A Survey on Sentiment Based Text Categorization

[1]Doshi Bhaven Kusumkant, [2]Hemant Kumar Gupta
[1]Student, [2]Assistant Professor
Lakshmi Narain College of Technology and Science(RIT), Indore.

Abstract - the data mining is a way of data processing for obtaining the information from raw data. These techniques are used in analysis of diverse nature of data i.e. image, text, video, audio and more. In our observation, according to the type of data the utilization of algorithms are changed. In this presented work the applicability of unsupervised learning techniques over the text data is tried to explore. In this paper, the technique of text data analysis based on the hidden sentiments are tried to study. In this context first the recently developed approaches of sentiment analysis techniques are studied and in further using the existing techniques and methods a new text categorization model is proposed for design and implementation. Finally the future work is also proposed in this paper.

Keywords - NLP, unsupervised learning, text clustering, comparison, text mining, FCM algorithm.

I. INTRODUCTION

Text contains the emotions, author usages the text to express their emotions and feelings. Therefore the text data analysis is an essential part of natural language processing (NLP) [1]. Now in these days a number of platforms are exits where the users post the text to convey the emotional and other kinds of messages. Such kind of text data analysis is known as sentiment based text analysis. Basically that technique is derived from two concepts text mining and natural language processing [2]. In this presented work the main aim is to explore the scope of text mining in the domain of sentiment analysis using the unsupervised learning approach.

The text data is a complex type of data which is not in regular manner. There are various complexities in processing the normal text data among the size, length, language and the noise is the key issues in text data processing. In addition of the techniques which are employed to recover the features of text data is also need to be very accurate to obtain the essential features [3]. In literature a number of techniques for accurately process the text is available. Based on the obtained conclusion from the literature we tried to provide a social media text categorization technique. The categorization process is basically unsupervised learning technique for data analysis. But due to the unstructured format of data the data mining clustering algorithms are not directly employable [4]. In this context a complete text categorization system is introduced in this work. That takes input the text (i.e. social media text), and process it to find the sentiment labels by using the unsupervised learning approach. Therefore the twitter dataset and the FCM algorithm [5] is studied in this work for designing the proposed system.

This section involve the overview of the proposed work, in next section the essential key terms are explained.

II. BACKGROUND

This section provides the overview of the key terms that are frequently used in this paper. In addition of we tried to explain the context of the used keywords:

1. Data mining

The data mining is a technique of evaluation of data. The data is evaluated for finding the specific kinds of patterns that can be used for some application. For example the student performance data is analyzed for projecting the future growth of student. In this work the text data mining is studied which also known as text is mining [6].

2. Text mining

When the data mining techniques and algorithm are applied on text data for finding significance knowledge such kind of mining is known as text mining. The use of text mining is found in various applications, such as digital library, security, review analysis and more [7].

3. Sentiment classification

The different platform users when create a text post, the author express their emotions using the posts. Therefore the text analysis is also performed for identifying the author's mood or emotions. Such kind of text mining is known as the sentiment based text classification. This kind of text data analysis is now in these days used for customer review analysis and enhancing the CRM systems [8].

4. Document clustering

The clustering is an unsupervised technique of data mining. When this technique is used to obtain the groups of similar pattern documents then the document clustering is performed on data. This kind of data processing is known as the document clustering [9].

5. Unsupervised learning

The unsupervised learning techniques are those which are directly applicable on the data to categorize the patterns based on their internal similarity or distance. These similarity or distance matrixes are used to make decision about the class label assignment of data instances. These techniques are not required to have the class labels for training the algorithm [10].

6. FCM (Fuzzy C-means)

Fuzzy c-means (FCM) is a technique for data clustering that allows the data instances to have a place with multiple clusters. The fuzzy c-means algorithm usages the fuzzy membership functions in place of distance matrix. In addition of that the algorithm works for minimization of an objective function. The objective function is given here as J_m [11]:

$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \parallel x_i - c_j \parallel , \qquad 1 \le m < \infty$

Where m is a real number higher than 1, u_{ij} is the degree of membership between x_i in centroid j, x_i is the i[th] value or instance of data set, c_j is the centroid of the cluster, and ||*|| is any norm expressing the similarity between any measured data and centroid.

Fuzzy partitioning is performed using an iterative optimization process for minimizing the objective function J_m, with the update of membership u_{ij} and centroid c_j using

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left[\frac{\parallel x_i - c_j \parallel}{\parallel x_i - c_k \parallel} \right]^{\frac{2}{m-1}}}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

The algorithm is composed of the following steps:

1: Initialize U=[u_{ij}] matrix, $U^{(0)}$

2: At k-step: calculate the centers vectors $C^{(k)}$=[c_j] with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3: Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left[\frac{\parallel x_i - c_j \parallel}{\parallel x_i - c_k \parallel} \right]^{\frac{2}{m-1}}}$$

4: If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$, then STOP; otherwise return to step 2.

Table 2.1 fuzzy C-means clustering algorithm

III. LITERATURE SURVEY

This section involves the essential contributions in the domain of text categorization and identifying the sentiment classes.

Relatively work has focused on learning representations for clustering. ***Junyuan Xie et al [12]*** propose Deep Embedded Clustering (DEC) method. That method concurrently learns feature representations and cluster assignments. For this task neural network is employed. DEC learns a mapping from data space to a lower-dimensional feature space in which it iteratively optimizes a clustering objective. The experimental evaluations on image and text corpora show significant improvement over other methods.

One of the most well-known binary (discrete) versions of artificial bee colony (ABC) algorithm is the similarity measure based discrete ABC, which proposed to deal with the un-capacited facility location (UFLP) problem. The discrete artificial bee colony simply depends on measuring similarity between the binary vectors through Jaccard coefficient. It is accepted as one of the simple, novel and efficient binary variant of ABC. ***Celal Ozturk et al [13]*** offers a new solution generation mechanism of discrete ABC is enhanced using all similarity cases using genetically inspired components. The superiority of given algorithm is simulated by comparing it with the basic discrete ABC algorithm, binary particle swarm optimization, genetic algorithm in dynamic (automatic) clustering, in which the number of clusters is determined automatically. Not only evolutionary computation based algorithms, but also classical approaches such as fuzzy C-means and K-means are employed to put forward the effectiveness of given approach. The results indicate that the discrete ABC with the enhanced solution generator component is able to reach more valuable solutions than the other algorithms.

Nameirakpam Dhanachandra et al [14] studied Image segmentation is the classification of an image into different groups. There are one of the most popular methods is k-means clustering. K-means algorithm is used to segment interest area from the background. But before applying K-means algorithm, first partial stretching enhancement is applied to improve quality of the image. Subtractive clustering is used to generate initial centers and these centers are used in k-means algorithm for the segmentation. Then finally medial filter is applied to the segmented image to remove any unwanted region from the image.

Various problems from different areas have been effectively solved by using FCM. But, for efficient use of the algorithm in diversified applications, some modifications or hybridization are needed. A survey on FCM and its applications has been carried out by ***Janmenjoy Nayak et al [15]*** to show the efficiency and applicability in a mixture of domains. Another intention of this survey is to encourage new researchers to make use of this simple algorithm in problem solving.

Yanhui Guo et al [16], introduces a new clustering algorithm, neutrosophic c-means (NCM), for uncertain data clustering. That is inspired from fuzzy c-means and neutrosophic set framework. To derive such a structure, a novel suitable objective function is defined and minimized, and the clustering problem is formulated as a constrained minimization problem, whose solution depends on objective function. In the objective function, two new types of rejection have been introduced: ambiguity rejection which concerns the patterns lying near the cluster boundaries, and distance rejection dealing with patterns that are far away from all clusters. These measures are able to manage uncertainty due to imprecise and/or incomplete definition of the clusters. The results are encouraging and favorably with results from other methods as FCM, PCM and FPCM algorithms. Finally, the method was applied into image segmentation. The results show that proposed algorithm can be considered as a promising tool for data clustering and image processing.

IV. PROPOSED WORK

The proposed work is aimed to design a text clustering algorithm that accurately handles the text data. This section provides approach and methodology to understand about the proposed work.

A. System Overview

The social media is a largest source of data in these days. It supports various different kinds of data and file formats such as text, audio, video and images. The social media data can be used for various kinds of applications using data mining techniques such as terror prevention, natural disaster management, branding, advertising and others. Therefore text analysis using the data mining techniques is an essential task in social media. The text data supports categorization, clustering or classification algorithm frequently. In this context based on the application requirements the supervised or unsupervised algorithm can be employed. In this presented work text categorization application is aimed to demonstrate.

Basically the proposed work is motivated from a research article [1] where the unsupervised learning technique is used to analyze the hidden sentiments in the social media text. In this context the proposed work include the twitter social media dataset is considered for experimentation. The twitter dataset contains different attributes such as author, time stamp, post and others. Among them we need only the posted content for data analysis and recovery the user or authors sentiments. In this data set the available data contains a significant amount of impurity and noise. Therefore some additional efforts are required to make the data and utilizable with the proposed sentiment based text categorization.

In this context the preprocessing techniques are used to improve the data quality. The preprocessing techniques are used for removal of unwanted data and noise from the datasets such as tags, keywords and others. After preprocessing text features are computed. The traditional feature extraction techniques are not much effective for sentiment based text analysis therefore here NLP (natural language processing) based features are computed. The features of a dataset represent the entire data contents from which it derived. Finally, fuzzy c-means clustering (FCM) algorithm is taken for improvements. And a comparative study among the modified FCM and traditional FCM algorithm is conducted in this work. This section provides the overview of targeted work in next section the proposed system is demonstrated.

B. Proposed system

The proposed system for sentiment based text categorization is simulated in figure 4.1. The figure contains required components for performing the required task.

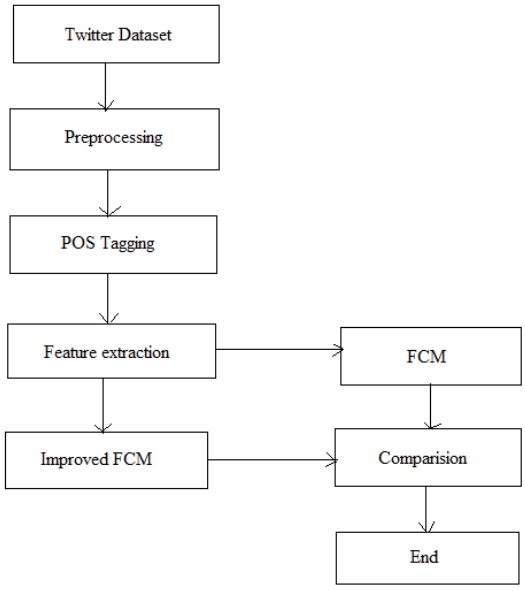

Figure 4.1 proposed text categorization system

The proposed data model is demonstrated in figure 4.1; this system accepts the input data as the social media text dataset. That is the initial input for the system. As we discussed before the given dataset contains noise and some unwanted data. Thus in order to clean and improve the data quality the preprocessing technique is employed. The preprocessing of the data includes

the three different aspects of data cleaning. That provides the three techniques of data cleaning first reducing the unwanted data attributes, reducing the stop words and finally removing special characters and tags.

After data preprocessing the feature selection process is involved for recovering small but significant contents. These contents or features are help to identify the emotions of the author or mood of author for sentiment based text categorization. Finally the extracted features are used for demonstrating the text categorization. Therefore first fuzzy c means clustering is implemented with system and then an improved version of FCM is applied which is implemented using the RBF (radial basis function). The RBF kernel implementation is used for nonlinear data classification. Thus it can be a promising approach. Therefore both the techniques are implemented and compared with the performance parameters.

This section provides the overview of the system and their functional aspects. To demonstrate the processes involved an algorithm steps are described in next section.

C. Proposed Algorithm

This section demonstrates how the proposed algorithm utilizes the improved FCM clustering algorithm for categorizing the text data. The table 4.1 contains the required steps:

Input: twitter dataset D, Number of clusters K Output: class label C
Process: 1. $D_n = ReadData(D)$ //where n is number of instances 2. $p_n = preProcessData(D_n)$ 3. $for(i = 1; i \leq n; i++)$ a. $POSTAG_i = NLP.ParseData(p_i)$ 4. $end\ for$ 5. $F_n = TransformData2D(POSTAG_n)$ 6. $[C, Data\ index] = IFCM.ClusterData(F_n, k)$ 7. Return C

Table 4.1 proposed text categorization algorithm

According to the given algorithm the social media text is accepted as input to the system which is denoted here as D, additionally the algorithm accept the number of clusters k. after the algorithm processes the algorithm for generating the class labels of input instances. In this context in first step the entire dataset is read using the system, and stored on a variable D_n where the n is number of instances to categorize, after that preprocessing technique is called to refine the data. The preprocessed data is stored in a variable p_n. Using the size n we initiate a loop which is used to tag the content of dataset into their part of speech information. The part of speech tagging is help to transform the data into a 2D vector. Finally improved FCM algorithm is applied to categorize the input dataset according to their sentiments.

V. CONCLUSION

Data mining is a young generation technology to understand the data and their applicability. Therefore the data is heart of entire data mining process and applications. Nature of the data mining algorithms can be supervised and unsupervised, the supervised learning are used as learn by example concept. Additionally unsupervised techniques are directly employed on data to differentiate data instance. Therefore it is beneficial for reducing learning time. The paper includes study of unsupervised learning. The aimed is to improve the performance of traditional text categorization approach. Thus an improved version of FCM algorithm is used for categorizing social media text.

In this context the twitter dataset is proposed for experimental usages. That contains the twits by different users. Initially the data is available in raw format thus data improvement required in this context a preprocessing technique is proposed for use. After that the NLP (natural language processing) based features are used for categorization. This paper provides the understanding of the proposed work for improving the text categorization approach for sentiment mining from social media text. In near future the proposed data model is implemented and their performance is reported.

REFERENCES

[1] Abdul Hannan, "Emotion Detection from Text", International Journal of Engineering Research and Development, Volume 11, Issue 07 (July 2015), PP.23-34

[2] Nikolaos Misirlis, Maro Vlachopoulou, "Social media metrics and analytics in marketing – S3M: A mapping literature review", International Journal of Information Management 38 (2018) 270–276

[3] TAO LI, CHUNQIU ZENG, YEXI JIANG, WUBAI ZHOU, and LIANG TANG, ZHENG LIU and YUE HUANG, "Data-Driven Techniques in Computing System Management", ACM Computing Surveys, Vol. 50, No. 3, Article 45, Publication date: July 2017.

[4] Thangaraj, M., & Sivakami, M. "Text classification techniques: A literature review", Interdisciplinary Journal of Information, Knowledge, and Management, 13, 117-135

[5] Janmenjoy Nayak, Bighnaraj Naik and H.S. Behera, "Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014",

[6] Abdulmohsen Algarni, "Data Mining in Education", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016

[7] Jiban K Pal, "Usefulness and applications of data mining in extracting information from different perspectives", Annals of Library and Information Studies Vol. 58, March 2011, pp. 7-16
[8] Sonia Xylina Mashal, Kavita Asnani, "Emotion Analysis of Social Media Data using Machine Learning Techniques", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, PP 17-20
[9] Ms.J.Sathya Priya, Ms.S.Priyadharshini, "Clustering Technique in Data Mining for Text Documents", International Journal of Computer Science and Information Technologies, Vol. 3 (1) , 2012, 2943-2947
[10] Anna L. Buczak, and Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, SECOND QUARTER 2016
[11] Dibya Jyoti Bora, Dr. Anil Kumar Gupta, "A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm", International Journal of Computer Trends and Technology (IJCTT) – volume 10 number 2 – Apr 2014
[12] Junyuan Xie, Ross Girshick, Ali Farhadi, "Unsupervised Deep Embedding for Clustering Analysis", Proceedings of the 33 rd International Conference on Machine Learning, New York, NY, USA, 2016 JMLR: W&CP volume 48. Copyright 2016 by the author(s)
[13] Celal Ozturk, Emrah Hancer, Dervis Karaboga, "Dynamic clustering with improved binary artificial bee colony algorithm", Applied Soft Computing 28 (2015) 69–80, © 2014 Elsevier B.V. All rights reserved.
[14] Nameirakpam Dhanachandra, Khumanthem Manglem and Yambem Jina Chanu, "Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm", Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Procedia Computer Science 54 (2015) 764 – 771
[15] Janmenjoy Nayak, Bighnaraj Naik and H.S. Behera, "Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014", © Springer India 2015 L.C. Jain et al. (eds.), Computational Intelligence in Data Mining - Volume 2, Smart Innovation, Systems and Technologies 32,
[16] Yanhui Guo, Abdulkadir Sengur, "NCM: Neutrosophic c-means clustering algorithm", Pattern Recognition 48 (2015) 2710–2724, & 2015 Elsevier Ltd. All rights reserved.