



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Multiview Point Based Similarity Measure for Text Classification and Clustering

R.Tamilselvi, R.Kavitha

M.Phil Research Scholar, Vivekanandha College For Women, Unjanai, Tiruchengode, Tamil Nadu, India.

Assistant Professor, Vivekanandha College For Women, Unjanai, Tiruchengode, Tamil Nadu, India.

ABSTRACT: Measuring the similarity between documents is an important operation in the text processing field. In this paper, a new similarity measure is proposed to text classification and clustering in sentiment analysis. The effectiveness of our measure is evaluated on several real-world data sets for text classification and clustering problems. The results show that the performance obtained by the proposed measure is better than that achieved by other measures. The current research is focusing on the area of text mining is Opinion Mining also called as sentiment analysis. One important problem in sentiment analysis of product reviews is to produce summary of opinions based on product features.

KEYWORDS: Classification, Clustering, K-Means, Support Vector Machine, Sentimental Analysis.

I. INTRODUCTION

Data mining is the analysis step of the "Knowledge Discovery in Databases" process, or KDD. An interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

DATA MINING CLASSIFICATION

Data Mining is a technique used in various domains to give meaning to the available data. Classification is a data mining technique used to predict group membership for data instances. Data mining algorithms can follow three different learning approaches: supervised, unsupervised, or semi-supervised. In supervised learning, the algorithm works with a set of examples whose labels are known. In unsupervised learning, in contrast, the labels of the examples in the dataset are unknown, and the algorithm typically aims at grouping examples according to the similarity of their attribute values, characterizing a clustering task. Finally, semi-supervised learning is usually used when a small subset of labelled examples is available, together with a large number of unlabeled examples.

DATA MINING CLUSTERING

Cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labeling of a large set of training tuples or patterns, which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity using clustering, and then assign labels to the relatively small number of groups. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

SENTIMENTAL ANALYSIS

Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. For example, businesses always want to find public or consumer opinions about their products and services. Potential customers also want to know the opinions of existing users before they use a service or purchase a product. With the explosive growth of social media on the Web, individuals and organizations are increasingly using public opinions in these media for their decision making. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level-whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy". To determine the sentiment in a text rather than the overall polarity and strength of the text.

II. RELATED WORK

To compute the similarity between two documents with respect to a feature, the proposed measure takes the following three cases into account: a) The feature appears in both documents, b) the feature appears in only one document, and c) the feature appears in none of the documents. For the first case, the similarity increases as the difference between the two involved feature values decreases. [3] Furthermore, the contribution of the difference is normally scaled. For the second case, a fixed value is contributed to the similarity. For the last case, the feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents. [3] Two documents are least similar to each other if none of the features have non-zero values in both documents. Besides, it is desirable to consider the value distribution of a feature for its contribution to the similarity between two documents. [3] The proposed scheme has also been extended to measure the similarity between two sets of documents. In this work, we are focusing on the performance resulted from the application of different similarity measures in different classification/clustering algorithms.

Sentiment analysis or opinion mining consist of many different fields like natural language processing, text mining, decision making and linguistics [7]. It is a type of text analysis that classifies the text and makes decision by extracting and analysing the text. Opinions can be categorized as positive and negative and measures the degree of positive or negative associated with that event such as people, organization and social issues[5]. So, it's basically people's opinion study, study of emotions and appraisals in the direction of any social issue, people or entity. Recently most of the researches have been done on the sentiment analysis of products and services. The analysis of events and issues, data is retrieved from social media like twitter etc[7].

The effect of using unlabelled data in conjunction with a small portion of labelled data on the accuracy of a centroid-based classifier used to perform single label text categorization. To use centroid-based methods because they are very fast when compared with other classification methods, but still present accuracy close to that of the state-of-the-art methods[1]. Efficiency is particularly important for very large domains, like regular news feeds, or the web. To propose the combination of Expectation-Maximization with a centroid-based method to incorporate information about the unlabelled data during the training phase[9]. To propose an alternative to EM based on the incremental update of a centroid-based method with the unlabelled documents during the training phase. And also showed how a centroid-based method can be used to incrementally update the model of the data, based on new evidence from the unlabelled data[9].

III. SCOPE OF RESEARCH

The goal is to develop a similarity measure for text classifier that performs sentiment analysis, by labeling the users comment to positive or negative. After complete the similarity measure for text classification the Clustering algorithms are used to measure the similarity or closeness between the text objects. To find out the similarity, cluster takes a collection of text which is "similar" in to a particular group/cluster and the "dissimilar" objects belonging to other clusters. The most well-known similarity function which is used commonly in the text domain is the cosine similarity function. Clustering is a useful technique that organizes a large quantity of unordered text documents into a



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

small number of meaningful and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms.

IV.METHODOLOGY

SUPPORT VECTOR MACHINE

Support Vector Machines construct a decision surface in the feature space that bisects the two categories and maximizes the margin of separation between two classes of points. This decision surface can then be used as a basis for classifying points of unknown class. That text data is ideally suited for SVM classification because of the sparse high-dimensional nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories. Define the hyperplane H such that:

$$x_i \cdot w + b \geq +1 \text{ when } y_i = +1 [\text{Positive Class}]$$

$$x_i \cdot w + b \leq -1 \text{ when } y_i = -1 [\text{Negative Class}]$$

SVM ALGORITHM

INPUT: Training Data, Testing Data

OUTPUT: Decision Value

METHOD

Step 1: Load Dataset

Step 2: Classify Features (Attributes) based on class labels

Step 3: Estimate Candidate Support Value

Candidate support value= {closest pair from opposite classes}

While (instances! =null)

Do

Step 4: Find a support vector

Support Value=Similarity between each instance in the attribute

Find Total Error Value

Step 5: If any instance < 0

Estimate

Decision value = Support Value\Total Error

Repeat for all points until it will empty

End If

Initialization step of the SVM algorithm is to load the data set. The goal is to produce a model which predicts target value of data instances in the testing set which are given only the attributes. After Classify the attributes based on the class labels. SVM chooses a maximum-margin hyper plane that lies in a transformed input space and splits based on the similarity. Then Estimate candidate support value. Support Vector Machine (SVM) can find an optimal solution by maximizing the distance between the hyper plane and the points close to decision boundary.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

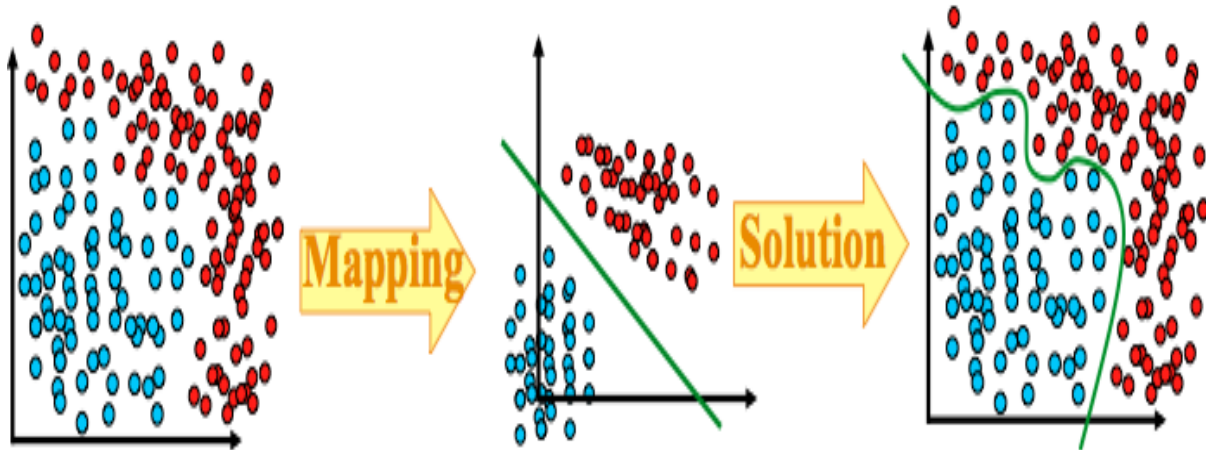


Figure: Classify the Attributes in Linear and Non-linear

K-MEANS CLUSTERING

The data in a cluster will have similar features or characteristic which will be dissimilar from the data in other clusters. This is how k-means algorithm partitions a dataset into clusters: it accepts the number of clusters to group data into and the dataset to cluster as input values. The k-means algorithm calculates the arithmetic mean of each cluster formed in the dataset. The arithmetic mean of a cluster is the mean of all individual records in the cluster. In each of the first K initial clusters, there is only one record. The arithmetic mean of a cluster with one recorded is the set of values that make the record.

K-MEANS CLUSTERING ALGORITHM

The k-means algorithm is used for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

INPUT

K: the number of clusters,

D: a data set containing n objects.

OUTPUT: A set of k clusters.

METHOD

Step 1: Choose k numbers of clusters to be determined.

Step 2: Choose C_k centroids randomly as the initial centers of the clusters.

Step 3: Repeat

3.1: Assign each object to their closest cluster center using Euclidean distance.

3.2: Compute new cluster center by calculating mean points.

Step 4: Until

4.1: No change in cluster center OR

4.2: No object changes its clusters.

Initially choose k number of clusters as the input parameter. Then partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Determine the centroids randomly. Let assume $k=3$ that means the objects to be partitioned into three clusters. then compute the distance from the centroids and assign elements to the cluster of the nearest centroids based on Euclidean distance. Then the cluster centers are updated. By the mean value of each cluster is recalculated based on the current objects in the cluster. Using the new

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

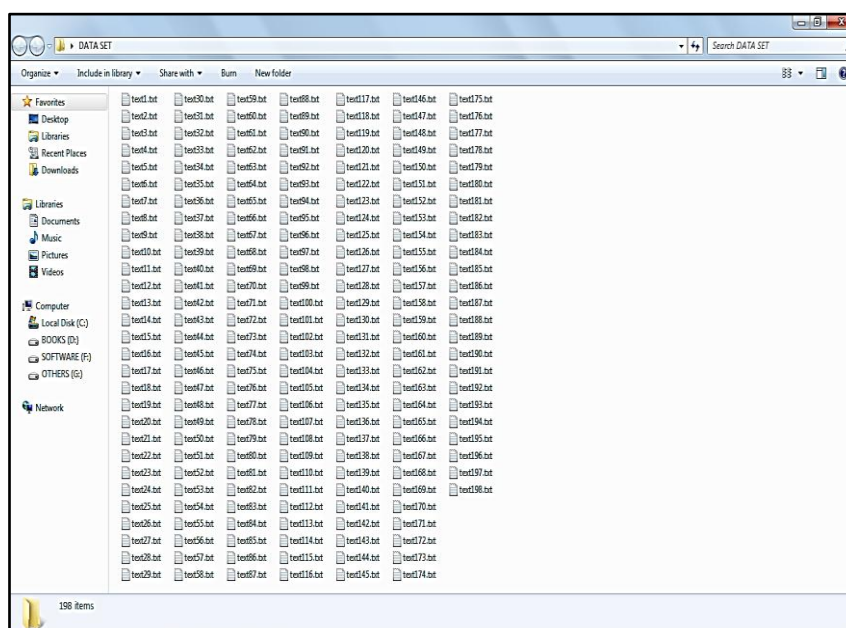
cluster centers, the objects are redistributed to the clusters based on which cluster center is the nearest. Repeat step 2 and 3 until all clusters converge as well as no changes in centroids.

SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is an application of Text Analytics to identify and extract subjective information in source materials. A basic task in sentiment analysis is classifying an expressed opinion in a document, a sentence or an entity feature as positive or negative. This program implements Precision and Recall method. Precision is the probability that a retrieved document is relevant. Recall is the probability that a relevant document is retrieved in a search. Or high recall means that an algorithm returned most of the relevant results. High precision means that an algorithm returned more relevant results than irrelevant. At first, both positive and negative reviews of a certain movie are taken. All of the words are stemmed into root words. Then the words are stored in different polarity such as positive and negative.

V.EXPERIMENTAL RESULTS

The Review of Individual Movie watchers are collected under a folder for classification.

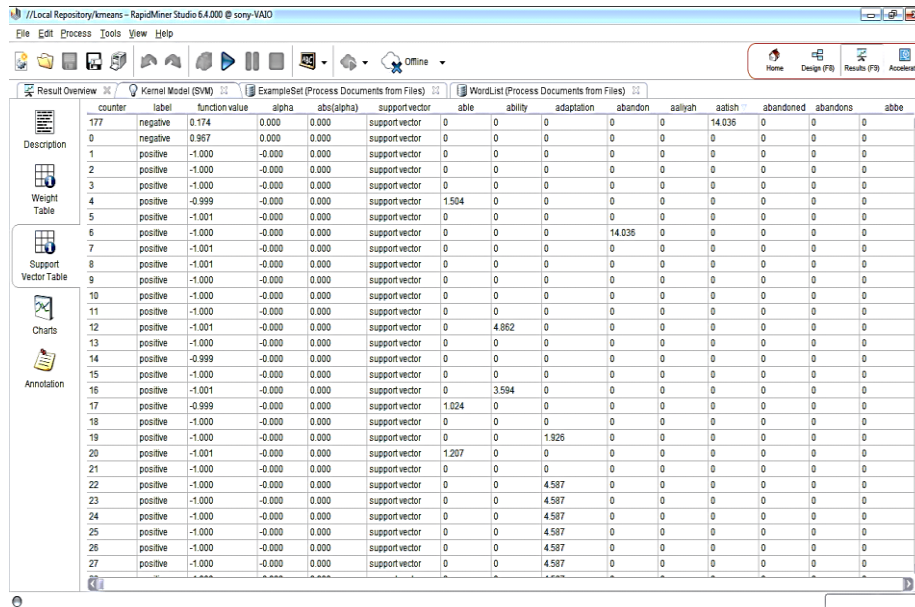


Support Vector Table is classified based on the concept of decision planes that define decision boundaries. It performs classification tasks by constructing hyper planes in a multidimensional space that separates the different class labels.

International Journal of Innovative Research in Computer and Communication Engineering

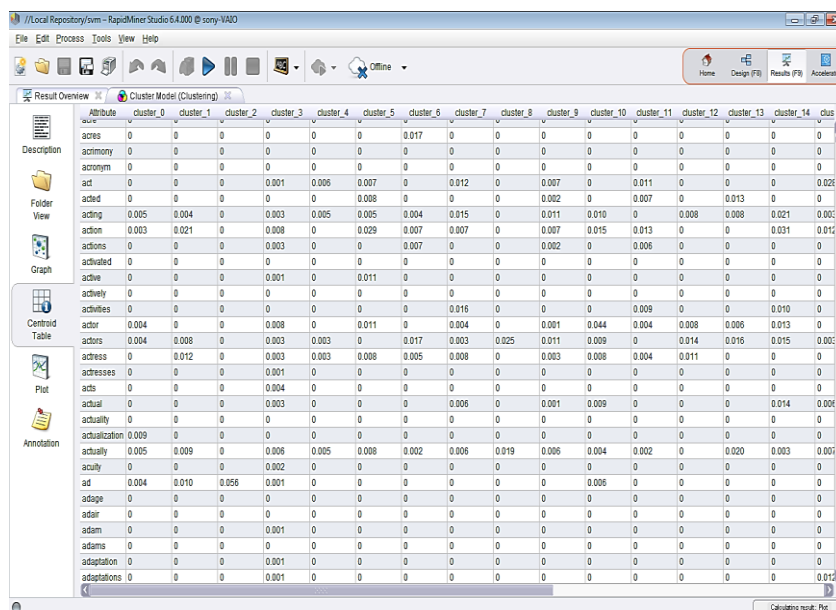
(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015



countnr	label	function value	alpha	abs(alpha)	support vector	able	ability	adapstion	abandon	aslysh	astsh	abandoned	abandons	abbe
177	negative	0.174	0.000	0.000	support vector	0	0	0	0	0	14.036	0	0	0
0	negative	0.987	0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
1	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
2	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
3	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
4	positive	-0.999	-0.000	0.000	support vector	1.584	0	0	0	0	0	0	0	0
5	positive	-1.001	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
6	positive	-1.000	-0.000	0.000	support vector	0	0	0	14.036	0	0	0	0	0
7	positive	-1.001	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
8	positive	-1.001	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
9	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
10	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
11	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
12	positive	-1.001	-0.000	0.000	support vector	0	4.862	0	0	0	0	0	0	0
13	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
14	positive	-0.999	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
15	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
16	positive	-1.001	-0.000	0.000	support vector	0	3.594	0	0	0	0	0	0	0
17	positive	-0.999	-0.000	0.000	support vector	1.924	0	0	0	0	0	0	0	0
18	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
19	positive	-1.000	-0.000	0.000	support vector	0	0	1.926	0	0	0	0	0	0
20	positive	-1.001	-0.000	0.000	support vector	1.297	0	0	0	0	0	0	0	0
21	positive	-1.000	-0.000	0.000	support vector	0	0	0	0	0	0	0	0	0
22	positive	-1.000	-0.000	0.000	support vector	0	0	4.587	0	0	0	0	0	0
23	positive	-1.000	-0.000	0.000	support vector	0	0	4.587	0	0	0	0	0	0
24	positive	-1.000	-0.000	0.000	support vector	0	0	4.587	0	0	0	0	0	0
25	positive	-1.000	-0.000	0.000	support vector	0	0	4.587	0	0	0	0	0	0
26	positive	-1.000	-0.000	0.000	support vector	0	0	4.587	0	0	0	0	0	0
27	positive	-1.000	-0.000	0.000	support vector	0	0	4.587	0	0	0	0	0	0

After completing the classification tasks then to compute the distance from the centroids and assign elements to the cluster of the nearest centroids based on Euclidean distance using K-Means Algorithm. Then the cluster centers are updated by the mean value of each cluster is recalculated based on the current objects in the cluster. Using the new cluster centers, the objects are redistributed to the clusters based the nearest cluster center. Then to group the similar word based on cosine similarity.



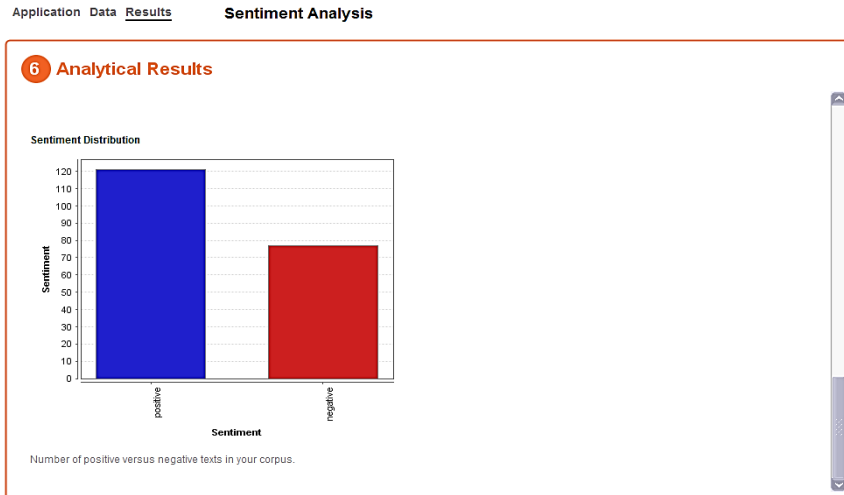
Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7	cluster_8	cluster_9	cluster_10	cluster_11	cluster_12	cluster_13	cluster_14
acres	0	0	0	0	0	0	0.017	0	0	0	0	0	0	0	0
acrimony	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
acronym	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act	0	0	0	0.001	0.006	0.007	0	0.012	0	0.007	0	0.011	0	0	0.026
acted	0	0	0	0	0	0.008	0	0	0	0.002	0	0.007	0	0.013	0
acting	0.005	0.004	0	0.003	0.005	0.005	0.004	0.015	0	0.011	0.010	0	0.008	0.008	0.021
action	0.003	0.021	0	0.008	0	0.029	0.007	0.007	0	0.007	0.015	0.013	0	0	0.031
actions	0	0	0	0.003	0	0	0.007	0	0	0.002	0	0.006	0	0	0
activated	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
active	0	0	0	0.001	0	0.011	0	0	0	0	0	0	0	0	0
actively	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
activities	0	0	0	0	0	0	0	0.016	0	0	0	0	0.009	0	0.010
actor	0.004	0	0	0.008	0	0.011	0	0.004	0	0.001	0.044	0.004	0.008	0.006	0.013
actors	0.004	0.008	0	0.003	0.003	0	0.017	0.003	0.025	0.011	0.009	0	0.014	0.016	0.015
address	0	0.012	0	0.003	0.003	0.008	0.005	0.008	0	0.003	0.008	0.004	0.011	0	0
addresses	0	0	0	0.001	0	0	0	0	0	0	0	0	0	0	0
acts	0	0	0	0.004	0	0	0	0	0	0	0	0	0	0	0
actual	0	0	0	0.003	0	0	0	0.006	0	0.001	0.009	0	0	0	0.014
actualization	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
actualization	0.009	0	0	0	0	0	0	0	0	0	0	0	0	0	0
actuality	0.005	0.009	0	0.006	0.005	0.008	0.002	0.006	0.019	0.006	0.004	0.002	0	0.020	0.003
actuality	0	0	0	0.002	0	0	0	0	0	0	0	0	0	0	0
act	0.004	0.010	0.056	0.001	0	0	0	0	0	0	0.006	0	0	0	0
adage	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
adair	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
adam	0	0	0	0.001	0	0	0	0	0	0	0	0	0	0	0
adams	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
adaptation	0	0	0	0.001	0	0	0	0	0	0	0	0	0	0	0
adaptations	0	0	0	0.001	0	0	0	0	0	0	0	0	0	0	0.012

It shows the Final Output is positive/negative based on sentimental Analysis.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015



VI. CONCLUSION AND FUTURE WORK

In this paper clustering and classification is used to identify the customers as well as user opinion in sentimental analysis. To improve the efficiency, it is proposed to measure the similarity between multiple documents. This technique gives more accurate results while compared to similarity measure for two sets of text documents. Using SVM algorithm is used to deal with very large datasets in linear and nonlinear classification. The support vector machine has been defined as input and output format based on hyper plane. Input is a vector space and output is 0 or 1 that means positive or negative. The support vector machine in text classification helps to find the positive class or negative from a particular text document. After the text classification, the k-means algorithm is used to group the text based on cluster centroids. The clustering process is divided into similar and dissimilar clusters. Similar or similar text gathered or occurring closely together are called similar clusters or positive clusters. The particular text not belonging to similar class is called dissimilar or negative cluster. In future work, other classification and clustering algorithms, such as naive Bayes tree, genetic algorithms, rough set approaches, fuzzy logic and hierarchical clustering will be used to deal with real-time multi-class classification tasks under dynamic feature sets for opinion mining.

REFERENCES

- [1] Ana Cardoso, Cachopo, Arlindo, L. Oliveira, "Semi supervised Single label Text Categorization using Centroid based Classifiers", SAC'07 March 2007, Seoul, Korea
- [2] Kamal nigam, andrewkachitesmccallum, sebastianthrun, "Text Classification from Labeled and Unlabeled Documents using EM".
- [3] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 7, July 2014.
- [4] S. Neelamegam, Dr. E. Ramaraj, "Classification algorithm in Data mining: An Overview", International Journal of P2P Network Trends and Technology (IJPTT) – Volume 4 Issue 8- Sep 2013.
- [5] Haseena Rahmath P, "Opinion Mining and Sentiment Analysis -Challenges and Applications", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 5, May 2014.
- [6] P. H. Govardhan, Prof. K. P. Wagh, Dr. P. N. Chatur, "Survey on Similarity Measure for Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013.
- [7] Ms. Gaurangi Patil, Ms. Varsha Galande, Ms. Kalpana Dange, "Sentiment Analysis Using Support Vector Machine", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014.
- [8] Shady Shehata, Fakhri Karray, Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, October 2010.
- [9] D. Renukadevi, S. Sumathi, "Term Based Similarity Measure for Text Classification and Clustering using Fuzzy c-means algorithm", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 4, April 2014.