Volume 4, Issue 9, September 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Pattern base Selection of Opponent in Contentious Articles Using Text Mining

Harish Patil
M.Tech. (CSE)
ICOT, Bhopal, India

Pinki Jain
M.Tech. (CSE)
ICOT, Bhopal, India

Jay Prakash Maurya
A.P. (CSE)
ICOT, Bhopal, India

Abstract— As the increase in the text documents in various fields extracting knowledge from the number of files is very important. This paper focus on identifying the opponent in the text document based on the pattern developed by the opponent. Pattern mining is always better then the term base mining. Here no need of external information required by the system, so a complete independent system that is develop that can analyze speakers then differentiate them into two group for opponent party. Finally identify the article favoring portion, means article is of which side. Results shows that process is excellent as compare to previous work, where external information is required.

Index Terms- Document analysis, Identifying Opponent, Pattern mining, Text mining.

I. INTRODUCTION

With the increase in the digital text document a new field of Mining is emerged where the process of fetching knowledge from the unorganized data is done. As the association is develop in the words on the different criteria depend on the requirement. Text mining is not similar in fashion that the possible outcome is already known, so without having the exact outcome from the data it predict the knowledge from it, which is base on the rules or step perform by the process of mining. So information which is of no use in the processing is removed from the document.

Text mining is nothing but a procedure to take out the important facts from the text documents. Selection of data for

Text mining is always the point of discussion. We can understand this point with an illustration that the process of

Text categorization comes under text mining because of its analogies but few of them think that text categorization should not link through text mining.

So, text mining merges various fields such as text retrieval, clustering, etc. It drastically reduces human efforts where human can do lots of mistakes and only expert can do a particular kind of work [3]. The two basic approaches are term base text mining and other is pattern base text mining. Depend on the requirement different technique is use the researchers.

As the journalism is the important part of the social worlds, which cover many contentious issues [10, 11]. There are many field where these contentious news articles come like as politics, environment, companies, stakeholders, civic groups, experts, commentators, etc. as articles compose of various speakers which act as different opponent and their argument act as the part of the articles. As the normal person who is unaware of the issues running in the article is not able to get understand of the contention issues. So a system should be developing for this so that direct understanding of the issues is present and opponents will be identifying easily.

By focusing above requirement this work will identify the main and sub opponents of the party, in order to develop such a system a relation need to be maintain between the opponents which can be analyzed by the article content. As the news makers start with the collection of the opponents then arrange all content accordingly [1]. This is done as per the readers mentality can be understand as the reader want to read the content keeping in mind that who are the opponent in the article what are their statements correspondingly. By writing thing in this manner or method without having much knowledge of the matter in prior reader can understand article easily. This can be said as the unsupervised understanding of the matter; it continuously identifies the opponents in separate parties and classified the article in the favoring party.

This work is different from the work that includes reading of sentiments which are base on the particular type of issue. So in these types of work they are working on the topic oriented view and differentiate matter on the sentiments. It provides opinion about the opponents on the basis of positive v/s negative ones.

Although contentious articles are very rarely differentiate on the above basis, so this paper work on the various topics which required separation of the articles speakers [12]. Some of the example of the different field such as in politics if X party wants to start a moment against Y party for different reasons. One more example of health care where insurance cover for different kind of disease done by the companies.

II. RELATED WORK

Different kind of text arrangement is done in previous work of text mining. Out of these one of the most common variety of text arrangement are bag of words here works are treat as the term and act as an element within the bag of words. The more important feature use in [12], is term frequency inverse document frequency, this act as the term weight for the

document feature. One more feature use in [4], is inverse document frequency, this act as the term weight for the document feature as well. By using this different feature value one reduce number of text for analysis in bag of words.

With the use of synonymy in [7] it is easy to develop a relation between the selected words. On the basis of these synonymy pattern evolution technique is develop in [5] which increase the performance of the text mining as compare to term base mining. Then sequence base pattern mining is evolved in the mining where these sequences are term as the phrases [2].

The sentiment characterization of contentious new articles is different from the sentiment studies. The discrimination of opponent is not done by the answer given by the opponent to the same topic but the facts deliver by it for the topic [3].

In [10] different combination of text retrieval techniques have use for the same, where a method is develop between them for their study. By implementing the use of supervised learning characterization process improves a lot in [7] a marking tool has been developed where projects are evaluate and score.

In [8] a decision support system is developed for them where projects are selected on the basis of the system outcomes. Similarly in [4] a fuzzy logic technique is used for project selection.

This paper uses the relation between the narrators of the different parties which has been use by various other paper as in [17]. But there approach was not unsupervised as develop in this work. There they required initial training for learning the pattern of the articles inserted But here no need of any kind of training module required.

III. PROPOSED WORK

Whole work experience is explained by the block diagram shown in figure 2 showing different steps, while explanation of each is done in below heads.

Preprocessing

Preprocessing is a process used for conversion of document into feature vector. Just like text categorizations the preprocessing also has controversy about its division. The preprocessing is divided into two parts – text preprocessing and document indexing [10].

Text preprocessing is consisting of words which are responsible for lowering the performance of learning models. Those words which are responsible for lowering the performance of learning model are called "noisy" words [6].

The words which are misspelled or the abbreviated word or common word {i.e. \is",\or",\ a"} are taken into as noise words. The word does not hold any info which is useful for classification.

Pruning

In machine learning pruning means removal of undesired features from feature space. For text mining pruning is the most essential part because maximum words of text corpus are low frequency words [13]. If we follow Zipf's law then we discover that in a corpus of natural language text the rank of word is provided on the basis of frequencies. Classification of word frequency is an inverse power law having exponent nearly one [10].

This signifies that in maximum time words in corpus arises very few times in any training corpus. Those words which are arise very few times statically unimportant having low information gain. However the occurrence of any word in training in future document is very less. When categorization is done pruning mostly produces feature space of small size.

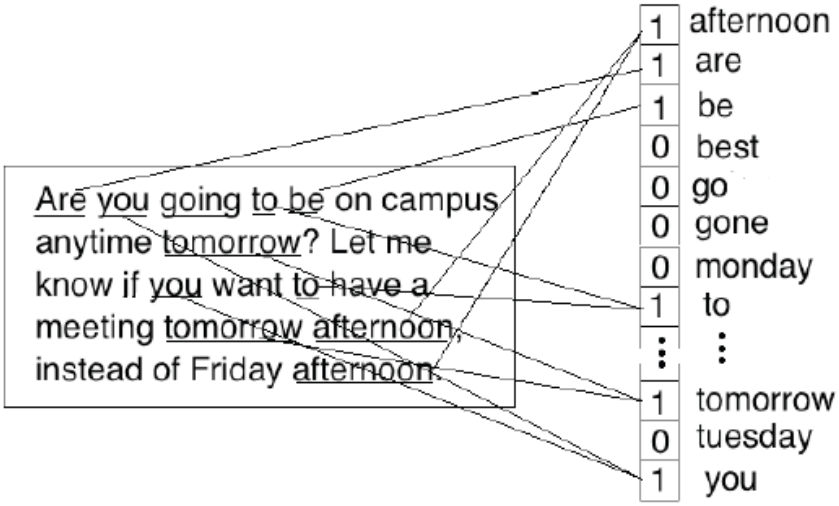

Fig.1 Example of the bag-of-word text representation with the occurrence of word as feature value

Article is the collection of sentences and each sentence is collection of words. So for the better understanding of the article it is dividing into sentences. Then for each sentence is process so getting effective results.

Sentence Separation:

On the basis of two different technique of sentence separation from the article it is done. First is to find the full stop '.' In the document other is the capital letter after the full stop. This can be understood as the "abc pqr. Mno xyz." So sentences are "abc pqr." and "Mno xyz".

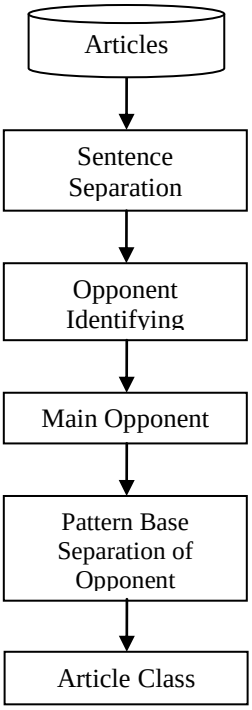

Fig. 2 Block diagram of Text Mining processing

Identifying Opponent:
Once sentences are obtain, now from it stop words are removed which is obtained by the dictionary of that language. As the assumption is done that name of the opponent is not as same as the words present in the dictionary [10, 11].
So, one bag of words which is explain in fig. 1 where important words are collect and maintain. In order to understand it bitterly read the below sentence.
"Rahul was a good cricketer of country"
From above sentence words as {was, a, good, cricketer, of country} were present in the dictionary but "Rahul" is not present so this word in the bag. In the similar fashion rest of sentences are processed. It is possible that sentences contain same name more than once so a counter is also maintain for the same. Suppose "Rahul" word repeat in the article ten times
Then for it act as a frequency of the word with their importance in the article.
One important issue cover in this opponent identification part is that few sentences contain opponent names with their surname then it is consider as the single opponent.

Main Opponent
This stage find the main opponent of the group means speakers who give more and more statements are delivered [12]. Here only purpose is to search the main speakers of the party for this entire one should find the most frequent speaker of the article which is obtain by the bag of words vector. As it maintain the list of opponent and there frequency that how many time that opponent appear in the article. So by this one can easily find the main opponent of the parties. This can be understood as let the BOW= {'Rahul', 'Sachin', 'Sourhab'……..} there frequency is {5, 4, 2……} then top frequency scorer is considering as the main opponent of the article.

Pattern Base Separation of Opponent
In order to develop the relation between the other opponents with the other main opponents. It means separation of the opponent with other is done by finding a pattern of the three important factors that is to read a sentence and follow below steps:
a). Find any of main opponent in the sentence.
b). Find any of other opponent in the sentence.
c). Find the Prons words use in the sentence.
d). Find the Cons words use in the sentence.

Now if the sentence contains main opponent and other opponent then count the prons words number and cons words number in the sentence [12]. Now if Pros > cons then depend on this pattern main opponent is in favor of the other opponent. Similarly if Cons > prons then depend on this pattern main opponent is in oppose of the other opponent. It don't matter that who is the speaker of the sentence.
This can be understand as let us say "Saurhab was scolding Rahul very badly " in this sentence the speakers are 'Saurhab' and 'Rahul' both prons words in the sentence is zero while cons word is scold so above condition as Cons > prons then depend on this pattern Saurhab is in opposing party of Rahul.

Article Class

This is final stage of the work where it is conclude that either article is of first party class or of second party class. So in order to decide this it is find that article contain sentence and how many are of each side depend on the number of sentence present on each side it is declared as that article belong to which class [8].

Given an article A, and the two sides B and C,

$$A->B=\frac{Q_b+S_b}{S_u}>=\frac{Q_{bc}*\alpha+S_{bc}*\beta}{S_u}$$

$$A->C=\frac{Q_C+S_b}{S_u}>=\frac{Q_{bc}*\alpha+S_{bc}*\beta}{S_u}$$

where

S_U : Number of all sentences of the article

Q_b : Number of sentences from the side B.

Q_{bc} : Number of sentences from either side B or C.

S_b : Number of sentences classified to B.

S_{bc} :: Number of sentences classified to either B or C.

α, β :serves as a threshold for the ratio of sentences.

Proposed Algorithm: Opponent and Document Classification

1. S←Sentence_Seperation(A) // S: Sentence Matrix
2. BOW←Identifying_disputant(S) // Collect disputant from the sentences and there count.
3. BOW ←Sort_decend(BOW) // Arrange in decreasing order of the count present
4. M←Top_two(BOW) //M Contain two main opponent
5. Loop d= 1:BOW-2 // For each other disputant
6. Loop s = 1:S
7. If contain_disputant(S, M, D)
8. P←count_pros(S)
9. C← count_cron(S)
10. If P>C
11. Class←{M, D}
12. Otherwise
13. Class←{M', D}
14. Endif
15. Endif
16. EndLoop
17. EndLoop

IV. EXPERIMENT & RESULT

This section presents the experimental evaluation of the proposed work. All algorithms and utility measures were implemented using the MATLAB tool. The tests were performed on a 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional.

Dataset:

In order to evaluate proposed work articles of different field has been used one is from the Politics, other is from country issues. Although some article are not of similar topic but on same theme. Table 1 contains the list of articles use for different category.

Table1 represent the Document set wise actual separation

	First Party	Second Party	Total
Set1	3	4	7
Set2	3	6	9

Evaluation Parameter

In order to evaluate results there are many parameter such as accuracy, precession, recall, F-score, etc. Obtaining values can be put in the mention parameter formula to get better results.

$$\text{Precision} = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$\text{Recall} = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

$$F\_Score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

In above true positive value is obtain by the system when the input article is in favor of B opponent and system also says that article is in favor of the B opponent. While in case of false positive value is obtain by the system when the input article is in favor of B opponent and system says that article is in favor of the C opponent.

Results:

There is article classification done on the basis on the disputant's relationship with other disputants. As mentions in D part of the paper.

Table2 represent the Document set wise proposed work separation

Articles in Favor		
	First Party	Second Party
Set1	2	5
Set2	4	5

Table3 represent the Results of first Party of set wise.

First Party			
	Precision	Recall	F-Measure
Set1	0.66	0.29	0.402
Set2	0.75	0.44	0.556

Table4. Represent the Results of Second Party of set wise.

Second Party			
	Precision	Recall	F-Measure
Set1	1	0.5	0.599
Set2	0.857	0.75	0.806

As table 2, 3 & 4 results shows that by the use of proper threshold of the disputant selection and dictionary it is possible to have values of precision above 0.75 which is quite good progress done by the proposed algorithm as compare to the previous work in [8], where most of the values are below the average of the results obtained. It is depend on the different reviewers and article that result may vary.

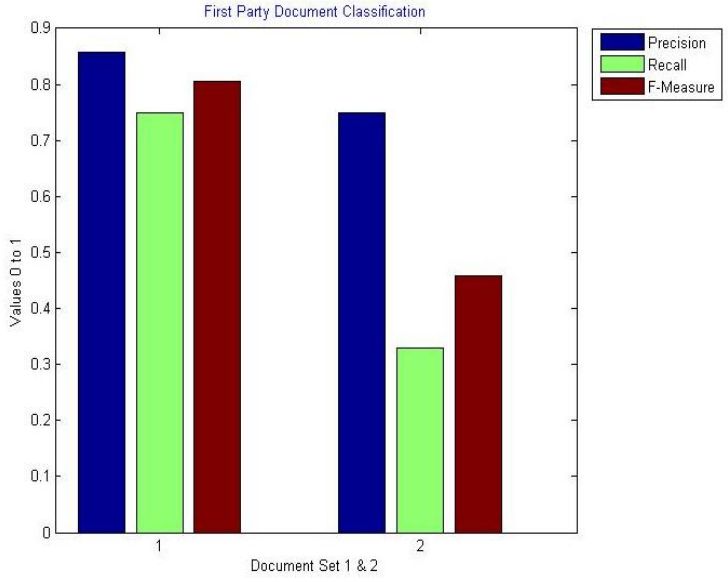

Fig. 3 Document Classification in First Party

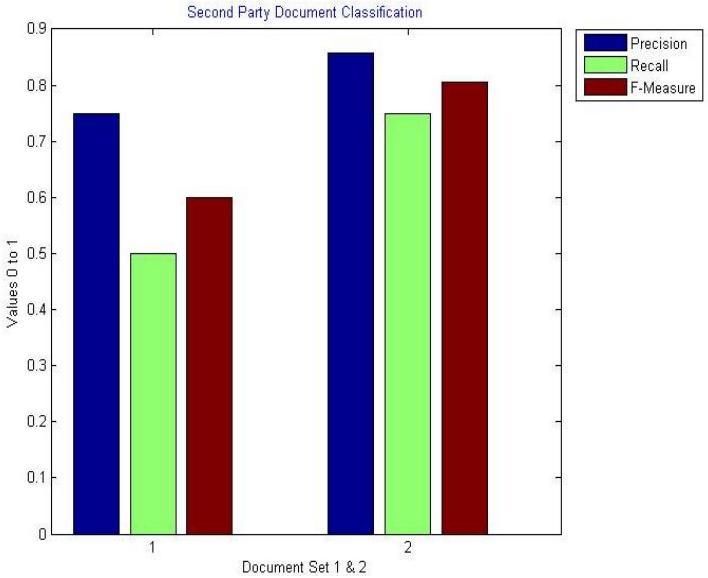

Fig. 4 Document Classification in Second Party

Figure 3 & 4 represent the Document classification of the two document sets on the basis of the evaluation parameter, which shows that in both the sets values obtain after the testing is quite impressive and acceptable as compared to the [8] where manual background knowledge need to be feed in the system.

Above results shows that as the use of proper threshold of the opponent selection and dictionary it is possible to have values of precision above 0.8 which is quite good progress done by the proposed algorithm as compare to the previous work in [8], where most of the values are below the average of the results obtained. It is depend on the different reviewers and article that result may vary. Graphs shows that separation of document is also quit impressive in both the sets for first party and second party

V. CONCLUSION

With the drastic increase of the digital text data on the servers, libraries it is important for researcher to work on it. Considering this fact paper has focus on one of the issue of the opponent identification which is build by the different organization such as news, debate, online articles, etc. Here many researchers has already done lot of work but that is focus only on the content classification which in this paper not only content but opponent are also identified then classify. In few work opponent classification are done on the basis of the background information, but this paper overcome this dependency as well here it classify all the opponent without having prior knowledge. Results shows that using an correct dictionary and proposed algorithm it works better than previous one done. As propose work give a precision efficiency value of 0.85 which is quite impressive as well as in document separation maximum of 0.8 precision value is achieved. As there is always work remaining in every because research is a never ending process, here one can implement similar thing for different other language.

REFERENCES

[1] B. Baker, How to Identify, Expose and Correct Liberal Media Bias. Media Research Center, 1994.

[2] D.A. Schon and M. Rien, Frame Reflection: Toward the Resolution of

[3] Freddy Chong Tat Chua, Hady W. Lauw, and Ee-Peng Lim Generative Models for Item Adoptions Using Social Correlation. IEEE transaction on knowledge and data engineering Vol. 25, NO. 9, 2013.

[4] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval" Information Processing and Management 24, 1988. 513-523. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) Readings in I.Retrieval. Morgan Kaufmann. pages.323-328.1997.

[5] G. Salton, "Automatic Text Processing: The Transfor-mation, Analysis, and Retrieval of Information by Computer" Addison-Wesley Publishing Company.1989.

[6] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu. "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection". IEEE transaction on system man and cybernetics-part A: System and Humans Vol. 42 no. 3, 2012.

[7] M. Wasson, "Using leading text for news summaries: Evaluation results and implications for commercial summarization applications" In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL. Pages.1364-1368, 1998.

[8] Souneil Park, Jungil Kim, Kyung Soon Lee, and Junehwa Song, Member, IEEE. Disputant Relation-Based Classification for Contrasting Opposing Views of Contentious News Issues. IEEE transaction on knowledge and data engineering Vol. 25 no. 8, 2013.

[9] S. Somasundaran and J. Wiebe, "Recognizing Stances in Ideological Online Debates," Proc. NAACL HLT Workshop Computational Approaches Analysis and Generation Emotion in Text (CAAGET '10), pages. 116-124, 2010.

[10] Xiang Wang, Xiaoming Jin, Meng-En Chen, Kai Zhang, and Dou Shen Topic Mining over Asynchronous Text Sequences. IEEE transaction on knowledge and data engineering no. 1 , 2012.

[11] Xibin Gao and Munindar P. Singh, Fellow, IEEE Mining Contracts for Business Events and Temporal Constraints in Service Engagements. . IEEE transaction on knowledge and data engineering Vol. 24 no. 1 ,2013.

[12] Yuefeng Li, Ning Zhong, and Sheng-Tang Wu "Effective Pattern Discovery for Text Mining". IEEE transaction on knowledge and data engineering Vol. 24 no. 1 , 2012.

[13] Vivek Tiwari , Vipin Tiwari, S. Gupta. Association Rule Mining: A Graph based approach for mining Frequent Itemsets. IEEE International Conference on Networking and Information Technology (ICNIT 2010) Manila, Philippines,IEEE Catalog Number:CFP1023K-PRT, ISBN:978-4244-7577-3, 2010.

