ISSN:2321-1156
International Journal of Innovative Research in Technology & Science(IJIRTS)

DISPUTANT CLASSIFICATION FROM NEWS ARTICLES USING TEXT MINING

Ashutosh Gupta ; Assit Prof. Arvind Mewada TIT Bhopal

Abstract

Today discovery of knowledge from the text data is an interesting research area, as the content has variety of wring context so the analysis fo the data and to produce an assured outcome from the document is not an easy task. Contentious news issues, like data os health care reform debate which contain the disputant need for classification is one of the new field for the text mining. These paper focuses on the disputant categorization of the different article pass as the input. Here things are completely automatic means finding of the disputants after analyzing it from the dictionary, then these disputant are categorize into the opponent. Result shows that articles which have a disputant collection can be arranged without having a prior knowledge of the disputants, or background information. Here proposed work shows better result than the previous work in which prior information need to provide.

Introduction

Text Mining is the process of extracting knowledge from the text document or un arrange written material. Here the main task of finding the association of the extracted information from the new thoughts. 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.

Here the very first step while taking the information is to remove all irrelevant information from the search space that is not meeting the actual requirement. While 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 depending 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 applications of the text mining is analyze the document for the natural language processing that whether the document contains 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 opponents views. Here information is like finding the main two opponents then what are the different sentence that is in favors or oppose of the main opponent in the document. One more information that can be generate from the system is differentiating other disputant 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 disputant in the article or input document can be finding then decide the main two disputant in the document after that classify other disputant in the document on the basis of the two main disputant. Finally conclude that article is in favors of which party..

Related Work

Many varieties of text mining are planned within the past. A standard one is that the bag of words that uses keywords (terms) as elements within the vector of the feature space. In [7], 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 [2]. 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 [5] was wont to confirm synonymy and subordination relations between keywords. Also, the pattern evolution technique was intro-

duced 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 [1] 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.

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. Research in mass communication has showed that opposing disputants 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 [1].

In [4] 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. Boosting classifier was used for performing supervised learning and used on the test data set. In [3] proposed a fuzzy logic approach to project selection. Butler et al. [9] used a multiple attribute utility theory for project ranking and selection. In [7] established a dynamic programming model for project selection, while Meade and Presley [8] developed an analytic network process model. In [91] proposed a hybrid AHP and integer programming approach to support project selection.

Several works have used the relation between speakers or authors for classifying their debate stance [5], [18]. However, these works also assume the same debate frame and use the debate corpus, for example, floor debates in the House of Representatives, online debate forums. Their approaches are also supervised, and require training data for relation analysis, for example, voting records of congress people.

Proposed Work

As the text document contain many information that is relavent to the current search but might not be. So first divide the whole document in the form of sentence collection, after this follow below steps

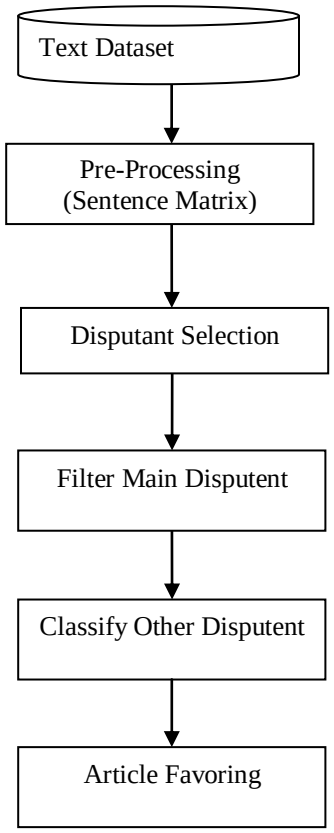

Figure 1. Block diagram of Text Mining processing

Pre-Processing:

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. Disputant 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 disputant 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 disputant.

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 Disputant. Here one more thing is introduce that is to find the term frequency TF of the disputant as it contain list of only those disputant that are above some threshold value of frequency in the article.

B. Filter Main Disputant

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

C. Classify other Disputant

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

i) Collect all sentences that include the main disputants in the article in C set.

ii) For each Other disputant OD searches that it is present in the sentence.

iii) If other disputant present in the sentence then find the number of prons and cons words present in the sentence.

iv) If prons is greater than the cons then the disputant is in favors of the main disputant M□OD.

v) Otherwise it is oppose of the main disputant present in that sentence M'□OD.

D. Article favoring

In this step it is conclude that article is in favour of either of the disputant. 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 disputant 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 * ά + β *Sbc)/Su

classify a to c if (Qc + Sb)/Su >= (Qbc * ά + β *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

1. S←Pre_Process(A) // S: Sentence Matrix
2. D←Disputant Collection(S) // D: Disputant Matrix
3. M←Main_Disputant //M Contain two main opponent
4. Loop d= 1:D-M // For each other disputant
5. Loop s = 1:S
6. If contain_disputant(s,M,d)
7. P←Search_pros(S)
8. N←Search_cron(S)
9. If P>N
10. Class←{M,d}
11. Otherwise
12. Class←{M', d}
13. Endif
14. Endif
15. EndLoop
16. EndLoop

Experiment and 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.

Table 1: represent the Document set wise actual separation

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

Evaluation Parameter

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 identifying as positive.

Results

There are article classifications done on the basis on the disputant's relationship with other disputants. As mention in D part of the paper.

Table 2 : represent the Document set wise proposed work separation

	Article in favour	
	First Party	Second Party
Set1	3	3
Set2	3	7

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

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

Table 4 : Represent the Results of Second Party of set wise.

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

Above results shows that as 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.

Conclusion

In this paper it is obtained that a remarkable improvement is done by the proposed work for the identification of the disputants as well as the classify them without having any kind of baground knowledge or supervised learning. This proposed work shows that the testing produce more effective results from the previous one where 0.75 is the accuracy obtain. So with the continous updation of the dictionary this can produce similar results. There is plenty of work is required to do in this field where one can apply its algorithm such as in different other language as the processing will change most of the steps.

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]. 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.

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

[5]. 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.

[6]. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu "Effective Pattern Discovery for Text Mining". IEEE Transaction on knowledge and data engineering, Vol. 24, No. , January 2012.

[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 Transaction on System 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 Transaction on knowledge and data engineering, Vol. 25, NO. 12, December 2013.

Biographies

FIRST A. ASHUTOSH GUPTA received the B.E. degree in Information Technology branch from the University of Guru Ghasidas vishwa vidayalaya, Bilaspur, Chhattisgarh India, in 2010, pursuing the M.Tech. degree in Computer Science and Engineering branch from the University of Rajiv Gandhi Technical University, Bhopal, Madhya Pradesh India. His teaching and research areas include Data Mining.

FIRST B. ARVIND MEWADA received the B.E. degree in Information Technology branch from the University of Rajiv Gandhi Technical University, Bhopal, Madhya Pradesh India, in 2007, M.Tech. degree in Computer Science and Engineering branch from the MANIT in , Bhopal, Madhya Pradesh India,in 2010. His teaching and research areas include Data Mining. manit.106@gmail.com