www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3, Issue 10 October, 2014 Page No. 9038-9043

Contentious News Article Categorization by Identifying Opponents

Pradip Patil, Prof. Srikant Lade
Student,RKDFIST,Bhopal,
RKDFIST,Bhopal

Abstract— With the use of mining in the field of the text researchers get new era for working, which discover knowledge from the text documents. One of the application of text mining is to categorize the opponent from the contentious news article is focus in this paper. Here as per the different opponent present in the article are identify which is base on the dictionary and frequency of the opponent in the article then all the opponent are also classify into two main party where each opponenet relation is find with the other is based on the words they use in the sentence.

To evaluate this work articles from different debate category has been passed and got results that is very highly acceptable.

Index Terms- **opponent classification,** document analysis, **decision support systems, text mining**.

I. INTRODUCTION

As the opening of the field of Natural language processing is done in the text mining new fields has been emerge where algorithms is develop to generate knowledge from the text document. There are many application of this field where different categorization, evaluation, searching, conclusion is done so that work will be find superior and things will be easy.

It is different from the normal search procedure where it is already known to the user that what is the actual thing need to find, but in the text mining it is not define and not known that what will be the output from the collection of the text documents.

Now by keeping the search procedure continue it is required that with time new rules may involve as the language change very frequently from person to person . If process continue for similar things then results obtain from it is enexpected as there is no direct criteria for the article writing because every writer pattern of writing is different.So in text mining the main goal is to find the unknown information from the document that is not yet discover.

From the above discussion it can be said that it is an combination of different field that include text information retrieval, clustering, categorization, topic tracking, etc. So text mining is providing the a solution to replace the human effort by the machine learning process, which simply retrieve document then process it and finally provide information from it. This information retrieval is depend on the generated pattern or relationship between the sentences, because without these it might not possible for the system to discover any fruitful information from the document or bunch of documents.

One of the wide application of the text mining is analyze the document for the natural language processing that whether the document contain information of which category. This is a kind of separation of the document from one category to other

By allotting it from obtain relationship from the category.

In the similar fashion finding the information from the continues issues document such as kind of debate, discussion on opponenet views. Here information is like finding the main two opponent then what are the different sentence that is in favour or oppose of the main opponent in the document. One more information that can be generate from the system is differenciating other opponent as well. Decide from which party they belong all these thing can be develop on the basis of the different relation which they develop among the system.

This paper is focus on developing a system where each opponent in the article or input document can be find then decide the main two opponent in the document after that classify other opponent in the document on the basis of the two main opponent. Finally conclude that article is in favour of which party.

II. Related Work

Rainer Malik et. al. have used a combination of algorithms of text mining to extract keywords relevant for their study from various databases and also identified relationships between key terminologies using PreBIND and BIND system [6]. Boosting classifier was used for performing supervised learning and used on the test data set. Henriksen and Traynor [7] presented a scoring tool for project evaluation and selection. Ghasemzadeh and Archer [8] offered a decision support approach to project portfolio selection. Machacha and Bhattacharya [5] proposed a fuzzy logic approach to project selection.

In [2], the TFIDF weight theme is employed for text illustration in Rocchio classifiers. Additionally to TFIDF, the worldwide IDF and entropy weight theme is projected in [9] and improves performance by a median of 30 %. Varied weight schemes for the bag of words illustration approach got in [4]. the matter of the bag of words approach is the way to choose a restricted range of options among a vast set of words or terms so as to extend the system expeditiously avoid over lifting [1].

Term based metaphysics mining ways conjointly provided some thoughts for text representations. As an example, stratified agglomeration [7] was wont to confirm synonymy and subordination relations between keywords. Also, the pattern evolution technique was introduced in [5] so as to boost the performance of term based metaphysics mining. These analysis works have primarily targeted on developing economical mining algorithms for locating patterns from an outsized knowledge assortment. within the presence of those setbacks, sequent patterns employed in data processing community have clothed to be a promising various to phrases [3] as a result of sequent patterns get pleasure from sensible applied mathematics properties like terms. to beat the disadvantages of phrase based approaches, pattern mining based approaches or pattern taxonomy models (PTM) [1]) are projected, that adopted the conception of closed sequent patterns, and cropped nonclosed patterns.

Research in mass communication has showed that opposing opponents talk across each other, not by dialogue, i.e., they martial different facts and interpretations rather than to give different answers to the same topics [3].The discourse of contentious issues in news articles shows different characteristics from that studied in the sentiment classification tasks. First, the opponents of a contentious issue often discuss different topics, as discussed in the example above.

Butler *et al.* [9] used a multiple attribute utility theory for project ranking and selection. Loch and Kavadias [7] established a dynamic programming model for project selection, while Meade and Presley [8] developed an analytic network process model. Greiner *et al.* [91] proposed a hybrid AHP and integer programming approach to support project selection.

III. Proposed Work

Articles having contentious news data are the input of the work and main purpose of this paper is to find the different opponent present in the article then find relation between the opponent. For classifying the article in the two party. So first divide the whole document in the form of sentence collection, after this follow below steps

Generate Sentence:

As article is a collection of sentences and to analyze any text data first it need to make in as per the requirement of the system. So here input document is arrange in form of bag of sentences or matrix.

a). Opponent Collection

Now from each sentence remove all the words that are use for framing the sentence or those words which are found in the dictionary of that language. It is assumed that the words that are not present in the library are opponent or name of some person. In this way all the words that are not matched with the dictionary words are collect in the set D. So D is the set of possible opponent.

This can be understand as let a Sentence S = “Mr Barack is the young president of entire history ”, in current sentence all words like {Mr, is, the, young, president, of, entire, history} are present in the dictionary but barrack word is not present so it is consider as the Opponent. Here one more thing is introduce that is to find the term frequency TF of the opponent as it contain list of only those opponent that are above some threshold value of frequency in the article.

b). Filter Main Opponent

In this step one all the opponent collect in the set D are count as the set contain same opponent number of time so the opponent with the greater number of repeatation is consider as the main opponent. While the opponent with lower order of opponent repeatation is consider as the other opponent. Now this can be understand as let D ={a,b,c,a,c,b,a,d,e,a,b,r….} in D unique opponents are {a,b,c,e,r} where Repeatation of the opponent are (a, 4), (b, 3), (c, 2), (e, 1) (r, 1). So from the D set if M represent the main opponent set then M = {a, b} as the greatest number of time 'a' is repreat then 'b' is present in the opponent list. This repeatation represent the presence of the opponent in the different sentence of the document so the document which cover most frequent opponent are identify here.

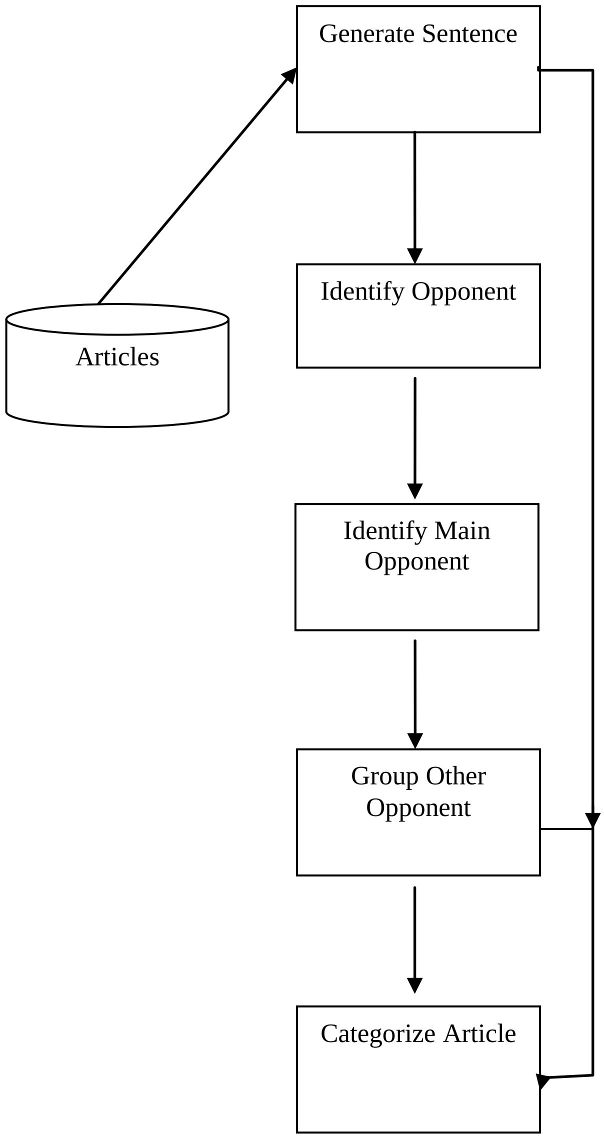

Fig.1 Block diagram of Proposed Method

c). Classify other Opponent

Once main opponent are identified by the system another step is to find the relation between another opponent with the main opposing party, this is develop in-order to classify other opponent in the opposing party. For this main logic include following points:

- i) Collect all sentence that include the main opponents in the article in C set.
- ii) For each Other opponent OD search that it is present in the sentence.
- iii) If other opponent present in the sentence then find the number of prons and cons words present in the sentence.
- iv) If prons is greater then the cons then the opponent is in favour of the main opponent M←OD.
- v) Otherwise it is oppose of the main opponent present in that sentence M'←OD.

d). Article favoring

In this step it is conclude that article is in favour of either of the opponent. An article is classified to a specific side if more of its quotes are from that side and more sentences are similar to other side. A quote is identified to a particular by passing it into SVM. Here feature need to be generate for the SVM that is developing the pattern on the basis of the opponent partion and verbs use in the quote. By using proper pattern rules false sentence classification be reduce.

: Given an article a, and the two sides b and c,

classify a to b if $(Qb + Sb)/Su >= (Qbc * \acute{\alpha} + \beta *Sbc)/Su$

classify a to c if $(Qc + Sb)/Su >= (Qbc * \acute{\alpha} + \beta *Sbc)/Su$

classify a to other, otherwise,

where

SU: Number of all sentences of the article

Qb: Number of quotes from the side i.

Qbc: Number of quotes from either side i or j.

Sb: Number of sentences classified to i by SVM.

Sbc:: Number of sentences classified to either i or j.

Parameter tuning. Two parameters ά & β are used for article classification. The parameter ά serves as a threshold for the ratio of quotes from a specific side: for example, if an article is written purely with quotes and ά is set to 0.8, the article is classified to a specific side if more than 80 percent of the quotes are from that side. The parameter β serves as a threshold for the ratio of sentences that are classified to be similar to the arguments of a specific side: for example, if an article does not include quotes from any side and β is set to 0.7, the article is classified to a specific side when more than 70 percent of the sentences are determined to be similar to a specific side's quotes.

Proposed Algorithm

Input: A // Article

Output: D, M, Class

```
S←Pre_Process(A)   // S: Sentence Matrix
D←Opponent_Collection(S)   // D: Opponent Matrix
M←Main_Opponent   //M Contain two main opponent
Loop d= 1:D-M  // For each other opponent
    Loop s = 1:S
        If contain_opponent(s,M,d)
            P←Search_pros(S)
            N←Search_cron(S)
            If P>N
                Class←{M,d}
            Otherwise
                Class←{M', d}
            Endif
        Endif
    EndLoop
EndLoop
```

IV. Experiment & Result

This section presents the experimental evaluation of the proposed perturbation and de-perturbation technique for privacy prevention. To obtain AR this work used the Apriori algorithm [1], which is a common algorithm to extract frequent rules. All algorithms and utility measures were implemented using the MATLAB tool. The tests were performed on an 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional. Experiment done on the customer shopping dataset which have collection of items, cost, Total amount, etc. attributes.

Dataset

Here two set of documents are use for the evaluation pupose first is of Debate and other is article on current issues. Article is divide into two category only that is of either side of the parties.

	First Party	Second Party	Total
Set1	3	4	7
Set2	4	6	10

Table1 represent the Document set wise actual separation

Evaluation Parameter

Opponent in Tables			Total
	First Party	Second Party	
Set1	5	7	12
Set2	8	11	19

Table2 represent the Document set wise proposed work separation .

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

Precision = true positives / (true positives+ false positives)

Recall = true positives / (true positives +false negatives)

F-score = 2 * Precision * Recall / (Precision + Recall)

In above true positive means that the submit positive document is identify as positive document and false negative means submit positive document is identify negative document and vice versa. False Positive means submit negative document is identify as positive.

Results:

There are article classification done on the basis on the opponents relationship with other opponents. As mention in D part of the paper.

First Party			
	Precision	Recall	F-Measure
Set1	0.4	0.22	0.28
Set2	0.75	0.5	0.6

Table 3. represent the Results of first Party of set wise.

Second Party			
	Precision	Recall	F-Measure
Set1	0.714	1	0.833
Set2	0.545	0.6	0.5716

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

Above results shows that as the use of proper threshold of the opponent selection and dictionary it is possible to have values of opponent identification average precision value above 0.602 which is quite good progress done by the proposed algorithm as compare to the previous work in [19], 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.

Article Classification in First Party			
	Precision	Recall	F-Measure
Set1	1	0.428	0.599
Set2	0.75	0.33	0.459

Table5 represent the Results of first Party of set wise.

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

Table6. Represent the Results of Second Party of set wise.

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 0.85 which is quite good progress done by the proposed algorithm as compare to the previous work in [19], 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.

V. CONCLUSION

Experiment results shows that a remarkable improvement is done by the proposed work for the identification of the opponents as well as the classify them without having any kind of background knowledge or supervised learning. This proposed work shows that the testing produce more effective results from the previous one where 0.85 is the maximum accuracy obtain. So with the regular improvement of the dictionary this can produce similar results with new sentence also as they may include those words which are new in that language. In future many of the format which are referring the opponent such as ‘He, She, I, etc’ need to be identified as this is still remaining.

VI. REFERENCES

1. D.A. Schon and M. Rien, Frame Reflection: Toward the Resolution of Intractable Policy Controversies. BasicBooks, 1994.

2. 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), pp. 116-124, 2010.

3. 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. Pp.1364-1368. 1998.

4. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu “Effective Pattern Discovery for Text Mining”. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[5] 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. pp.323-328.1997.

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

7. 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 TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 42, NO. 3, MAY 2012.

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 TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 12, DECEMBER 2013.