International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Text Mining with Topical and Informative Measures using Semantic Ontology

G Shobarani, K Arulanandham

Abstract: ***The text mining has been identified as dominant throughout the era in many scientific problems. Number of techniques have been identified and proposed earlier, each uses different features towards mining text information. However, they suffer to achieve higher performance in mining relevant text information or documents. To support the development of text mining tasks, an efficient topical and informative measures based algorithm is presented, which uses semantic ontology and the taxonomy as dictionary. The text features of the documents has been extracted to generate set of terms. For any document, the text features are used to remove the noisy features like stop words, stemming and tagging. With the noise removed pure terms, the method estimates the Topical Depth Similarity (TDS) and informative Depth Similarity (IDM) measures. The measures has been estimated towards each document to perform text mining. The input query has been estimated for the TDS measure to identify the category of the query. According to the category of query, the method estimates the TDS and IDS measures to mining the text. The proposed method improve the performance of text mining with reduces false classification ratio.***

Keywords:Text Mining, Semantic Ontology, Taxonomy, TDS,IDS, Feature Extraction, Similarity Measures.

I. INTRODUCTION

The growth of information technology allows the users to maintain various data in the form of text, images and videos. However, it is contemporary that maintaining the combined data in any document. For example, the content related to data mining has been placed in a document which represents the concept in text and images. Similarly, any document would contain various form of information to support the purpose. So, the volume and size of documents in any data base gets increased. The increase in size and volume, challenges the organization in maintaining the document in a sophisticated and organized manner. The organization of document in any storage is most important which reflect on the performance of data retrieval. Consider a medical organization which maintains various documents related to different diseased persons. But at a point, an medical practitioner would like to read the medical documents related to lung cancer, thenhe has to search throughout the documents which is hectic job and would take more time. If the documents are organized properly, it will take fraction of seconds to fetch the documents needed. This is where the clustering comes to play, and it is a process of grouping the documents of any dataset according to the number of groups based on the features.

Manuscript published on 30 September 2019.
*****Correspondence Author(s)
G Shobarani, Ph.D Research Scholar, Bharathiar University, Coimbatore, Tamilnadu, India
K Arulanandham, Head, Department of Computer Applications, Government Thirumagal Mills College, Gudiyattam, Tamilnadu, India

However, there exist number of features in the document, the text features are the most important and key one which decides the topic of the document. The image feature would speak about the concept but the automated system could not identify the category instantly. This is why, text mining has been considered for the grouping and mining of documents.
The text mining has been approached with several techniques like K-means clustering which estimates the distance between the document text according to the text features. Similarly, the Term Frequency and Inverse Document Frequency measures has been used in extracting different documents in many methods. Similarly, there are methods which use only the text feature present in the document to perform text mining. It will not be efficient when the method consider only the text terms of the document, it is necessary to consider the semantic meanings also. The document would speak about the topic as well as it would represent some semantic meanings about the category. Further, the category of the document would be classified into number of levels. For example, the category "Computer" can be classified into Computers, Computers / Programming Language, Computers/Programming Language / Java. Similarly, the categories of the document can be classified up to any number of levels. The text mining algorithm should consider the exact category and the sub classes of the data set. However, different features are used for text mining, considering the topical and informatics measures are more important. This paper presents a text mining algorithm which considers the topical and informatics depth similarity measures. The topical measure represents the documents depth in covering the topical measures where the informatics similarity measure represents the depth of document in covering the most information related to the category. The next section discusses the proposed algorithm in detail.

II. PROCEDURE FOR PAPER SUBMISSION

There are number of algorithms has been discussed for the text mining problem and various approaches of text mining is discussed in this section.
In [1], the author presents various text mining techniques and applications in detail. The method present different discovery patterns to analyze documents from huge volume of data set. The text mining process has been discussed as patterns and features based algorithm in extracting related documents.
Using text mining to classify research papers [2], uses natural language processing tools towards the classification of research papers. The method has been adapted support vector machine and naïve bayes algorithms in classification.
In [3], the author presents a trend recognition algorithm for journal papers.

The popular TF and IDF algorithms have been used. The TF measure fetches the topical strength of the document where IDF measure fetch the topical strength of other categories. The method has been adapted in quality control of Japanese documents.

In [4], the Korean deciphering monetary board data has been mined using field specific keywords to support the communication of central bank of Korea. The method fetches the indicators based on text to support the explanation of BOK monetary decisions. The method produces efficient results on textual classification and media based measure to support economic policy.

In [5], a sentimental analysis approach to support social media has been presented. The method analyzes the social media data sets by constructing a dictionary of words. Based on the words of dictionary, the method selects a set of tags on any topic. Based on the polarity of tags, the method performs classification of tweets. The US election data set has been used for evaluation.

Text-based predictions of beer preferences by mining online reviews [6], discusses the process involves decomposing the texts into the words composing them and using the frequency of those words to predict the text polarity. The text categorization approach may use single words composing the texts or longer combinations of the words. Longer combinations of the words have the advantage of better representing the complexity of human language compared to single words. Moreover, less frequent terms may also better discriminate between reviews than more common words. This study tests these two hypotheses in the context of beer reviews. It shows that the words used in the reviews can be used to predict consumer's preferences for beer. Moreover, it shows that the use of less frequent terms in the predictive models outperforms the use of more frequent terms.

In [7], the author presented an opinion classification algorithm towards the development of tourist reviews. Earlier systems used implicit, infrequent measures in classifying the opinion. In earlier approach, a aspect based classification is presented which suffers with higher irrelevancy. To improve the performance, a fuzzy based algorithm is presented which extract the aspects according to the opinions of user. The method produces accurate classifications than previous algorithms.

In [8], a sentiment analysis algorithm is presented which uses lexicon to analyze the performance of teachers. The method supports the evaluation of teacher's performance by obtaining feedback from students. From the feedback obtained from students, the text mining techniques have been used to analyze the feedback to promote performance evaluation of students.

In [9], the author describes the idea of converting the Ecommerce data into classification problem. The method applies text mining algorithms to identify the similar records of e-commerce data. The method verifies six different strategies towards the selection of key words and grams. Finally, the classification is performed with SVM, Random Forest, and Logistic Classifier.

In [10], the authorpresent a prospective analysis based on the study of PubMed search history enables us to determine the possible directions for future research.

In [11], the author present a performance analysis algorithm for Airline marketing data using text mining approaches. The method first extracts the key words from the market data and identifies the most prominent terms. Further the method estimates the influence of key word towards corporate performance analysis.

In [12], a semantic pattern mining algorithm has been presented towards text mining. The method generates number of semantic pattern mining and measures the frequency of semantic patterns. Such pattern frequency has been sorted using suffix array sorting to perform semantic pattern mining.

In [13], a user interest prediction model is presented which combines Latent Dirichlet Allocation (CLDA), which uses big data. The method identifies and learns the topics from blogs of big data. The method learns the short text from long text towards the improvement of text mining.

In [14], the author presents a bug report classification algorithm using text mining. The method uses text mining algorithm in interest prediction in multiple stage approach. First the method analyzes the text features from bug reports and estimates the probability to classify them in three levels. The learned features are applied with machine learning to perform prediction.

In [15], a sentiment analysis of reviews towards the scholarly papers has been presented. The method automatically predicts the recommendation generated for any article by analyzing the statements of reviews. The method estimates the polarities to perform classification. Towards the development, a multiple instance learning network with abstractbased memory mechanism (MILAM) is presented.

In [16], the author reviews on various approaches of web document clustering. The paper discusses various aspects and algorithm which has been used for mining text from web documents. Finally, a harmony search algorithm has been presented towards the grouping of web documents. Different algorithms have been compared for their performance in different parameters.

All the above discussed methods suffer to achieve higher performance in text mining and document classification.

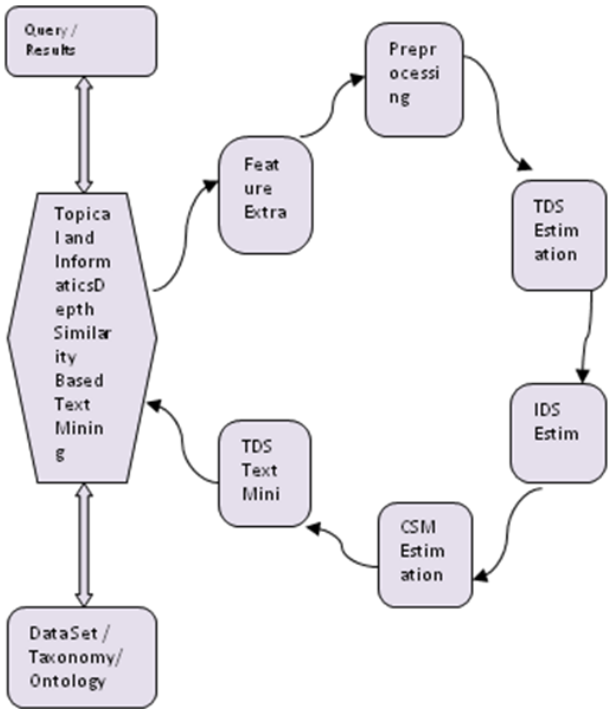
III. REVIEW CRITERIA

The proposed topical and informatics similarity based text mining algorithm starts with the feature extraction from the document set available. For each document the text terms has been extracted, and a term set has been generated. Generated term set for a document has been preprocessed to obtain the root features. With the root features, the method estimates the TDS and IDS measures towards each category. Finally, the category of the document has been identified. Towards text mining, the method estimates the same measures for the input query to identify the class of query. Based on the taxonomy and ontology of the class identified, the documents are measured for their similarity on topical and informatics features. According to these values, the categorical support measure (CSM) has been measured.

Retrieval Number: K23360981119/19©BEIESP
DOI: 10.35940/ijitee.K2336.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Figure 1: Architecture of TIDS Text Mining Algorithm



Based on the value of CSM, the document either selected to produce the result on text mining or rejected from the result. The detailed approach is presented below.

The architecture of proposed TIDS text mining algorithm has been presented in Figure 1 and shows various components which is explained in detail in this section

A. Feature Extraction:

The feature from the input document set has been extracted here. For any document from the data set, the text features are extracted. Extracted features have been split into terms to produce terms set Tes of any document Di. Extracted term feature has been used to identify the topic, estimating the topical and informatics measure in the upcoming stages of the algorithm.

B. Preprocessing:

This acts as the gateway for text mining algorithm which uses the term set produced in the feature extraction. The algorithm reads the term set obtained and removes the stop words and performs stemming on the terms of terms set Tes obtained in the previous stage belong to a document Di. Consider the term set Tes contains the text features of the document, then the method eliminates the stop words found in the term set Tes based on the terms of SwS (stop word set). The remaining terms of Tes has been applied with tagging which is performed with the support of Part of Speech Tagger provided by Stanford university. The preprocessed term set has been used for further estimation of measures.

Consider the term set of the document is Tes then the method eliminates the stop words as follows:

Term set Tes

$$= \int_{i=1}^{size(Tes)} Tes(i) \in SwS, Tes \cap Tes(i), Tes(i) = Tes(i)$$

Here SwS is the stop word list used to eliminate the stop words from the term set Tes, if the term is present in the list then it will be eliminated from the term set.

Second, the stemming is the process of trimming the terms of Tes, which is performed by removing the “ing” and “ed” at the end of terms of Tes.

Tagging is performed for all the terms of term set Tes, it returns a tag which specifies that the term is Noun, Verb. If the term is Noun then it will be selected otherwise it will be eliminated as follows:

Term set Tes =

$$\int_{i=1}^{size(Tes)} Pos.Tag(Tes(i)) == Verb, Tes \cap Tes(i), Tes(i) = Tes(i)$$

The result of tagging is the outcome used to estimate the measures towards text mining.

C. TDS Estimation:

The topical depth similarity measure represent the weightage of the document in representing the topic of the class. In general it has been measured according to the terms of the topic present in the document. For example, if the document to be covered under the topic, diabetic, it should speak the terms like blood sugar, HbA1C, fasting sugar, Diabetic, and so on. For each category or topic, the method maintains the taxonomy, which lists the terms related to the concept. With the taxonomy, and the term set of the document, the method estimates the topical depth similarity measure (TDS). First the method estimates the number of terms of Term set Tes contained in the class taxonomy CT. Second, for each term Ti found in the document, the method estimates the recurrence appearance factor (RAF). According to the value of both, the method estimates the TDS value.

TDS Estimation Algorithm:

Input: Term Set Tes, Class Taxonomy CT

Output: TDS.

Start

Read input term set Tes and taxonomy CT.

Compute the terms of Tes appear in CT.

ATes = $\int_{i=1}^{size(Tes)} \sum Tes(i) \in$ CT

For each term Te from ATes

Compute Number of appearance of Te

NoA = $\int_{i=1}^{size(Tes)} Tes(i) == Te$

End

For each term Te

Compute recurrence appearance factor RAF = $\frac{NoA}{size(Tes)} \times \frac{NoA(Te)}{\sum NoA(OtherTerms)}$

End

Compute topical depth similarity measure Tds = $\frac{ATes}{size(CT)} \times \frac{\sum RAF}{size(Tes)}$

Stop

The above discussed algorithm shows how the topical depth similarity measure has been estimated which is been used toward text mining.

D. IDS Estimation:

The informative depth similarity measure represents the depthness of topic being discussed in the document. The document would contain set of terms related to the concept but it cannot be assured that the document briefs the related information.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The modern documents would represent the semantic meanings in the document. Also, it is necessary for the document to discuss the semantic features. For example, a document discusses the topic clustering would contain the term "grouping" but it should also cover the semantic meanings like, "merge, same class" and so on. The IDS measure represents the document strength in presenting the semantic features. It has been estimated according to the semantic meanings or terms present in the document. To measure that, the method first computes the list of semantic meanings present in the document. Then the method estimates the number of relations the document towards the semantic class. Using these two, the method computes the IDS measure.

Algorithm:
Input: Term Set Tes, Ontology CO
Output: IDS
Start
Read input term set Tes and Ontology Co
Compute the number of meanings present.
Nmp = $\sum_{i=1}^{size(Tes)} Tes(i) \in Co$
Compute number of semantic relations present.
NoSp= $\int_{i=1}^{size(Tes)} \sum_{j=1}^{Nmp} Nmp(j) \rightarrow Tes(i) \in Co$
Compute IDS = $Nosp / Nmp$
Stop

The procedure of estimating the informatics depth similarity measure has been presented in the above algorithm. Estimated IDS measure has been used to perform text mining in the next stage.

E. TIDS Text Mining:

The problem of text mining has been approached with the TIDS (Topical informatics depth similarity) based text mining algorithm. For a query q given, the method first extract the features as term set Ts, and performs preprocessing to remove the noise features. With the preprocessed feature set, the method estimates the Topical depth similarity measure (TDS) towards each class of documents. Based on the value of TDS, the method identifies the class of query. With the identified class, the method estimates the TDS and IDS measures towards each class of concept. Using these two measures, the method computes the TIDS value. Based on the value of TIDS, the method performs text mining.

TIDS Algorithm:
Input: Query Q, Data Set Ds, Taxonomy CT, Ontology CO
Output: Result Set Rs
Start
Read query Q.
Term set Tes = Feature Extraction (Q)
Tes = Preprocessing(Tes)
For each class c
Compute TDS measure QTds = $\int_{i=1}^{size(C)} TDSEstimation(c, Tes)$
End
Choose the class with higher TDS value.
Class C= $\int_{i=1}^{size(C)} Max(C(TDS))$
For each document Di of Ds
Tes = Feature Extraction(Di)
Tes = Preprocessing(Tes)
Tds = Estimate-TDS(Tes)
IDS = EstimateIDS(Tes)
Compute TIDS = TDS×IDS
If TIDS>Th Then
Add Di to Result set.
Rs = $\sum(Dk \in Rs) \cup Di$
End
End
Stop

The above discussed algorithm shows how the text mining is performed. The method estimates the topical depth similarity and informatics depth similarity measures to estimate the TIDS value. Based on the value of TIDS, the method selects the document to the result set.

IV. RESULTS AND DISCUSSION

The proposed topical and informatics depth similarity based text mining algorithm has been hard coded and its performance has been evaluated using different data set. The method has been implemented using Advanced java and the performance of the method has been evaluated under different parameters. This section present the results obtained by the proposed algorithm and present the comparative results.

Table 1: Evaluation Details

Parameter	value
Tool Used	Advanced Java
Data Set	Reuters, KDD, UCI
Number of Class	15
Number of Documents	10 lakh

The Table 1, shows the details of evaluation being used to measure the performance of various methods of text mining. The performance of the algorithms has been measured under the following parameters.

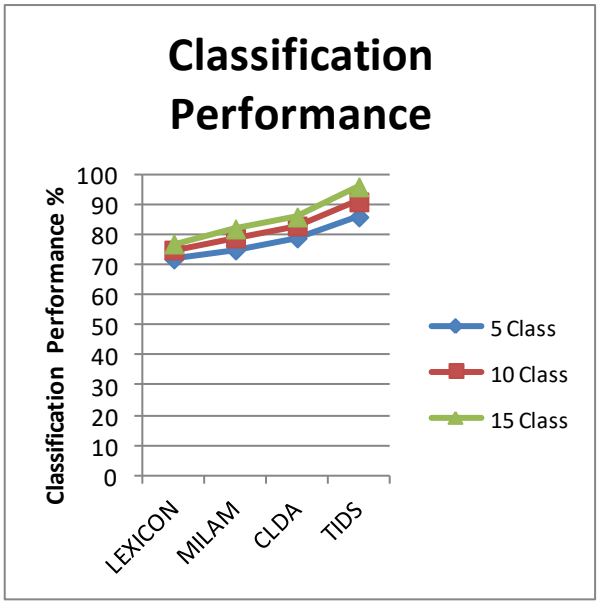


Figure 2: Performance on Classification

The performance on classification accuracy has been measured for different algorithms. The result produced has been compared and presented in Figure 2. The proposed TIDS algorithm has achieved higher classification performance than other methods.

The text mining performance of different methods has been measured and compared with the result of proposed methods. The proposed TIDS algorithm has produced higher text mining performance than other methods. The performance analysis result is presented in Figure 3.

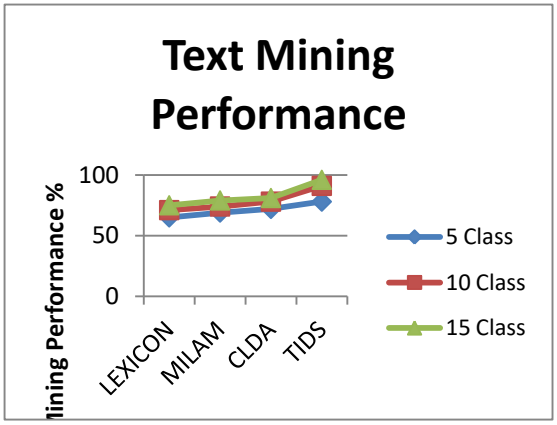


Figure 3: Performance on text mining

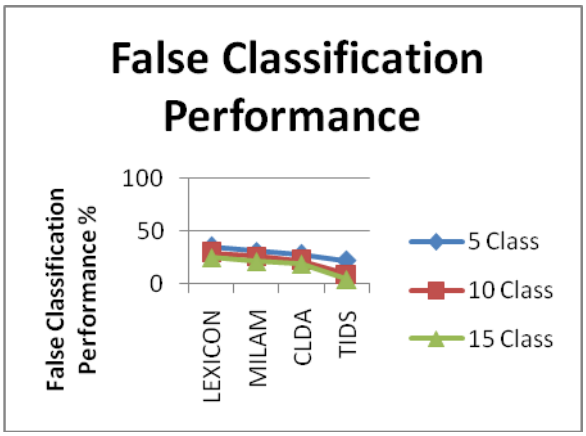


Figure 4: Performance on False classification ratio

The false classification ratio has been measured for the proposed TIDS algorithm and compared with the result of other methods. The proposed TIDS algorithm has reduced the false ratio than other methods.

The performance on time complexity has been measured and compared with the results of other methods and presented in Figure 5. The proposed TIDS algorithm has achieved less time complexity than other methods.

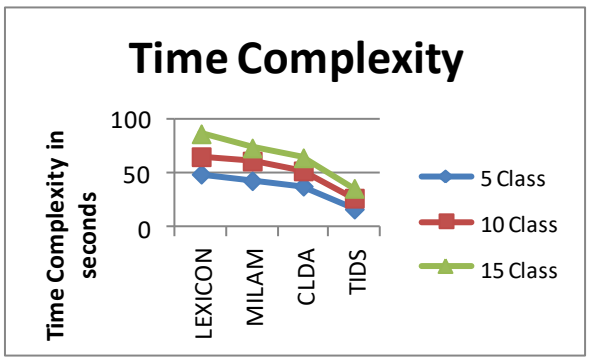


Figure 5: Performance on Time Complexity

V. CONCLUSION

In this paper, an efficient topical and informatics depth similarity based text mining algorithm has been presented. The method first extracts the features of the query and estimates the TDS measure to identify the class of query. Second, for the class identified, the method extracts the features from the documents set, and preprocess to remove the noise. Then, the method estimates the TDS and IDS measures for each document towards the each class with class taxonomy and class ontology. Using these two measures, the TIDS value has been measured for each class. Finally, based on the value of TIDS the method select the documents to the result set. The proposed TIDS algorithm has achieved higher performance in text mining and classification. The false ratio and time complexity has been hugely reduced.

REFERENCES

1. RamzanTalib, Text Mining: Techniques, Applications and Issues, (IJACSA), Volume 7, Number 11, 2016.
2. Sulova, Snezhana, Using text mining to classify research papers, (SGEM), Volume 17, 2017.
3. M. Terachi, R. Saga and H. Tsuji, "Trends Recognition in Journal Papers by Text Mining," (IEEE) 2006, pp. 4784-4789.
4. Park, Ki Young, Deciphering Monetary Policy Board Minutes Through Text Mining Approach: The Case of Korea, (SSRN), 2019.
5. Imane El Alaoui, A novel adaptable approach for sentiment analysis on big social data, (Springer, Journal of Big Data), 2018.
6. AbdelazizLawani, Text-based predictions of beer preferences by mining online reviews, (IBM), 2019.
7. Muhammad Afzaal, Fuzzy Aspect Based Opinion Classification System for Mining Tourist Reviews, (HINDAWI), 2016.
8. Quratulain Rajput, SajjadHaider, and SayeedGhani, Lexicon-Based Sentiment Analysis of Teachers' Evaluation, (HNDAWI), 2016.
9. Gianpiero Bianchi, Renato Bruni, Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms, Mathematical Problems in Engineering, (HINDAWI), 2018.
10. Ekaterina Ilgisonis,AndreyLisitsa,Creation of Individual Scientific Concept-Centered Semantic Maps Based on Automated Text-Mining Analysis of PubMed, Advances in Bioinformatics (HINDAWI), 2018.
11. Jae-Won Hong,Seung-Bae Park, The Identification of Marketing Performance Using Text Mining of Airline Review Data, (Mobile Information Systems, HINDAWI), 2019.
12. X. Song, X. Wang and X. Hu, "Semantic pattern mining for text mining," IEEE(Big Data), 2016, pp. 150-155.
13. LirongQiu, Jia Yu, CLDA: An Effective Topic Model for Mining User Interest Preference under Big Data Background, (HINDAWI), 2018.
14. Yu Zhou, Yanxiang Tong, Combining text mining and data mining for bug report classification, (SEP), Volume28, Issue3, 2016, pp 150-176.
15. Ke Wang, Xiaojun Wan, Sentiment Analysis of Peer Review Texts for Scholarly Papers, (SIGIR), 2018, pp 175-184.
16. Forsati R., Mahdavi M. Web Text Mining Using Harmony Search, Studies in Computational Intelligence, vol 270,2010.

AUTHORS PROFILE

Ms. G. Shobarani studied Bachelor of Science and Master of Computer Applications from University of Madras and Master of Philosophy from Bharathidasan University. She is currently working as Assistant Professor in Department of Computer Science, KMG College of Arts and Science, Gudiyattam, Tamilnadu, India. Her main research interest is in the area of Data Mining, Data Warehousing, Machine Learning and Programming Languages. She has around 18 years of teaching experience and 11 years of research experience.

Dr. K. Arulanandam is currently working as Assistant Professor & Head in the Department of Computer Science and Applications, Government Thirumagal Mills College, Gudiyattam. He has published several papers in reputed National and International journals. He has 17 years of experience in teaching and research.