A Novel Hybrid Approach on Document Clustering and and its Importance on Sentiment Analysis and Opinion Mining

Sandeep Singh Thakur [1], Aman Kumar [2]

[1]M.Tech Scholar, Department of Computer Science and Engineering, LRIET,Solan, H.P.

[2]Assistant Professor, Department of Computer Science and Engineering, LRIET,Solan, H.P.

Abstract — Document clustering, one of the conventional information mining procedures, is an unsupervised learning worldview where clustering techniques attempt to distinguish inherent groupings of the text documents, so that a set of clusters is produced in which clusters display high intra-group comparability and low inter-cluster closeness. The significance of document clustering rises up out of the huge volumes of printed records being made. Here in this paper, we have implemented a hybrid approach in clustering texts using fuzzy c-means and PSO(particle swarm optimization). It shows better results that the individual performances of the algorithms.

Keywords: Text mining, Opinion mining, Sentiment analysis

Introduction

Clustering is considered as one of the most important unsupervised learning problem. In the clustering process, objects are organized into groups of similar members. Hence, a cluster is a collection of objects which are similar to each other but dissimilar to the objects of other clusters. A text clustering divides a collection of text documents into different category groups so that documents in the same category group describe the same topic. A text clustering is the elemental function in the process of text mining. Automated document processing can include operations such as document comparison, document categorization, and document selection. Document clustering has very significant uses in many areas of data mining and information retrieval. Clusters of documents are generated automatically on the collection of documents[9].

In additional method of document clustering, single, unique, or compound words of the document set are used as features. But the additional method does not consider semantic relationships into account. The problems such as the synonym problem and the polysemous problem exist in the traditional method; therefore, a bag of original words cannot represent the exact content of a document and cannot produce meaning l clusters. Therefore, to improve document clustering, there is a need of clustering techniques that also consider meaning of words into clustering process[2].

One way to solve this problem is to enhance document representation with the background knowledge represented by ontology. Another way is to use Latent Semantic Analysis (LSA) technique. Polysemy and synonym problems are fundamental problems in unsupervised learning techniques[5]. A synonymous term maps to the same concept as different words in the document. A polysemous term is a term that has multiple, disjoint meaning. These two problems can be solved by using LSA in additional keyword based retrieval. The main use of this LSA technique is to illustrate the core semantic structure of a document by doing their representation in high dimensional space..

Semantic document clustering has an important benefit of being able to remove irrelevant documents by recognizing conceptual mismatches. Word Sense Disambiguation (WSD) is also used to resolve the ambiguity by pointing which concept is represented by a word or a phrase in a context. Use of ontology makes it easier to identify related concepts and their linguistic representatives given a key concept, whereas LSA tries to uncover the hidden conceptual relationships among the words and phrases as per their linguistic usage patterns. Word Net 2.0 is a lexical database (a collection of words) that can be

used to get information of words or phrases. It is used as knowledge base in automatic text analysis and artificial intelligence. Word Net 2.0 contains many set of synonym words of same concept and their relationships with different syntax.

1.2 OVERVIEW OF TEXT MINING

In text mining, it t is corresponding to content examination which suggests do content advancing and in the midst of the progress make data. It can do packing, arrange, gathering, association examination, dispersal examination and example envisioning to a broad number of substance on the document. These years, text mining has been a fundamental field in orders, for instance, data mining, machine learning, information recuperation, and so forth.

1.3 Text Mining Process

- Gathering unstructured information from various sources.
- Pre-processing and cleansing operations are performed to recognize and evacuate bug. procedure make a point to catch the genuine substance of content accessible and is performed to remove stop words stemming (procedure of recognizing the foundation of certain word) and ordering the information
- Processing and controlling operations are connected to review and further clean the informational collection via programmed handling.
- Pattern investigation is executed by Management Information System (MIS).

Data prepared in the above steps are utilized to concentrate on profitable and valuable data for compelling and decision making process and pattern investigation.

Figure 1.1 Text Mining Process

The advancement of technology and innovation in recent years, particularly the gigantic prevalence of the social media, enormous data rises and individuals utilize different instruments to extricate the learning they require from monstrous data assets. With respect to different un-organized or semi-organized content information, for example, books, different content information of messages, online journals, micro blogging services and messages which are delivered , daily paper and social media explore whose trademark are information decentralized, auxiliary assorted and hard to extensively examine. So endeavors create different content mining apparatuses, for example, Text Analyst, Word Stat, PolyAnalyst, ICrossReader, etc.

1.4 Text Mining Tools

'Text Analyst' is a kind of programming made by Megaputer Intelligence, Inc to oversee content and semantic examination. Appear differently in relation to other substance examination and information investigate structures, the rule promotion vantage of Text Analyst is it can totally thusly isolate content semantic framework without experts' advanced made specific point word reference. In view of neural framework advancement, Text Analyst executes any application fields' substance semantic examination by using thus made substance semantic framework.

Text Analyst for the most part has the accompanying capacities:

- **Filtering text meaning**: Filtering text meaning can shape and yield redress substance or corpus semantic framework which can rapidly address substance's hugeness and can a base to do help examination. Corpus course.
- **Structuring topics**: Structuring topics recognize the most basic thoughts from a semantic framework; changing over the semantic framework into a tree in sliding solicitation of noteworthiness of embedding subject by overcoming those associations which addresses feeble ties and using substitution of deviant relationship into coordinate relationship remembering the true objective to reveal the level dynamic framework relationship in content topics of re-chase
- **Clustering**: additionally expelling connections of subject structure quality that is beneath a specific edge esteem so that a corpus' joint topical

structure can be separated into a few sections that speaks to significant autonomous subjects; at that point single record can be appropriated into various theme bunches keeping in mind the end goal to improve document clustering in corpus.

- **Text abstracting**: Text abstracting can judge and research a singular sentence in the substance by using semantic frameworks; the more number of basic semantic ramifications in sentences and the more grounded association between thoughts, the higher criticalness of sentence's own semantic significance

- **WORDSTAT**: Word Stat is a remarkably delineated substance analyzing programming for analyzing content data, for instance, journal articles, masterful works, interviews, open request answering, electronic exchanges and so on which made by Provalis Research Corp.

Wordstat has many capabilities including the followings:

i) Supporting diverse tongues: This consolidates the going with: English, French, German, Portuguese, Spanish, Italian, Chinese (standard and enhanced); supporting Latin-1, twofold byte character set and UTF-8 code.

ii) Text Processing: word frame dissecting by lexicon changing, calling outside content preprocessing; utilizing client characterized rejection rundown to specifically prohibit pronouns and conjunctions et cetera; utilizing the current or client characterized word reference to group words or expressions; arranging words in view of Boolean and Proximity rules, constraining content investigation or barring remarks and notes area as indicated by content.

iii) Feature extraction: evacuating advancement things, things, association names and essential wrong spellings by word pioneer; discovering rehashing expressions and expressions by expression pioneer. **iv) Automatic text classification:** owning archive arrange machine learning computations; giving versatile segment decision to the best quality subset's customised picks; giving various check procedures; gathering model can be secured and so forth.

Literature Review

Recently, document clustering became very useful for every single aspect in the Information technology industry.

Sandip D Mali et al [1] proposed another framework called Sent iView which a vocabulary based approach for sentiment investigation. They have gotten high accuracy because of pre-processing and expulsion of non-opinion tweets from data.

Calvin and Johan Setiawan [2] proposed a model where sentiment extremities of Twitter surveys are measured utilizing Naive Bayes classifier strategy. The model demonstrates a promising come about on characterizing the ubiquity in light of consume satisfaction and along these lines characterizing the best supplier to be utilized.

Sanjana Woonna et al [3] proposed a framework that examinations tweets into three classifications which are positive, negative and neutral utilizing supervised learning approach After the execution, the outcomes demonstrated which viewpoints individuals like or aversion and how feelings on motion pictures changes over a timeframe.

Zhao Jianqiang et al [4] talked about the impacts of content pre-processing strategy on sentiment characterization execution in two sorts of classification task, and summed up the grouping exhibitions of six pre-processing techniques utilizing two feature models and four classifiers on five Twitter datasets. T

Aashutosh Bhatt et al [5] proposed a framework that plays out the classification of customer reviews after by discovering estimation of the surveys. A rule based extraction of item highlight assumption is likewise done. The outcomes demonstrated that characterization of reviews alongside sentimental investigation expanded the exactness of the framework turn provides accurate reviews to the user.

3.1 PROBLEM FORMULATION

There are several clustering techniques

1. Connectivity-based clustering (hierarchical clustering)

2. Centroid-based clustering

3. Distribution-based clustering
4. Density-based clustering

which are used for statistical analysis of data. Most prominently PSO (Particle Swarm Optimisation) and FCM (Fuzzy C-means) are used in data clustering. But PSO algorithm is easy to fall into local optimum in high-dimensional space and has a low convergence rate in the iterative process. The computational complexity is accepted when it is applied to solve the high-dimensional and complex problems.

On the other hand FCM has disadvantages like

1) A priori specification of the number of clusters.

2) With lower value of β we get the better result but at the expense of more number of iteration.

3) Euclidean distance measures can unequally weight underlying factors.

So we are proposing a hybrid algorithm which would be able to overcome the disadvantages of the above mentioned algorithms.

The FCM algorithm is more speedy than the PSO algorithm as it needs fewer function evaluations, but it usually falls into local optima. Here in our paper, we have integrated he FCM algorithm with PSO algorithm to form a hybrid clustering algorithm.

FCM–PSO applies FCM to the particles in the swarm and in every number of iteration, the fitness value of each particle is improved.

It can be described as below:

DATASET DESCRIPTIOPN

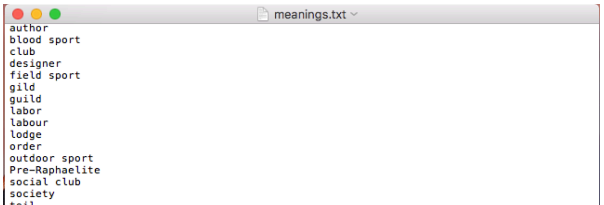
Repository link of our dataset is given below:

http://mlg.ucd.ie/datasets/bbc.html

Description of our dataset can given as:

- Consists of 737 documents from the BBC Sport website corresponding to sports news articles in five topical areas from 2004-2005.
- Class Labels: 5 (athletics, cricket, football, rugby, tennis)

RESULTS



4.1.1 GUI (graphical user interface) of our proposed method

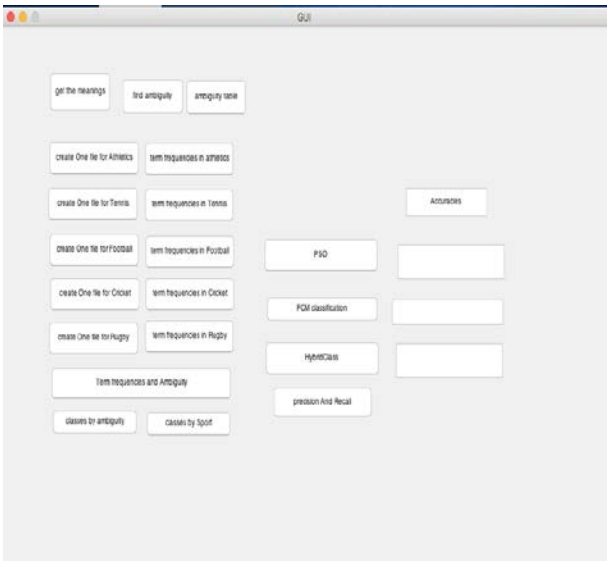
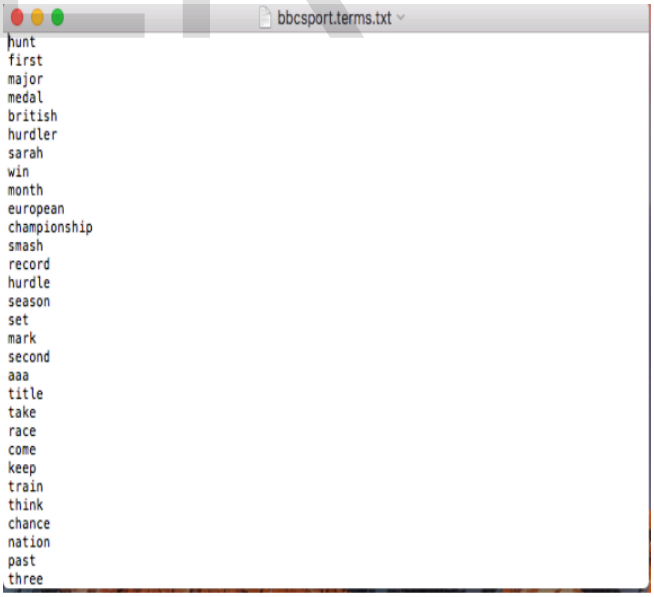


Figure 4.1 Graphical User Interface

4.1.2 Once we click on the 'get the meanings' button, the text file containing the terms for meaning search, is imported and each term is searched one by one in the repository. Snapshot of the terms file is given below:



4.1.3 A text file named 'meanings.txt' is created and the meanings of the corresponding terms are stored in it.

A snapshot of the meanings.txt file is given below:

Figure 4.3 Meanings.txt File

The API used to find the meanings is:

words.bighugelabs.com

4.1.4. We are going to find the ambiguity of the terms inside the meanings.txt file. That will give us the occurrence frequency of the terms inside the meanings file.

We are going to find the occurrences of each term inside each of the folders of the bbcsports news documents.

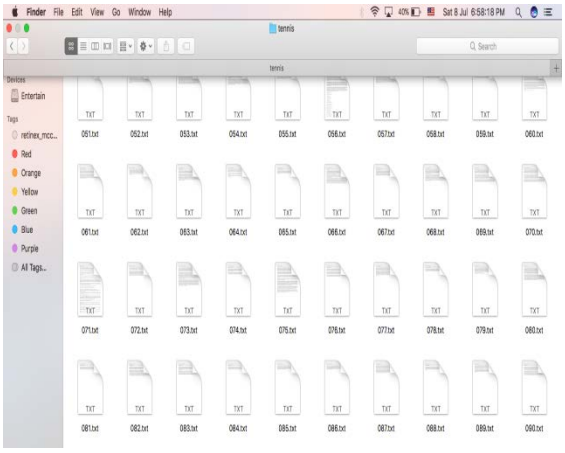
Highest occurrences of each term inside the tennis, rugby, football, cricket and athletics are needed to find and they will determine in which category the term belongs to. The clustering algorithms will determine the further accuracy of the terms and their categories.

Table 4.1 Ambiguity table of the terms

Ambiguity	Terms
21	'hunt'
21	'first'
0	'major'
7	'medal'
0	'british'
0	'hurdler'
0	'sarah'
21	'win'
14	'month'
0	'european'
7	'championship'
14	'smash'
49	'record'
0	'hurdle'
0	'season'
28	'set'
49	'mark'
35	'second'
0	'aaa'
21	'title'
7	'take'
21	'race'

4.1.5 Inside each of the folder, there are number of text files. All of the files are combined into a single file for simplification of the computation. Suppose, in Tennis folder, there are 100 files. Each of them is combined into a single file named Tennis.txt.

The Figure 4.4 shows all the files inside the folder Tennis.



4.1.6 Combined document Tennis.text

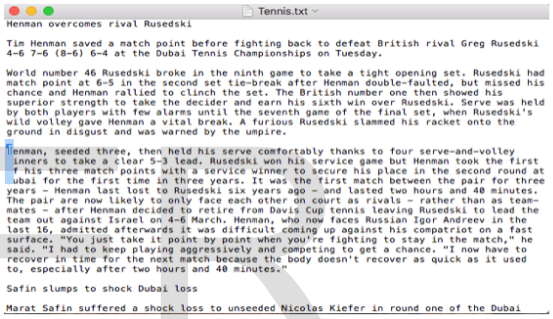


Figure 4.5 Combined document Tennis.text

4.1.7 Combined document Athletics.txt

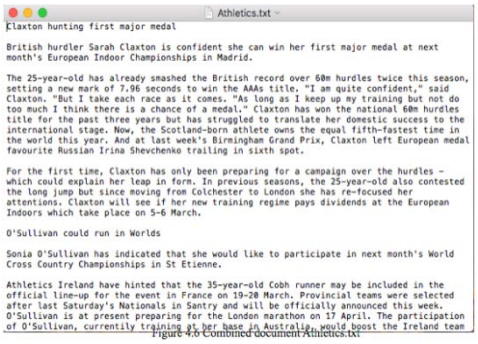


Figure 4.6 Combined document Athletics.txt

4.1.8 Combined document of the Football.txt

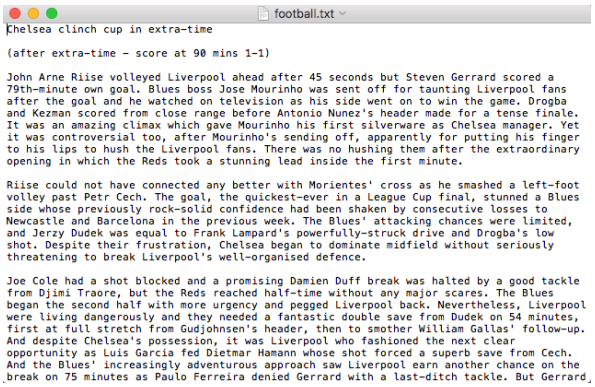
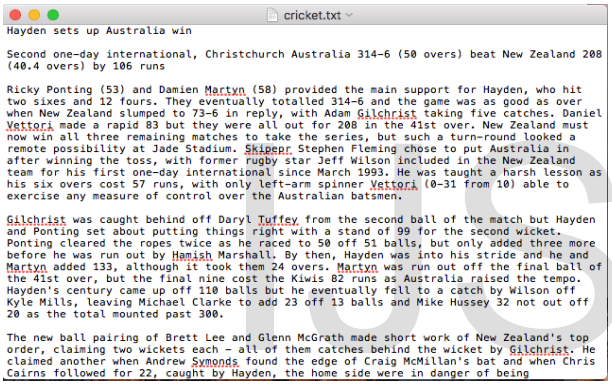


Figure 4.8 Combined document of the Football.txt

4.1.9 Combined document of the cricket.txt



4.1.10 Combined document of the rugby.txt

Figure 4.10 Combined document of the rugby.txt

4.1.11 When we click on Term frequency and ambiguity button, we get a table named Ambiguity And Frequencies and it looks like as below:

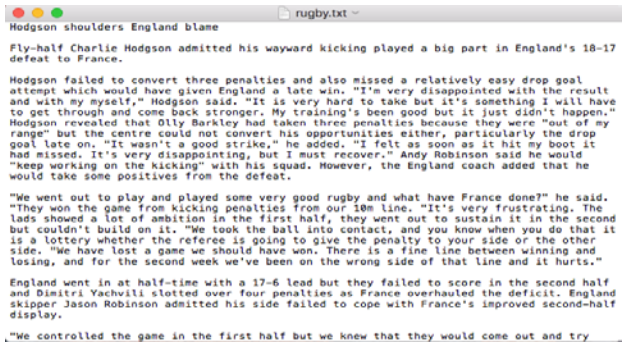


Table 4.2 Ambiguity and Frequencies

4.1.12 Classes of each term, either ambiguous or non-ambiguous

AmbiguiyAndFrequencies =

Ambiguity	Terms	Athletics	Tennis	Football	Cricket	Rugby
21	'hunt'	0	0	0	4	0
21	'first'	296	296	81	179	120
0	'major'	16	16	9	5	17
7	'medal'	2	2	1	0	2
0	'british'	0	0	0	0	0
0	'hurdler'	0	0	0	0	0
0	'sarah'	0	0	0	0	0
21	'win'	424	424	133	107	191
14	'month'	42	42	29	15	11
0	'european'	0	0	0	0	0
7	'championship'	8	8	1	0	13
14	'smash'	2	2	4	6	0
49	'record'	14	14	10	20	11
0	'hurdle'	0	0	0	0	0
0	'season'	56	56	54	11	49
28	'set'	344	344	21	55	43
49	'mark'	26	26	12	21	11
35	'second'	186	186	44	63	69
0	'aaa'	0	0	0	0	0
21	'title'	136	136	14	0	11
7	'take'	136	136	52	55	55
21	'race'	26	26	7	19	10
7	'come'	110	110	39	54	59
0	'keep'	26	26	64	41	17

Table 4.3 Each term, either ambiguous or non-ambiguous

'Ambiguous'	'hunt'
'Ambiguous'	'first'
'Nonambiguous'	'major'
'Ambiguous'	'medal'
'Nonambiguous'	'british'
'Nonambiguous'	'hurdler'
'Nonambiguous'	'sarah'
'Ambiguous'	'win'
'Ambiguous'	'month'
'Nonambiguous'	'european'
'Ambiguous'	'championship'
'Ambiguous'	'smash'
'Ambiguous'	'record'
'Nonambiguous'	'hurdle'
'Nonambiguous'	'season'
'Ambiguous'	'set'
'Ambiguous'	'mark'
'Ambiguous'	'second'
'Nonambiguous'	'aaa'
'Ambiguous'	'title'
'Ambiguous'	'take'
'Ambiguous'	'race'
'Ambiguous'	'come'
'Nonambiguous'	'keep'
'Ambiguous'	'train'
'Nonambiguous'	'think'

4.1.13 Terms by its classes.
Here, 1= Athletics, 2 = Tennis, 3 = Football, 4 = Cricket, 5 = rugby

Table 4.4 Terms with its classes

Terms	Classes
'hunt'	4
'first'	5
'major'	1
'medal'	5
'british'	5
'hurdler'	5
'sarah'	5
'win'	5
'month'	5
'european'	5
'championship'	1
'smash'	4
'record'	4
'hurdle'	5
'season'	5
'set'	5
'mark'	5
'second'	5
'aaa'	5
'title'	5
'take'	5
'race'	5
'come'	5
'keep'	3

4.1.14 Now we are implementing the classification methods. At first, we are implementing PSO (particle swarm optimization).

i) The predicted class values of terms after implementing PSO are given below:

1	4
1	5
1	1
1	5
1	5
1	5
1	5
1	5
1	5
1	5
1	1
1	4
1	4
1	5
1	5
1	5
1	5
1	5
1	5
1	5
1	5
1	5
1	3
1	1

ii) Confusion Matrix of PSO Classification

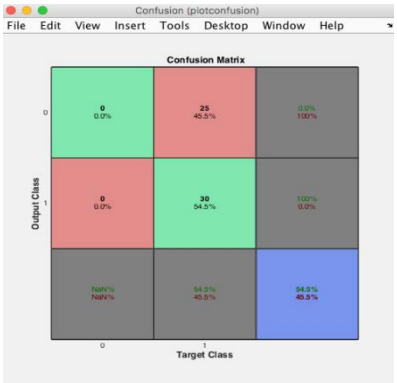
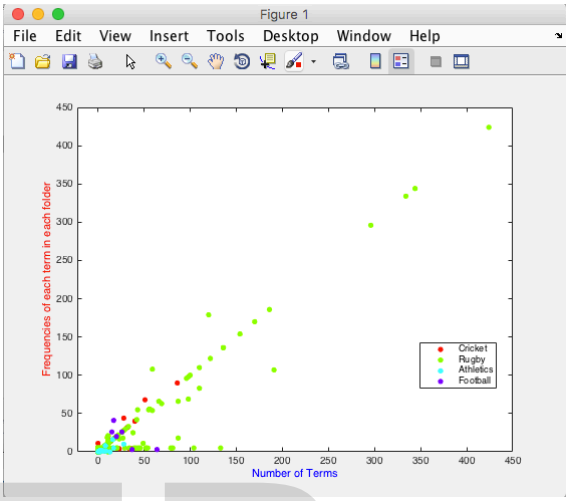


Figure 4.11 Confusion Matrix of PSO

iii) Scattered plot of the terms using PSO classification

4.1.15 Now we are implementing FCM (Fuzzy C-means).



i) Predicted values of Fuzzy C-means:

1	4
1	5
1	1
1	5
1	5
1	5
1	5
1	5
1	5
1	5
1	5
1	1
1	4
1	4
1	5
1	5
1	5
1	5

ii) Confusion matrix of Fuzzy C-means.

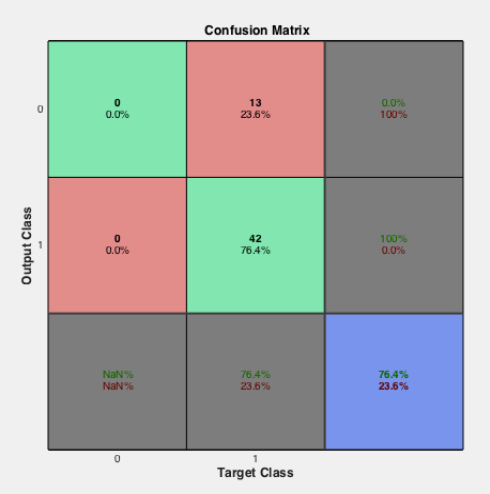


Figure 4.13 Confusion matrix of Fuzzy C-means

iii) Scatter plot of Fuzzy C-means

4.1.16 After implementing PSO and FCM classification, we have integrated the FCM algorithm with PSO algorithm to form a hybrid clustering algorithm.

i) Predicted class values terms using hybrid approach

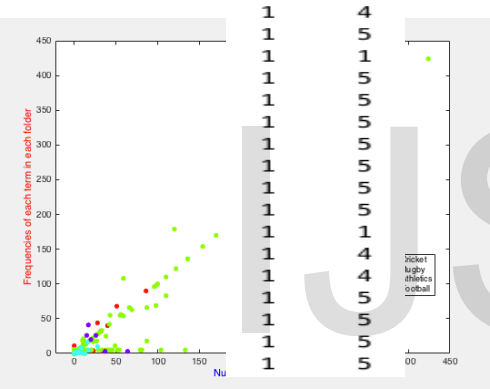
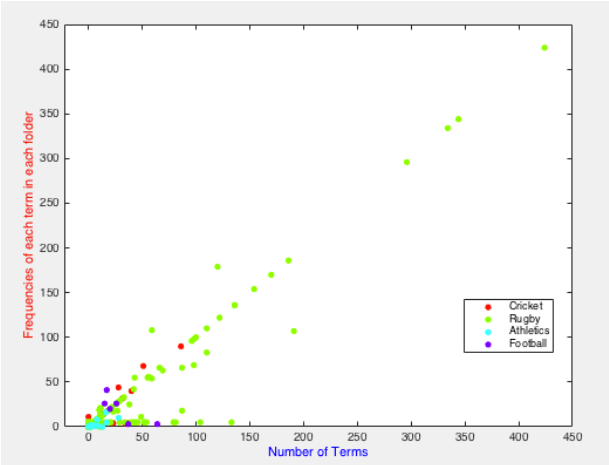


Figure 4.15 Confusion matrix of hybrid classification



	Accuracies
PSO	54.5455
FCM classification	76.3636
HybridClass	87.2727

iii)

Scatter plot hybrid classification

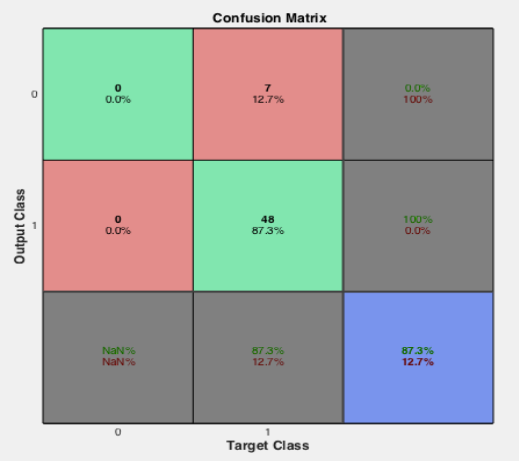


Figure 4.16 Scatter plot hybrid classification

4.17 Accuracy table of each classification:

Figure 4.17 Accuracy table of each classification

From the results discussed above, we can find it out that our hybrid approach gives much better classification results then PSO and FCM.

Conclusion

In this paper, at first we discussed text mining and its relation with sentiment analysis and opinion mining. We then discussed some recent works on text mining. We stated our approach and methodology. The results we achieved claims that the hybrid approach we used to classify the documents showed very good results regarding classification accuracy.

In future, we can focus on other genetic algorithms for hybridization of clustering algorithms.

References

[1] J. Deonna, F. Teroni, The emotions: A philosophical introduc- tion, Routledge, 2012.
[2] B. J. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: Tweets as electronic word of mouth, Journal of the American Society for Information Science and Technology 60 (11) (2009) 2169–2188. doi:10.1002/asi.21149.

[3] P. Melville, W. Gryc, R. D. Lawrence, Sentiment analysis of blogs by combining lexical knowledge with text classification, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 1275–1284.
[4] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, in: Mining Text Data, Springer U.S., 2012, pp. 415– 463. doi:10.1007/978-1-4614-3223-4_13.

[5] B.Pang,L.Lee,Opinionminingandsentimentanalysis,Foun- dations and Trends in Information Retrieval 2 (12) (2008) 1– 135. doi:10.1561/1500000011.

[6] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intelligent Sys- tems 28 (2) (2013) 15–21. doi:10.1109/MIS.2013.30.
[7] S. Asur, B. A. Huberman, Predicting the future with social media, in: IEEE/WIC/ACM

International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010, pp. 492–499. doi:10.1109/WI-IAT.2010.63.

[8] C. Clavel, Z. Callejas, Sentiment analysis: From opinion mining to human-agent interaction, IEEE Transactions on Affective Computing (2015) 74–93doi:10.1109/TAFFC.2015. 2444846.

[9]. M. D. Munezero, C. S. Montero, E. Sutinen, J. Pajunen, Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text, IEEE Transactions on Affective Computing 5 (2) (2014) 101–111. doi:10.1109/TAFFC.2014.2317187.

[10]. L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis, in: ACM International Conference on Mul- timodal Interfaces (ICMI), New York, New York, USA, 2011, p. 169. doi:10.1145/2070481.2070509.

[11].M. W ̈ollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, YouTube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intelligent Systems 28 (3) (2013) 46–53. doi:10.1109/MIS.2013.34.

[12]. J. R. Fontaine, K. R. Scherer, E. B. Roesch, P. C. Ellsworth, The world of emotions is not two-dimensional, Psychological Science 18 (12) (2007) 1050–1057. doi:10.1111/j.1467-9280. 2007.02024.x.

[13]. K. R. Scherer, A. Schorr, T. Johnstone, Appraisal processes in emotion: Theory, methods, research, Oxford University Press, 2001.

[14]. D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: ACM International Conference on Multimedia, 2013, pp. 223–232. doi:10.1145/2502081.2502282.