Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (7s): 927-934

Finding Related Posts On Social Media Through Content Semantic Similarity

Vasavidevi Potta[1*], Dr. Gandi Satyanarayana[2], Dr. Akula Chandra Sekhar[3]

[1*]Research Scholar, Department of Computer Science and Engineering, Avanthi, Institute of Engineering and Technology, Cherukupally (Village), Vizianagaram (Dist)-531162
[2]Professor and Head of the Department of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Cherukupally (Village), Vizianagaram (Dist)-531162
[3]Professor, Department of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Cherukupally (Village), Vizianagaram (Dist)-531162

Abstract

Sentiment research on social media provides businesses with a quick and easy way to track public opinion about their brand, business, directors, and other topics. In recent years, a variety of features and approaches for training sentiment classifiers on datasets have been investigated, with mixed results. In this research, we have proposed an approach for detecting emotion in text and predicting sentiment using semantics as extra characteristics for various datasets and a study on present methods for opinion mining Forum posts has the specific problem of finding related posts to a post at hand. By considering across the related documents the contents of posts are generally consider are whole .Here similarity process are done between two posts with respective segments and should be of same intention. All posts are generally fragmented in the form of group to attain the goal bunches. Now similarities are generally cross view in the forums in the form of sections and that will of same intention. Finding related forum posts are done in the form of division strategy is delineated

I. Introduction

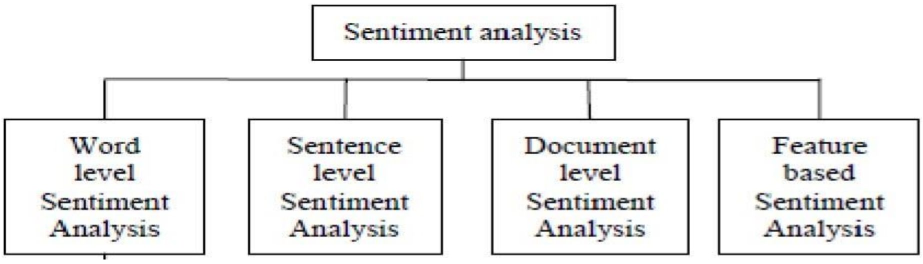
Social media has changed the opinion of people of sharing their views and sentiments in today's world. Nowadays they share their feeling through posts, status, blogs and social networking sites. Presently, millions of people use these social media sites like Facebook, Instagram, Twitter, and other sites to talk their emotions, opinions, and points of view about their day to day practices, by this we get a knowledge of ongoing things in the world through these internet groups. People use these communicating media to inform and influence other people around the globe. Over conversation media a incredible amount of statistics is produced by social media sites. Tweets, Stories, status updates, posts, etc deliver sentiment-rich data from posts, comments, views and reviews. Additionally, mass media gives a platform for commerce to connect with their customers for marketing resolutions. For the most part, people make decisions based on user-created content found on the social media/online. For example if a person is willing to buy stuff, they will review the comments online and then decide whether to buy stuff or not. The volume of material formed by operators is far too large for a usual user to inspect. As an outcome, multiplicity of sentiment analysis techniques are often used to preset this process. People can anticipate through online reviews about purchasing thing or not. Dealers and companies use this information to obtain a better knowledge of their products or facilities so that they may be better suited to their clients' demands. Dispensation, finding, and understanding the factual data available are the main concerns of textual information retrieval systems. The most prevailing contents in Sentiment Analysis comprise of opinions, judgments, feelings, attitudes, and emotions (SA). The tremendous expansion of existing information on the networking sites, such as posts, blogs, Status and social networks, gives various difficult opportunities for new application development. Example, SA may be used to predict suggestions of commodities offered by a reference system based on standards such as favorable or unfavorable remarks about those things SA (Sentiment analysis) uses natural language processing (NPA) to extract feelings, opnions and views from the data like text, audio data, tweets and other media.It is the procedure of categorizing textual opinions into classifications such as "positive," "negative," or "neutral." Subjective analysis, opinion mining, and assessment extraction are additional expressions that are for it.

It comprises of a different responsibilities such as extraction of sentiments, classifying the sentiments, categorization subjectivity in the content, overall summarization of opinions and spam detection, etc. Its purpose is to investigate people's sentiments, attitudes, opinions, and emotions about various items, including as products, people, subjects, organizations, and services.

In order to classify sentiment, several steps must be followed, namely data gathering, data preprocessing, feature extraction, emotion arrangement and valuation. The data is obtained from various sources that are in raw form. Then by finding the mood that you need to uphold in a structured system and its done by preprocessing the data. After the preprocessing, the feature extraction is carried out. After the characteristic has been mined from the data, the sentiment classification must now be carried out. To carry out process, different methods can be used to classify feelings, such as: lexicon, machine learning and hybrid technique, are the basic step for SA. Feature extraction and sentiment classification methods. Sentiment analysis is a difficult task. Some of the important tests in sentiment analysis of local language tweets are sarcasm detection, negation handling and emoticon detection. The main work that will be performed in this paper is to perform semantic analysis on the data including the emoticon detection.

Levels in Sentiment Analysis

Mainly there are four levels that are Document level, Sentence or phrase level, Aspect level and Word level. In Fig1. Four levels of sentiment analysis is shown in the form of a diagram



Document level- It is concerned with assigning a feeling to specific papers. In this level, the entire text is classified as good or bad. By determine the polarization of a text, identify the feeling polarizations of discrete phrases or disputes and aggregate them. Additional techniques include reference determination and additional complex language issues. Some of the tasks that are used is each text concentrates on a single item and contains viewpoint from a single view owner.

Sentence level- Separate judgments are tagged with their corresponding sentiment polarity in this semantic level. Its categorization divides sentences into three categories: positive, negative, and neutral. Finding the sentiment placement of separate word in a sentence and then combining them to get the sentiment of the entire sentence or phrase is the general technique.

Aspect level- It is concerned with assigning a feeling to each phrase as well as identifying the entity to whom the sentiment is addressed. Sentiment categorization at the aspect or feature level is concerned with detecting and extracting product attributes from the source data. This makes use of techniques such as dependency parsers and discourse structures. The following are some of the tasks that are involved like determining the view on features and locating the object characteristics and structures

Word level- For emotion categorization at the verdict level, most recent approaches have relied on the preceding polarization of texts and expressions. Adverbs and adjectives are the best shared features used in text sentiment grouping, but adverbs are also used.

Forums are generally a online discussion site, where people hold there conversations by posts. It is like a message board and different from chat rooms. A traditional approach for finding related document that perform content comparisons across content of posts, the contents are compared by different posts. The relatedness of two posts can then be based on a comparison across segments that serve the same goal. Every posts are generally considered as segments. Segments are generally said as parts (or) sections .In This the relatedness between two posts should be based on similarities respective to segments. The segmentation methods play important role by developing work with monitoring the no of text features ,it identify by parts of post. While this process performing significant jumps are occurred because of that segmentation are occur. Now segmentation of all posts are generally clustered in the form of intention cluster so that the similarities are calculated across segmentation with same intention.

Generally existing forums range from domains like health (e.g., Med help), law (e.g., Expert Law) and technology (e.g.,HP support forum as m). The relatedness of forums are compared based on segmentation process.. Work are done this direction has been done for questions in Q&A archives but not for richer- content posts. The compression can be performed by information retrieval method TF/IDF or BM25 variants or language- model based methods or using topics generated by topic modeling techniques like LDA paraphrasing techniques or even auxiliary external services with the latter been used especially for documents with short and poor content, e.g., tweets.

Here for finding forum posts that are related we generally introduce a novel method . In that method every and each post are considered as set of segments and then the are compute similarity contents across every segments with same intention.

1. Now the segments are identified and grouped into clusters so that the text features are explicit.
2. Now weights are assigned to text features.
3. Now multi segment ranking process are provided to top k forum are done along with there related reference documents, in this cluster play important role, in each cluster similarities are performed between each post along with reference documents

II. RESEARCH METHODOLOGIES

Data Mining is now a day an interesting domain to work on. Previously, many methods were proposed for text document clustering. In this research, we mainly focused on Inter passage approach by using Senti Word analysis for text document clustering

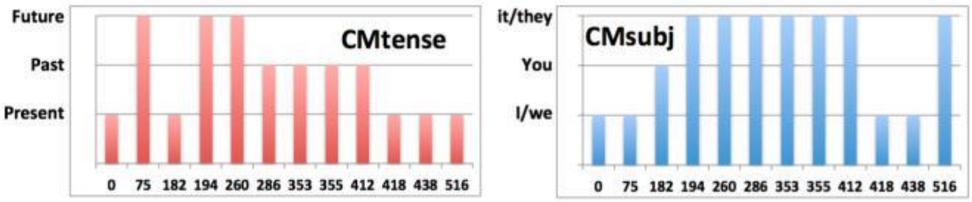
This section reviews some related work to investigate the strengths and limitations of previous methods and to identify the particular difficulties in computing semantic similarity. Related works can roughly be classified into following major categories:

1.Word co-occurrence methods.
2.Similarity based on a lexical database.
3.Method based on Twitter data.

III. PROPOSED SYSTEM:

3.1.SEGMENTATION OF POSTS

The challenging task is finding right segmentation process, from segmented documents, that are occurred in large body of work, if it is in the form of document d, there are $2^{jdj\ 1}$ possible segmentations are occur.
Every segmentation process should be in the form of coherent largely disconnected from its adjacent segments.
Segmentation is the intention-based, these two properties translate to a segmentation where every segment:
conveys a single clear intention conveys by the adjacent segments.



	Possible Segmentations
Boxes	75, 182, 201, 259, 285, 338, 355, 371, ch418, 436, 488, 535
(a) CM_{tense}-Based	([0,75],[76,182],[183,201],[202-285],[286-418],[419-535])
(b) CM_{subj}-Based	([0,182],[183,201],[202,418],[419,488],[489,535])
(c) CM_{qneg} Shift	([0,182],[183,201],[202,438],[439,535])
(d) Intention-Based	([0-182],[183,418],[419-535])
(e) Thematic	([0-49],[50-535])

Fig. . CMs and Segmentations

3.2. SEGMENT GROUPING

Segments with similar intentions are created same group and segments with different intentions in different groups. It is modeled with vector of features, array of information are taken here. Now each cluster are generally communicates respectively the same goal.

I to denote a cluster, and C to denote the set of the generated clusters.

Vector of weights that are based on the feature values are created by us. Vector with the letter F . Now consider two types of weights that capture the strength of the use of each CM categorical value, of each feature.

Now each CM value with in the segment are measured then the comparisons are done to categorical values that belong to same communication appearing in segments.

Using the notion of the distribution table DSb_{CMr} of a communication mean CMr introduced in Section we define the vector F_s of weights, one weight for each feature.

3.3.SEGMENTATION REFINEMENT:

They should have same document with same segments that are end up with same cluster with same intention., if they have the same intention but are not same cluster then consequence document. The segments that belong to the same document in a cluster are concatenated into one.

3.4.MATCHING:

Document matching is one of the best technique plays important role by collecting of documents that are generally related to reference document dq .

Now the dq reference document are measure the relatedness between other documents d0 are lie in the form of IR technique.

3.4.1Matching with respect to a specific Intention:

Every document with some specific intention are projected on each cluster. The specific intention are made by measuring the related documents to reference document d0. Text comparison are computed the relatedness between the documents like IR technique i.e. TF/IDF model.TF/IDF method and its probabilistic variance BM25 consists of a term weighting scheme that weighs a term in a document considering the number searching That variance computes the weight of a term t in a document d^0.

. If s_q and s^0 are the segments of the documents d_q and d^0, respectively, in the intention cluster. where fsq (t) denotes the frequency of the term t in the segment sq, jIj the cardinality of the intention cluster, and jI^{t}j the number of segments in the intention cluster.

```
Algorithm 1 Single Intention Matching
Input: Cluster I, Doc. Collection D, Document dq2D
Output: List of n documents and their intention matching
  MI    ;
  for each sq2Sdq
   if sq62Icontinue; // See footnote1 scr
   0
   for each s02I
     d0 fd j s02Sdg // See footnote2 for
     each t2sq
      scr scr+fsq (t) w(t; s0) log(jIj j Itj)=jItj MI
      M[hd0; scri
  Return fhd0; scri j hd0; scri2MI ^ scr2 top-n scores
```

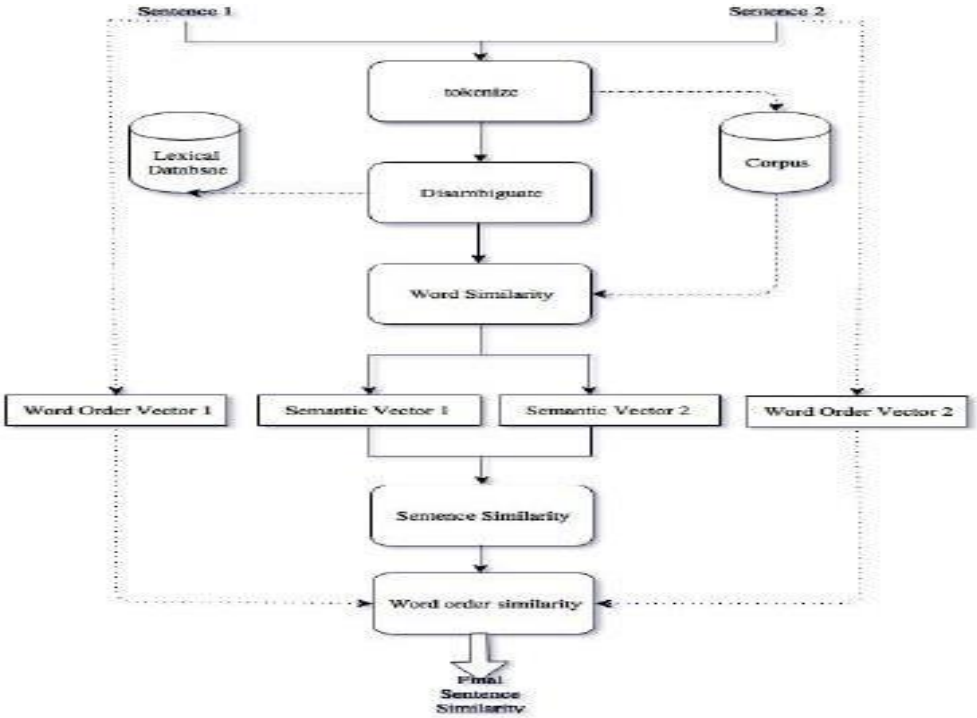
3.4.2.Matching with respect to All the Intentions:.

This algorithm consist of top-n lists generated across the different intentions, ., the set M are used to generate the k most related documents to the reference document d_q R is created as new list contains in every document in lists in M. Each document are associated with the sum of the scores with which this document appears in the various lists in M. The k elements in R with the highest score are returned as answer to the request of the matching documents to the reference document d_q.

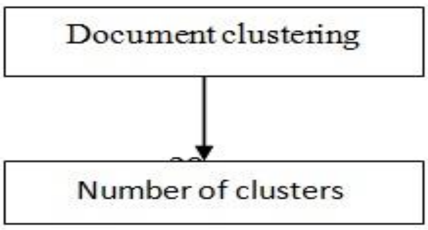
High value for n compared to the value of k, on the other hand, will favor documents that appear in many lists even with not very high scores. We have empirically found that a good choice is an n equal to 2 k

```
Algorithm 2 All Intentions Matching
Input: Document Collection D, Document dq2D, Int
       Intention Clusters C
Output: List of documents
  L   ;       M    ;
  for each I2C
     for each sq2Sdq
       if sq62Icontinue
       MI    SingleIntentionMatching(I,D,dq,n)
       L    L[fMI g
  for each MI 2L
     for each hd0; scri2MI
       if exists hd0; xi2M, with x2R
         M      M[hd0; scri
       else hd0; xi     hd0; x + scri
   Return fd0 j hd0; scri2M ^ scr2 top-k scores in Mg
```

Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (7s): 927-934

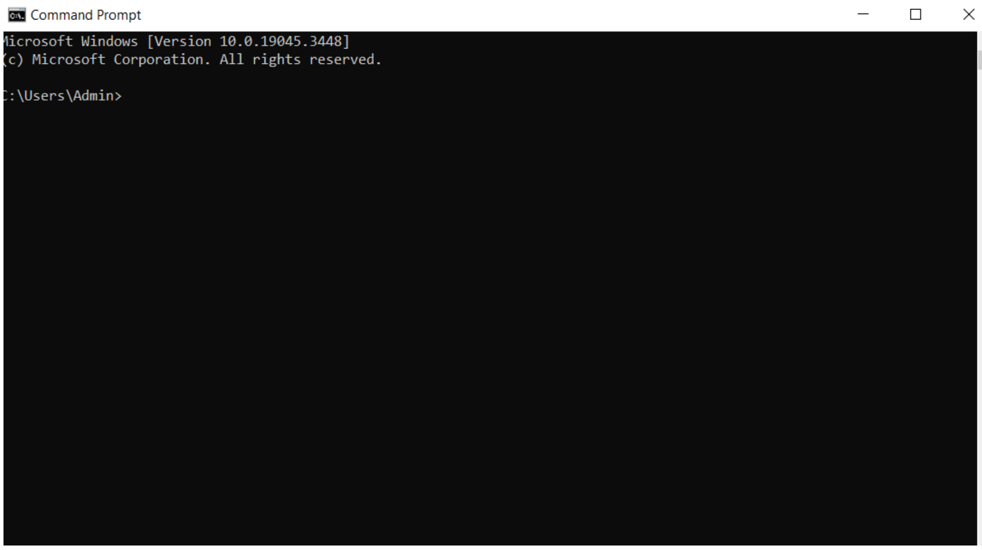


3.4.3.The Proposed Method As Follows:



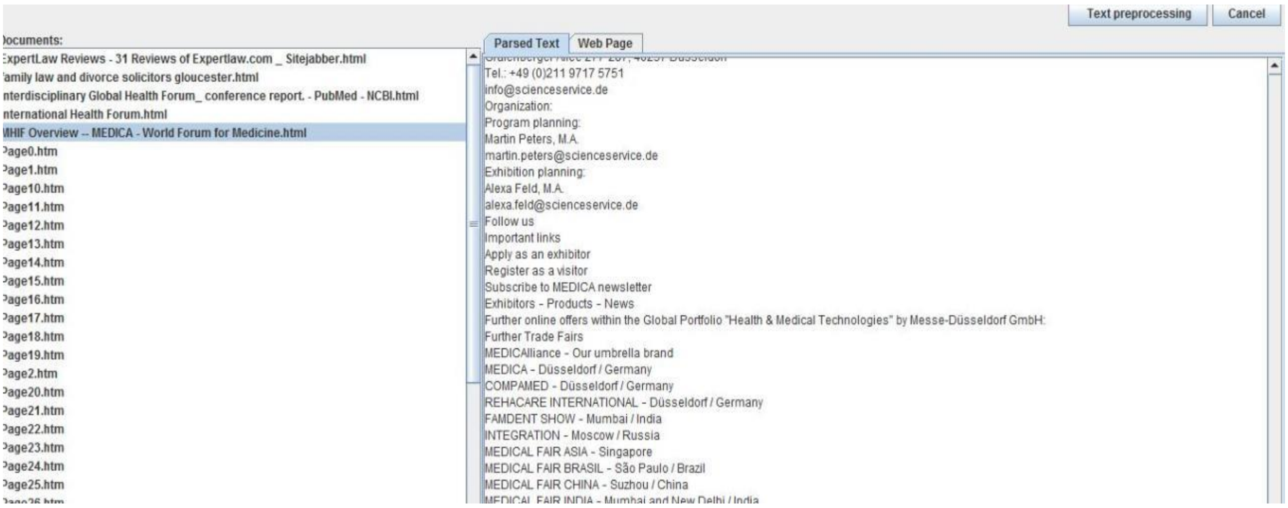
IV. RESULTS AND DISCUSSION:

OPEN THE COMMAND PROMPT AND RUN

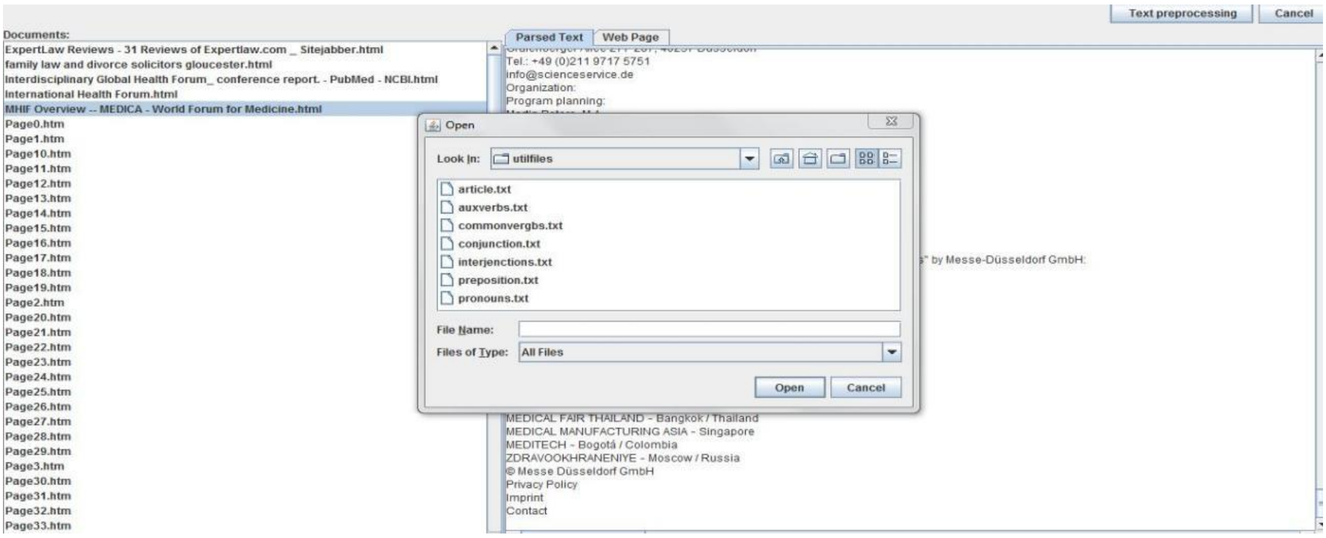


Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (7s): 927-934

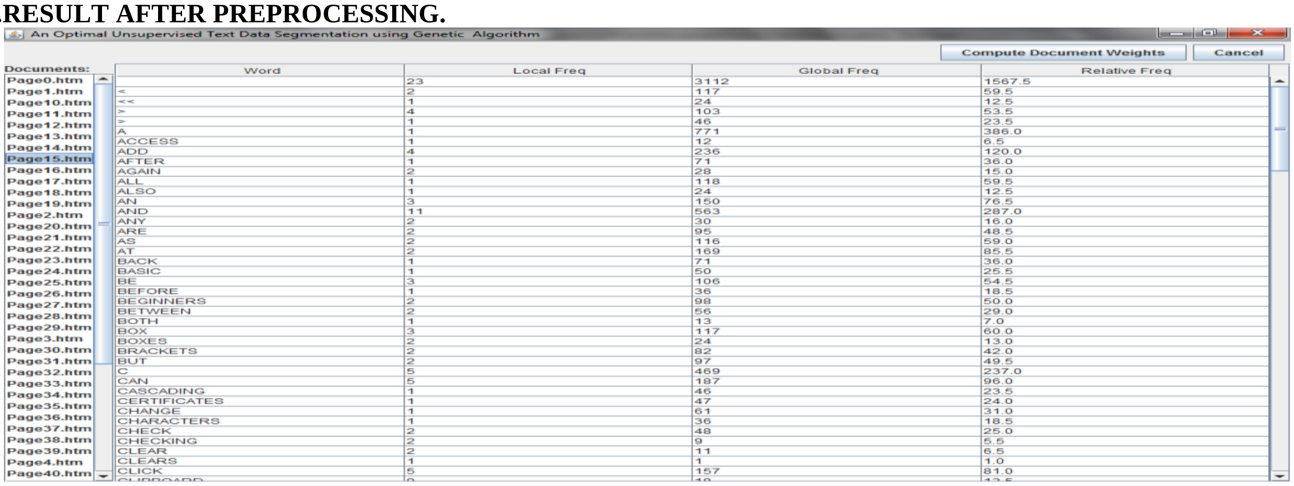
i.IT IS THE FIRST PAGE



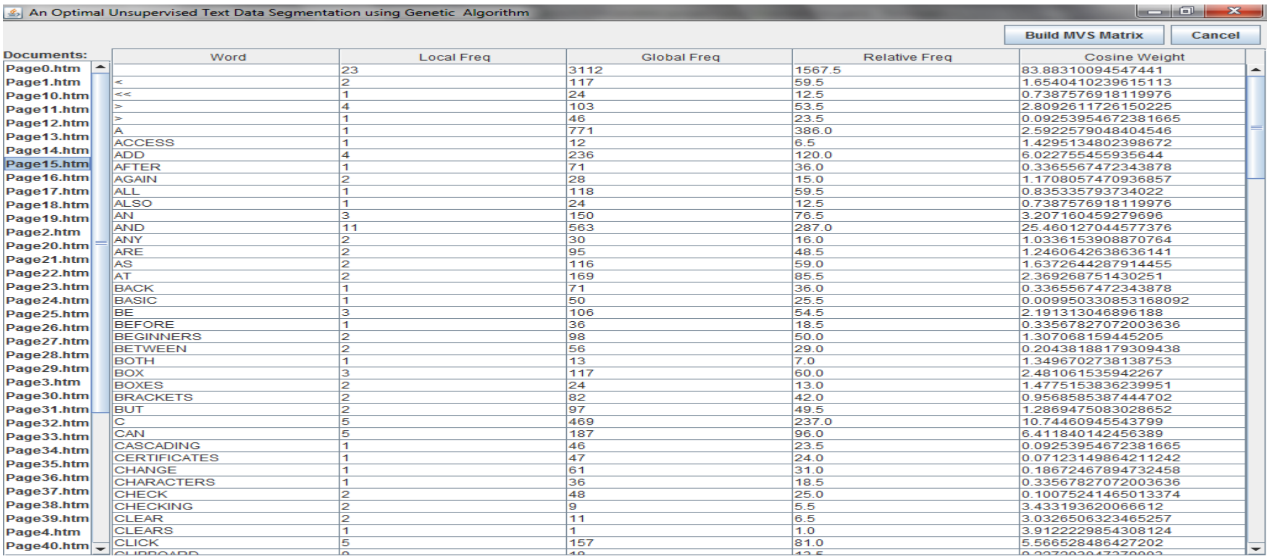
ii.NOW PRE PROCESSING IS TO BE PERFORMED



iii.RESULT AFTER PREPROCESSING.

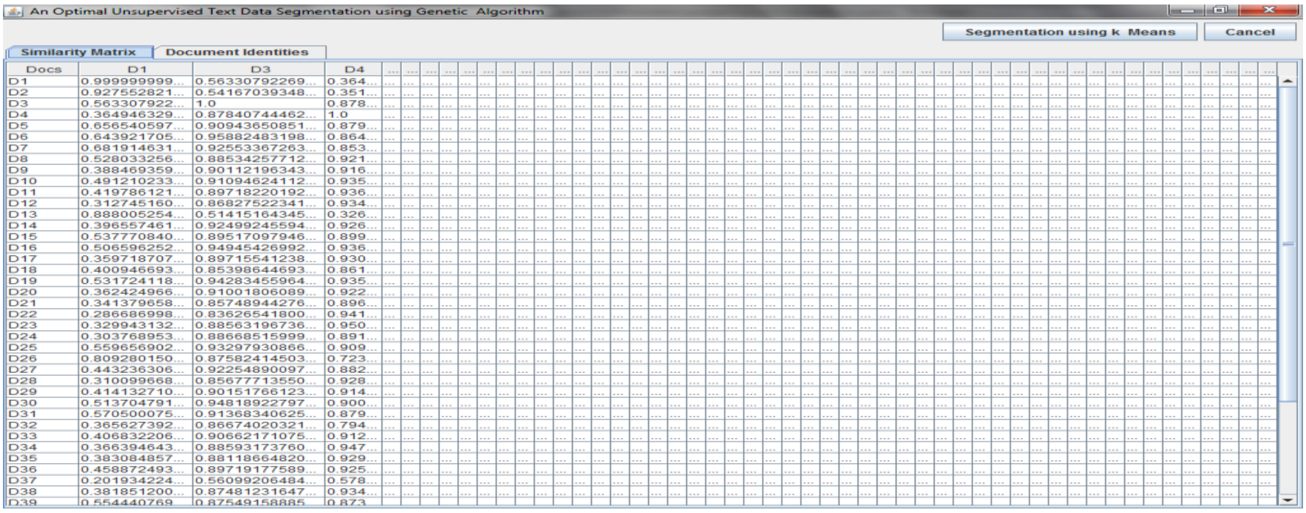


iv.COMPUTE DOCUMENT WEIGHTS



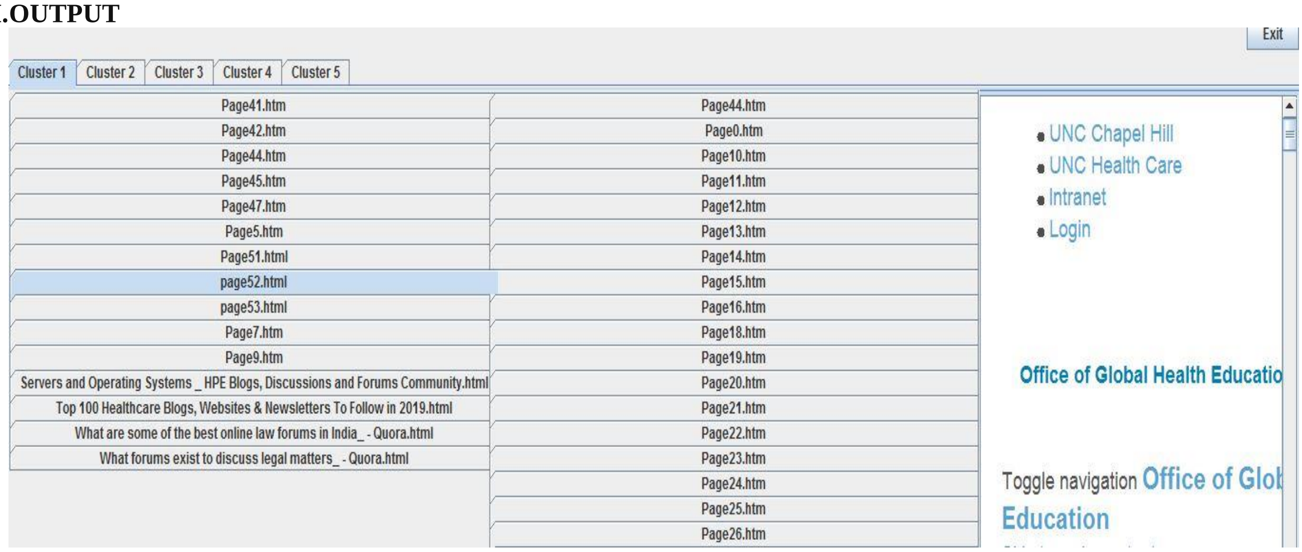
Word	Local Freq	Global Freq	Relative Freq	Cosine Weight
	23	3112	1567.5	83.88310094547441
<	2	117	59.5	1.6540410239615113
<<	1	24	12.5	0.7387576918119976
>	4	103	53.5	2.8092611726150225
>	1	46	23.5	0.09253954672381665
A	1	771	386.0	2.5922579048404546
ACCESS	1	12	6.5	1.4295134802398672
ADD	4	236	120.0	6.022755455935644
AFTER	1	71	36.0	0.3365567472343878
AGAIN	2	28	15.0	1.1708057470936857
ALL	1	118	59.5	0.835335793734022
ALSO	1	24	12.5	0.7387576918119976
AN	3	150	76.5	3.207160459279696
AND	11	563	287.0	25.460127044577376
ANY	2	30	16.0	1.0336153908870764
ARE	2	95	48.5	1.2460642638636141
AS	2	116	59.0	1.6372644287914455
AT	2	169	85.5	2.369268751430251
BACK	1	71	36.0	0.3365567472343878
BASIC	1	50	25.5	0.009950330853168092
BE	3	106	54.5	2.191313046896188
BEFORE	1	36	18.5	0.33567827072003636
BEGINNERS	2	98	50.0	1.307068159445205
BETWEEN	2	56	29.0	0.20438188179309438
BOTH	1	13	7.0	1.3496702738138753
BOX	3	117	60.0	2.481061535942267
BOXES	2	24	13.0	1.4775153836239951
BRACKETS	2	82	42.0	0.9568585387444702
BUT	2	97	49.5	1.2869475083028652
C	5	469	237.0	10.74460945543799
CAN	5	187	96.0	6.411840142456389
CASCADING	1	46	23.5	0.09253954672381665
CERTIFICATES	1	47	24.0	0.07123149864211242
CHANGE	1	61	31.0	0.18672467894732458
CHARACTERS	1	36	18.5	0.33567827072003636
CHECK	2	48	25.0	0.10075241465013374
CHECKING	2	9	5.5	3.433193620066612
CLEAR	2	11	6.5	3.0326506323465257
CLEARS	1	1	1.0	3.9122229854308124
CLICK	5	157	81.0	5.566528486427202

VI.BUILT VSM MATRIX



Docs	D1	D3	D4
D1	0.999999999...	0.56330792269...	0.364...
D2	0.927552821...	0.54167039348...	0.351...
D3	0.563307922...	1.0	0.878...
D4	0.364946329...	0.87840744462...	1.0
D5	0.656540597...	0.90943650851...	0.879...
D6	0.643921705...	0.95882483198...	0.864...
D7	0.681914631...	0.92553367263...	0.853...
D8	0.528033256...	0.88534257712...	0.921...
D9	0.388469359...	0.90112196343...	0.916...
D10	0.491210233...	0.91094624112...	0.935...
D11	0.419786121...	0.89718220192...	0.936...
D12	0.312745160...	0.86827522341...	0.934...
D13	0.888005254...	0.51415164345...	0.326...
D14	0.396557461...	0.92499245594...	0.926...
D15	0.537770840...	0.89517097946...	0.899...
D16	0.506596252...	0.94945426992...	0.936...
D17	0.359718707...	0.89715541238...	0.930...
D18	0.400946693...	0.85398644693...	0.861...
D19	0.531724118...	0.94283455964...	0.935...
D20	0.362424966...	0.91001806089...	0.922...
D21	0.341379658...	0.85748944276...	0.896...
D22	0.286686998...	0.83626541800...	0.941...
D23	0.329943132...	0.88563196736...	0.950...
D24	0.303768953...	0.88668515999...	0.891...
D25	0.559656902...	0.93297930866...	0.909...
D26	0.809280150...	0.87582414503...	0.723...
D27	0.443236306...	0.92254890097...	0.882...
D28	0.310099668...	0.85677713550...	0.928...
D29	0.414132710...	0.90151766123...	0.914...
D30	0.513704791...	0.94818922797...	0.900...
D31	0.570500075...	0.91368340625...	0.879...
D32	0.365627392...	0.86674020321...	0.794...
D33	0.406832206...	0.90662171075...	0.912...
D34	0.366394643...	0.88593173760...	0.947...
D35	0.383084857...	0.88118664820...	0.929...
D36	0.458872493...	0.89719177589...	0.925...
D37	0.201934224...	0.56099206484...	0.578...
D38	0.381851200...	0.87481231647...	0.934...
D39	0.554440769	0.87549158885	0.873

VII.OUTPUT



V. CONCLUSION

We have proposed a technique of Lexicon-based approach to solve sentiment analysis problems which it is expected to refine the sentiment analysis system using machine learning technique. For further study, we can concentrate on the learning of merging machine learning methods with opinion lexicon methods in demand to increase sentiment classification correctness and adaptability to a wide range of domain.

This paper presents an approach to calculate the semantic similarity between two words, sentences or paragraphs. A novel approach proposed by us for matching a reference post to the k most related posts in a collection. In our method , segmentation are done across the posts that convey similar with some intentions. We presented several experiments regarding the right segmentation criteria, the effectiveness of the segmentation algorithms and the formation of intention clusters that prove that a rather intuitive concept, that of the author intentions to communicate a certain message, can be effectively captured by an automated process.

Then extractive summarization is used to extract feature terms and the summarized sentences were ranked based on feature frequency of the posts, measuring the relatedness score after having distinguished the different

VI. REFERENCES

1. M. Chen, X. Jin, and D. Shen, "Short text classification improved by learning multi-granularity topics," in IJCAI, 2011, pp. 1776–1781.
2. J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," in Proceedings of the 28th ACM SIGIR Conference, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 617–618.
3. M. Chen, X. Jin, and D. Shen, "Short text classification improved by learning multi-granularity topics," in IJCAI, 2011, pp. 1776–1781.
4. J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," in Proceedings of the 28th ACM SIGIR Conference, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 617–618
5. S. Robertson, S.Walker, and M. Hancock-Beaulieu, "Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track," TREC '98, pp. 199–210, 1998.
6. .G. Salton, A. Singhal, C. Buckley, and M. Mitra, "Automatic text decomposition using text segments and text themes," in ACM Hypertext, 1996, pp. 53–65.
7. S. Louvigne, N. Rubens, F. Anma, and T. Okamoto, "Utilizing social media for goal setting based on observational learning," in ICALT, 2012, pp. 736–737.
8. .K.Wang, Z. Ming, and T. Chua, "A syntactic tree matching approach to find similar questions in community QAservices," in ACM SIGIR, 2009, pp. 187 – 194.
9. K. Jones, C. Van Rijsbergen, B. L. Research, and D. Department, Report on the Need for and Provision of an Ideal Information Retrieval Test Collection, ser. British Library Research and Development reports, 1975.
10. J. Kekalainen, "Binary and graded relevance in ir," Inf. Processing & Management, vol. 41, no. 5, pp. 1019 – 1033,2005.
11. Z.-Y. Ming, T.-S. Chua, and G. Cong, "Exploring domain specific term weight in archived question search," in Proceedings of he 19th ACM CIKM, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 1605–1608.
12. H. Wen, W. Zhongyuan, W. Haixun, Z. Kai, and Z. Xiaofang, "Short text understanding through lexical-semantic analysis," in IEEE ICDE, 2015
13. J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," in Proceedings of the 28th ACM SIGIR Conference, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 617–618.