discriminative features is updated after receiving the feedback of the labeling request.

A. Similar Scores

Each tweet is assigned to the debate using eq. (3) with the maximum similarity score. The tweets have a similar maximum score with more than one debate in order to label the request. In assigning those tweets uncertainty may occur to the exact debates. Moreover, the information source offers the true association label, which is helpful for modifying the list of discriminative features which are automatically extracted to avoid assigning more tweets to the incorrect debates because of the key terms.

B. Duplication of tweets

Here, the efficient method namely LSH [8] is applied in order to minimize the intricacy of finding the near duplicating tweets. In addition, a near-duplicate candidates list is provided by this method. Hence, all pairwise comparisons do not require. Subsequently, the Jaccard similarity is exploited in the short list of candidates in order to find the near-duplicate tweets. If the Jaccard similarity > 0.9 means the two tweets are nearer duplicates subsequently, that tweet will add to the cluster.

C. Tweets Hashtags

In Twitter, the hash tags are important features and each, and every person in the Twitter utilizes these hashtags to spot their tweets. Hence, a set of discriminative hashtags has been extracted. In the unsupervised retrieval method, the extracted hash tags exploited to retrieve the tweets. In Fig 2, Flowchart of choosing hashtags relevant to debates to enhance recall and accuracy.

(a) Filtering process of stop hashtags: Among thousands of hashtags, which are a good candidate for enhancing the precise of the result to recognize the hashtags. At first, if a hashtag is discriminative sufficient to be a pointer of any detailed topic in the domain of our difficulty. A virtual manuscript is constructed to recognize the stop hashtags for each debate. Here, the retrieval algorithm is exploited to concatenate the tweets. Subsequently, the hashtags are ranked in decreasing order by the debate frequency and finally that are filtered.

(b) Enhancing accuracy by means of hashtags: Here, the wrong debates needs to be identified in tweets by utilizing the hashtags to enhance retrieval accuracy. Hence, two cases are stated the tweet is represented as $T^{'}$ and the hashtag is represented as $h_0^{'}$. In the first case, if the debate frequency df of $h_0^{'}$ is $df\left(h_0^{'}\right) > 1$ then using eq. (4), the normalized entropy is computed. Subsequently, the rank is allotted in decreasing order, and the hashtags with the higher order are assigned as good candidates. In the second case, $df\left(h_0^{'}\right) = 1$ it represents that all of the tweets which are presents in the $T'(h_0^{'})$ are retrieved correctly otherwise all of them are wrong. Here, the virtual manuscript is built by concatenating all of the textual which are presents in the tweets in order to identify the cases.

$$\gamma(F'(T'(h_0^{'}))) = \frac{\sum_{j=1}^{m} -F^{'}(T'(h_0^{'}))(d_j)\log F'(T'(h_0^{'}))((d_j))}{\log\left(F'\left(T'\left(h_0^{'}\right)\right)\right)} \quad (4)$$

Here, a profile is constructed in order to calculate the similarity score among the virtual manuscript. Subsequently, the Jaccard similarity is computed by means of the binary weights in the vectors.

(c) Enhancing recall by means of hashtags: In D, to enhance the retrieval recall the hashtags which are good indicators have been exploited. Moreover, the best way to choose the most recurrent hashtags of non-retrieved tweets happening in the pool. Nevertheless, this is not a good indicator of similarity.

(d) Interaction with the information source: Here, we need to arbitrarily choose a small number in order to request a label for the tweets, for instance, 3 is assigned for $T'(h_0^{'})$. In the labeling request if the information source allocates all of the chosen tweets to one debate, subsequently this debate is interlinked with all of $T'(h_0^{'})$.

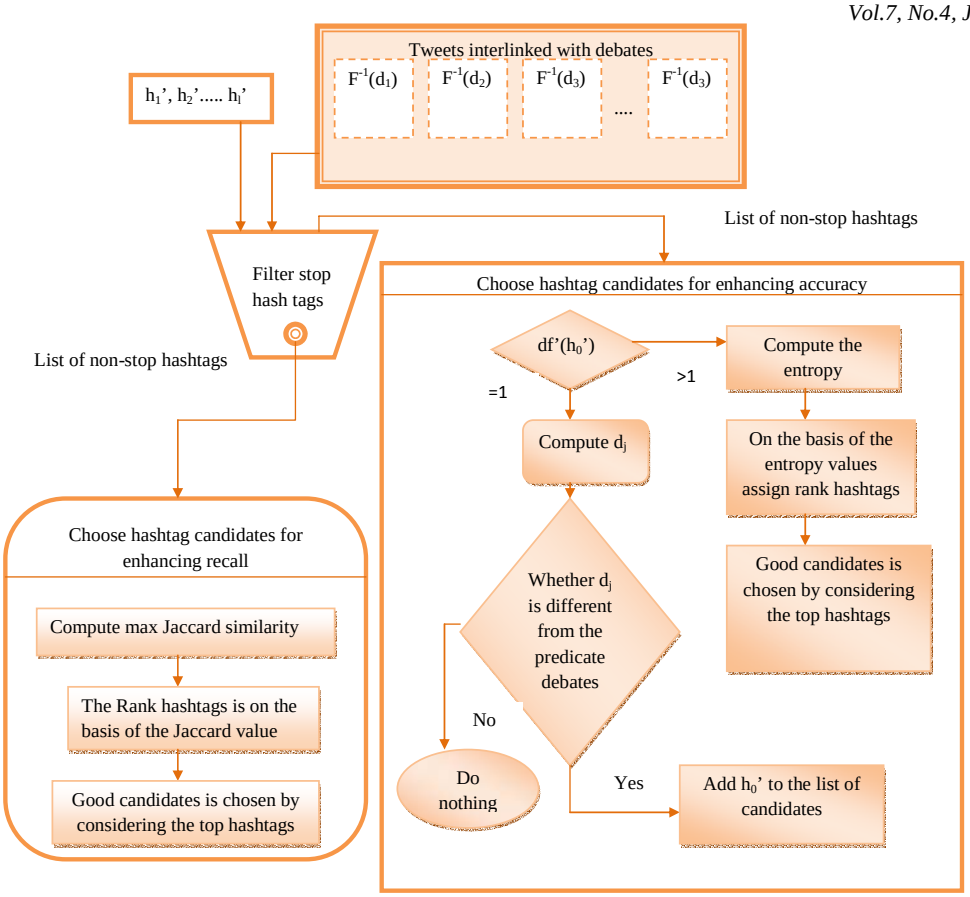

Fig. 2. Flowchart of choosing hashtags relevant to debates to enhance recall and accuracy

D. Interface with the information Tags

In order to reply to other user's tweets, Twitter has a structural feature which enables the users. One of the selection schemes is utilizing the relational information among tweets. Here, the reply back tweets tracing to the sources is the first step. Hence, a reply chain is constructed for every tweet like $t_i^{''}$ is represented as the reply of $t_i^{''}$.

Let us consider, the tweets in $R'(t_i')$, which are already interlinked with the debates in D. As for eq. (4), an entropy value is computed utilizing the same method. In $R'(t_i')$, if all the tweets are interlinked with the similar debate means subsequently, the entropy value is attained as zero. Otherwise, the entropy value is one if they have split uniformly between all the debates. Hence, the good candidates are represented as the source tweets, which have maximum entropy in the replay chains as there is a vital discrepancy among the replies. As a result, some reply tweets may interlink with the wrong debate.

In the selection process, the size of the reply chain is multiplied with the entropy value of every reply chain with a value. Hence, by the value in minimizing order, the source tweets are ranked. For labeling requests, the top tweets are chosen as candidates. Consequently, the label of the chosen source tweets is requested. In the reply chains, all tweets are interlinked such as those, which are non-retrieved but as their source tweets with the similar debate. Finally, this scheme aids to enhance both the recall and accuracy but exact replies with the incorrect assignments and interlinked tweets which are not retrieved.

V. Result and Discussion

In this experiment, the dataset tweets have been collected from the Twitter stream. This dataset comprises of 41,447 tweets, and original tweets are 16,297. Here, the re-tweets are not considered. Moreover, the tweets of one of the parliament member have been collected, and in Table II the 11 of the longest debates are discussed. In Table I, the estimate measures of this experiment have been presented. This Table depicts the choosing scheme on the basis of the hashtag, which is efficient. In addition, Table I demonstrates the unsupervised retrieval technique metrics measures and the ATR technique for the selection schemes.

Results in Table II indicate that the selection strategy based on the hashtags is the most effective one. Table II illustrates the performance of the retrieval methods for different debates and the accuracy as well as recall has been evaluated. Moreover, the number of relevant tweets and the subset of labeled tweets have considered for estimate the metrics such as recall and accuracy is revealed in Table II for

every debate. This table depicts for most debates the efficiency of the active retrieval for the proposed method enhancing the recall and the precision, whereas for some debates only one measure is enhanced considerably by the other method. In Meat inspection, the unsupervised technique is near to perfect, and the enhancement produced by means of active retrieval is not as important as in those debates. In Marine Mammal Regulations, the supervised retrieval is not better while comparing with others.

In Fig. 3 the precise value for unsupervised and ATR techniques on the key terms have been demonstrated. Here, the performance of the unsupervised method affects by choice of parameters. Moreover, the ATR method enhances the outcomes of the unsupervised retrieval. In addition, the ATR results after doubling demonstrate the number of labeling requests. Fig 4 demonstrates the outcome of all the schemes as given in Table I. Here; the smaller range is obtained for the similar score because only 20 tweets have an equal score. Moreover, the results depict the combination of four schemes is better than the applications of any of the schemes independently.

TABLE I. METRICS MEASURES FOR THE PROPOSED SELECTION SCHEMES.

Methods	Macro-Precision	Precision	Macro-Recall	Correctly Retrieved	Total request
Unsupervised- Debate	0.74	0.60	0.51	2169	0
Unsupervised- tweet and Debate	0.75	0.79	0.68	2875	0
Random Labeling Requests (1)	0.75	0.80	0.69	2926	100
Random Labeling Requests(2)	0.76	0.84	0.70	3049	100
ATR+ (1) + (2) + (3)	0.83	0.90	0.85	3259	100
ATR+ (1) + (2)	0.82	0.89	0.78	3222	60
ATR+(1)	0.80	0.83	0.73	3026	24
ATR	0.79	0.82	0.71	2991	15

TABLE II. ATR AND UNSUPERVISED TECHNIQUES FOR EACH DEBATE RETRIEVAL

Debate	ATR		Unsupervised Retrieval		#Tweets
	Recall	accuracy	Recall	Accuracy	
Fair Elections	0.93	0.97	0.65	0.98	670
Meat Inspection	0.99	0.98	0.96	0.98	536
Employment	0.93	0.93	0.63	0.98	365
Aboriginal Affairs	0.79	0.90	0.79	0.90	394
Veterans Affairs	0.62	0.96	0.56	0.99	125
Kidnapping Girls	0.85	0.95	0.78	0.89	122
CBC	0.97	0.78	0.97	0.78	101
Marine Mammal	0.75	0.83	0.51	0.67	73
Local Food	0.75	0.27	0.37	0.09	8
Non-Relevant	0.90	0.84	0.89	0.65	1209
Palliative	0.78	0.64	0.67	0.16	9
Housing	0.95	0.91	0.36	0.89	22

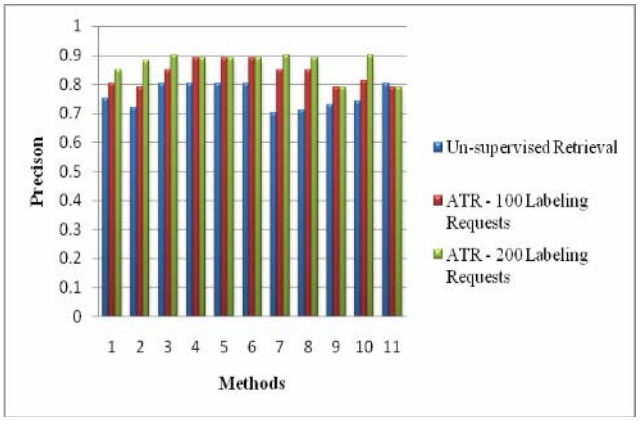

Fig. 3. Graphical representation of Precision before and after applying of ATR schemes

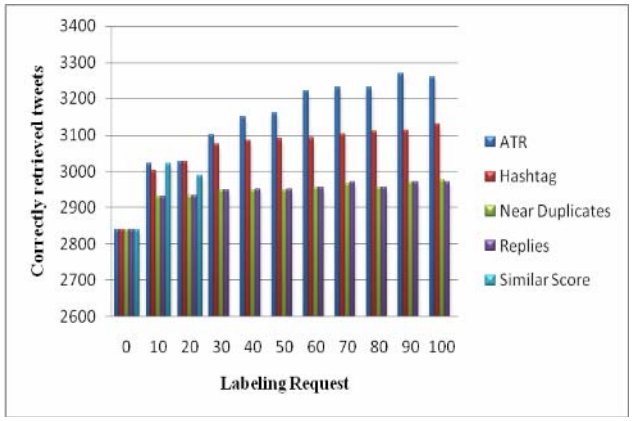

Fig. 4. Graphical representation of correctly retrieved tweets

VI. CONCLUSION

In this paper, a framework is presented for an active retrieval of tweet for real-time political debates. Here, the

proposed technique is utilized to extract key terms from the domain of interest. Moreover, the particular features of Twitter are employed to choose the labeling request by the proposed schemes. Besides, four schemes have been proposed such as recognizing near-duplicate tweets, hashtag, tweets with the equal score and inconsistent reply chain tweets related to the debates. The experimental results demonstrate that the proposed schemes are useful in order to choose labeling requests. Moreover, the ATR is a better method to choose the tweets, and it is a precise retrieval, whereas aspiring at keeping the number of user interferences to the least amount.

REFERENCES

[1] A. Gruzd and J. Roy, "Investigating political polarization on twitter: Acanadian perspective," Communications of the ACM, vol. 6, no. 1, pp. 28–45, 2014.

[2] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweetgeist: Can the twitter timeline reveal the structure of broadcast events?" in CSCW'10,2010, pp. 589–593.

[3] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical clustering of tweets," Proceedings of the ACM SIGIR: SWSM, 2011.

[4] K. Lee, D. Palsetia, R. Narayanana, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in ICDMW'11, 2011, pp. 251–258.

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in ACM SIGIR, 2010, pp. 841–842.

[6] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in ACM SIGIR, 1996, pp. 4–11.

[7] T. Miyanishi, K. Seki, and K. Uehara, "Improving pseudo-relevance feedback via tweet selection," in CIKM'13, 2013, pp. 439–448.

[8] M. Slaney and M. Casey, "Locality-sensitive hashing for finding nearest neighbors [lecture notes]," Signal Processing Magazine, IEEE, vol. 25, no. 2, pp. 128–131, 2008.

[9] D. F. Gurini and F. Gasparetti, "Trec microblog 2012 track: Real-time algorithm for microblog ranking systems," DTIC Document, Tech. Rep., 2012.

[10] R. Qiang, F. Fan, C. Lv, and J. Yang, "Knowledge-based queryexpansion in real-time microblog search," arXiv:1503.03961, 2015.

[11] W. Lucia and E. Ferrari, "Egocentric: Ego networks for knowledgebased short text classification," in CIKM'14, 2014, pp. 1079–1088.

[12] Y. Chen, Z. Li, L. Nie, X. Hu, X. Wang, T.-s. Chua, and X. Zhang, "A semi-supervised bayesian network model for microblog topic classification." in COLING, 2012, pp. 561–576.

[13] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: Realtime credibility assessment of content on twitter," in Social Informatics, 2014, pp. 228–243.

[14] T. Jaakkola and H. T. Siegelmann, "Active information retrieval," in NIPS 2001, 2001, pp. 777–784.

[15] X. Hu, J. Tang, H. Gao, and H. Liu, "Actnet: Active learning for networked texts in microblogging." in SDM, 2013, pp. 306–314.

[16] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "Aidr: Artificial intelligence for disaster response," in WWW'14, 2014, pp. 159–162.

[17] M.-H. Peetz, D. Spina, J. Gonzalo, and M. De Rijke, "Towards an active learning system for company name disambiguation in microblog streams," in CLEF 2013 Eval. Labs and Workshop Online Working

[18] K. S. S. Rao Yarrapragada and B. Bala Krishna, "Impact of tamanu oil-diesel blend on combustion, performance and emissions of diesel engine and its prediction methodology",Journal of the Brazilian Society of Mechanical Sciences and Engineering, pp. 1-15,2016.

[19] Rao Yerrapragada. K. S.S, S.N.Ch. Dattu .V, Dr. B. Balakrishna,"Survey of Uniformity of Pressure Profile in Wind Tunnel by Using Hot Wire Annometer Systems",International Journal of Engineering Research and Applications, vol.4(3), pp. 290-299, 2014.

[20] Kavita Bhatnagar and S. C. Gupta, "Investigating and Modeling the Effect of Laser Intensity and Nonlinear Regime of the Fiber on the Optical Link", Journal of Optical Communications, 2016.

[21] K Bhatnagar and SC Gupta, "Extending the Neural Model to Study the Impact of Effective Area of Optical Fiber on Laser Intensity", International Journal of Intelligent Engineering and Systems, vol.10, 2017.

[22] B.S. Sunil Kumar, A.S. Manjunath, S. Christopher, "Improved entropy encoding for high efficient video coding standard", Alexandria Engineering Journal, In press, corrected proof, November 2016.

[23] BSS Kumar, AS Manjunath, S Christopher," Improvisation in HEVC Performance by Weighted Entropy Encoding Technique" Data Engineering and Intelligent Computing, 2018.

[24] S Chander, P Vijaya, P Dhyani,"Fractional lion algorithm–an optimization algorithm for data clustering" ,Journal of computer science, 2016.

[25] P. Vijaya, Satish Chander,"Fuzzy Integrated Extended Nearest Neighbour Classification Algorithm for Web Page Retrieval", Proceeding of the International Conclave on Innovations in Engineering and Management, 2016.

[26] P Yadav and RP Singh, "An Ontology-Based Intelligent Information Retrieval Method For Document Retrieval", International Journal of Engineering Science, 2012.