Analyzing Twitter Data Using Unsupervised Learning Techniques

N. Lakshmi Devi
Asst.Professor, Dept of CSE, GMRIT, Rajam-532127, India.

K.S rividya
Asst.Professor, Dept of CSE, GMRIT, Rajam-532127, India.

Abstract – Social media has become important for social networking and content sharing. Twitter, an online social network, allows users to upload short text messages, also known as tweets, with up to 140 characters. A lot of people use sentiment analysis on Twitter to do opinion mining. The objective of Sentiment Analysis is to identify any clue of positive or negative emotions in a piece of text reflective of the authors' opinions on a subject. The main objective of opinion mining is to cluster the tweets into positive, neutral and negative clusters. An earlier work is based on supervised machine learning (Naïve bays, maximum entropy classification and support vector machines). The proposed work is able to collect information from social networking sites like Twitter and the same is used for sentiment analysis. The processed meaningful tweets are cluster into three different clusters positive, neutral and negative using unsupervised machine learning technique such as K-Means clustering and DBSCAN clustering. Manual analysis of such large number of tweets is impossible. So the automated approach of unsupervised learning used. This project deals with the issues of clustering algorithm. The results are discussed on datasets. Using JAVA we have calculated the percentage of correctness of K-Means and DBSCAN clustering algorithms.

Index Terms Sentiment Analysis, K-Means, DBSCAN, NLP

1. INTRODUCTION

Opinion Mining is a field of Web Content Mining that aims to find valuable information out of users opinions. Mining opinions on the web is a fairly new subject, and its importance has grown significantly mainly due to the fast growth of e-commerce, blogs and forums. With the high profits of e-commerce increasing year after year many people had changed the habit of going to a shop for the comfortable virtual shopping. The emergence of social media has given web users a venue for expressing and sharing their thoughts and opinion on different topics and events. Twitter, with nearly 600 million users and over 250 million messages per day, has quickly become a gold mine for organizations to monitor their reputation and brand by extracting and analyzing the sentiments of the tweets posted by public about them, their markets, and competitors.

Opinion mining or sentiment analysis has recently received a lot of attention in the natural language processing (NLP) community. Opinion mining whose goal is to determine whether the opinion expressed in a twitter is "thumbs up" or "thumbs down" is arguably one of the most popular tasks in document level sentiment analysis. Opinion mining uses some algorithm techniques to cluster the user opinions into positive and negative clusters. Earlier work is based on supervised learning such as. Supervised learning have been popularly used and proven its effectiveness in sentiment classification. It is highly depend on large amount of labeled data which results in time consuming and also expensive one. Based on the previous work unsupervised learning method are proposed to overcome the problem of supervised learning method which require large amount of unlabeled data. Unsupervised learning is the machine learning task of inferring a function to describe the hidden structure from unlabeled data. Approaches to unsupervised learning are clustering (e.g.,k-means clustering and hierarchical clustering).

2. RELATED WORK

2.1. Problem Statement

Simple clustering algorithms such as K-means and hierarchical are widely used for analyzing large data sets; but they are unable to deal with noise and high dimensional data.

Incorporating prior knowledge in clustering (semi-supervised) has been shown to improve the consistency between the data partitioning and domain knowledge. It improves the robustness and quality of clustering. It eliminates the noise data before partitioning and thus helps us in analyzing the large data sets easily.

2.2. Existing problem

In opinion mining task documents and example are represented by thousands of tokens, which make the clustering problem very hard for many clustering system. In feature extraction, the original features converted to more compact new space. All the original features are transformed into new reduced space without deleting them but replacing the original features through a smaller representative set. Feature selection is a process of removing the irrelevant and redundant features from a dataset in order to improve the performance of

unsupervised learning algorithm in terms of accuracy and time to build the model. Y.Mejova et al in his research work proposed that we can use presence of each character, frequency of occurrence of each character, word which is considered as negation etc. as feature for creating feature vector.

Drawbacks

Tree pruning has been shown to increase the prediction accuracy of a decision tree on one hand and reduce the complexity of the tree on the other.

The Pittsburgh approach was originally designed for single-class learning problems and hence only the antecedent of a rule was encoded into an allele of a chromosome.

The GAs are run once for each class. Specifically, they would search rules predicting the first class in the first run; they would search rules predicting the second class in the second run and so on. This problem has not been addressed by these algorithms.

3. PORPOSED SYSTEM

Various techniques have been used to do sentiment analysis or opinion mining of tweets. The proposed system contains various phase of development. A dataset is created using twitter. As we know that tweets contains slang words and misspelling. So we perform a sentiment level sentiment analysis on tweets. This is done in three phases. In the first phase pre-processing is done. Then feature vector is created using relevant features. Finally using different unsupervised learning techniques, tweets are cluster into positive and negative classes. The supervised approach can be categorized as corpus-based methods as it uses labeled data to train sentiment classifiers. Given the difficulties of supervised sentiment analysis, it is conceivable that unsupervised approach to sentiment classification is even more challenging. Unsupervised learning overcomes these difficulties. It divides the document into sentences and categorizes each sentence using word list of each category. Unsupervised learning does not require training set and test set.

The entire proposed modeling and architecture of the current research paper should be presented in this section. This section gives the original contribution of the authors. This section should be written in Times New Roman font with size 10. Accepted manuscripts should be written by following this template. Once the manuscript is accepted authors should transfer the copyright form to the journal editorial office. Authors should write their manuscripts without any mistakes especially spelling and grammar.

3.1 Cluster analysis

The size of data is increasing vigorously every day. The clustering methods are used to quickly retrieve the required information from the large data repositories. Clustering is the process of grouping the data objects into clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members or particular statistical distributions. Clustering can be formulated as a multi-objective optimization problem.

3.1.1 Main Elements of Cluster Analysis

1) Representation of data
2) Choice of objects
3) Variables choosing
4) What to cluster
5) Normalization of variables
6) Choosing similarity or dissimilarity measures
7) Criteria for clustering
8) Choice of missing data strategy
9) Algorithms and implementation
10) Number of clusters
11) Interpretation of results.

3.2 Clustering techniques

3.2.1 K-Means clustering

Given a data set, a desired number of clusters, k, and a set of k initial starting points, the k-means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is defined as the point whose coordinates are obtained by computing the average of each of the coordinates of the points of the samples assigned to the cluster. Formally, the k-means clustering algorithm follows the following steps. 1. Set k: Choose a number of desired clusters, k. 2. Initialization: Choose k starting points to be used as initial estimates of the cluster centroids. These are the initial starting values. 3. Classification: Examine each point in the data set and assign it to the cluster whose centroid is nearest to it. 4. Centroid Calculation: When each point is assigned to a cluster, recalculate the new k centroids. 5. Convergence condition: Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

Before the clustering algorithm can be applied, actual data samples are collected. The features that describe each data sample in the database are required a priori. The values of these features make up a feature vector (Fi1, Fi2, … , Fim), where Fim is the value of the mth feature of the ith job. The feature vector can be thought of as a point in M-dimensional space. Like other clustering algorithms, k-means requires that a distance metric between points be defined. This distance metric is used in step 3 of the algorithm given above. A common distance metric is the Euclidean distance. Given two sample points, pi and pj, each described by their feature vectors, pi = (Fi1, Fi2, … , FiM) and pj = (Fj1, Fj2, … , FjM), the distance, dij, between pi and pj is given by:

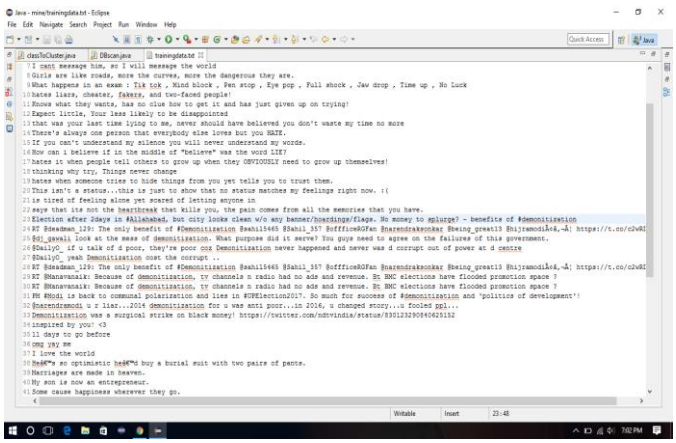
$$dy = \text{sqrt}(\text{sum}(\text{Fim} - \text{Fjm})2)$$

If the different features being used in the feature vector have different relative values and ranges, the distance computation may be distorted and hence can be scaled. The number of clusters to be found, along with the initial starting point values are specified as input parameters to the clustering algorithm. Given the initial starting values, the distance from each sample data point to each initial starting value is found using equation. Each data point is then placed in the cluster associated with the nearest starting point. New cluster centroids are calculated after all data points have been assigned to a cluster. Suppose that C_{im} represents the centroid of the m^{th} feature of the ith cluster. Then,

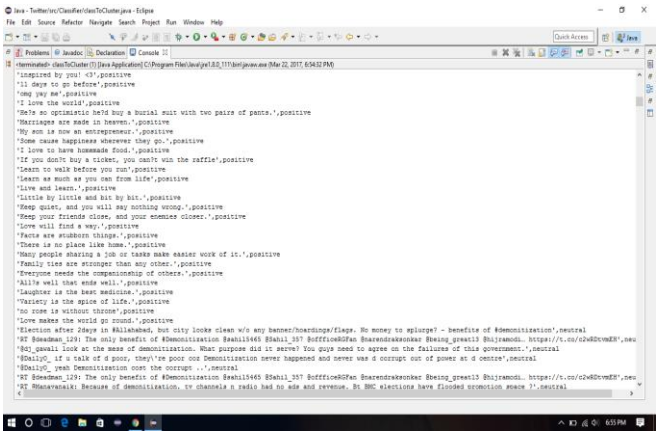
$$\text{Cim} = (\text{sum}(\text{Fi,jm})/\text{ni})$$

where F*i,jm is the m^{th} feature value of the j^{th} job assigned to the i^{th} cluster and where ni is the number of data points in cluster i. The new centroid value is calculated for each feature in each cluster. These new cluster centroids are then treated as the new initial starting values and steps 3-4 of the algorithm are repeated. This continues until no data point changes clusters or until a maximum number of passes through the data set is performed.

Input:



Output:



3.2.2 DBSCAN clustering

Density-based Spatial Clustering of Applications with Noise. DBSCAN is a density based clustering algorithm, where the number of clusters are decided depending on the data provided. Density based clustering algorithm has played a vital role in finding nonlinear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reach ability and density connectivity. Density Reachability - A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbor which are within distance ε. Density Connectivity - A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ε distance.

This is chaining process. So, if "q" is neighbour of "r", "r" is neighbor of "s", "s" is neighbour of "t" which in turn is neighbor of "p" implies that "q" is neighbo9r of "p".

This is unlike K – Means Clustering, a method for clustering with predefined ‘K’, the number of clusters. Since it is a density based clustering algorithm, some points in the data may not belong to any cluster. Again, this is unlike K – Means Clustering where all the points are assumed to be belonging to some cluster. Density = number of points within a specified radius r (Eps)

- Core point: A point is a core point if it has more than a specified number of points (MinPts) within Eps .At the interior of a cluster.
- Border point: A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point .At the outer surface of a cluster.

- Noise point: A noise point is any point that is not a core point or a border point

  Not part of any cluster.

Algorithm:

The process of the DBSCAN algorithm is described as follows:

Input: Let X = {x1, x2, x3, ..., xn} be the set of data points. DBSCAN requires two parameters: ε (Eps) and the minimum number of points required to form a cluster (minPts).

Output: A set of clusters

Steps:

1) Start with an arbitrary starting point that has not been visited.

2) Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).

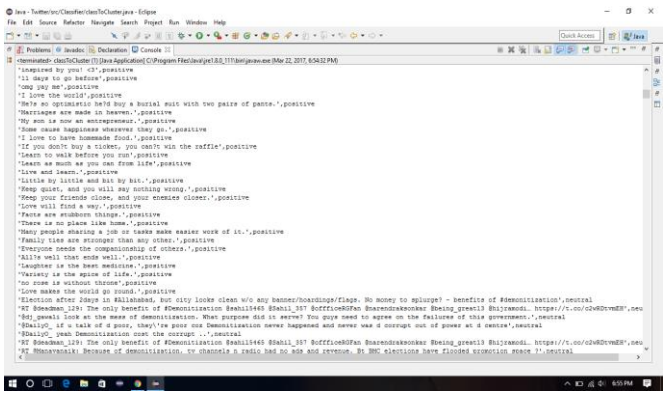
3) If there is sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).

4) If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.

5) A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

6) This process continues until all points are marked as visited.

Output:



4 RESULTS AND DISCUSSIONS

Clustering is one of the most essential steps in data mining. It is the process of grouping data items based on similarity between elements. K-means is a numerical, unsupervised, iterative method. It is simple and very fast, when the data set is small, it is proved to be a very effective way that can produce good clustering results. It cannot produce effective results when data set is large. DBSCAN algorithm gives good results in small as well as large data sets. DBSCAN substantially outperformed standard k-means in terms of percentage of correctness.

Accuracy:

Accuracy, or correctness of the cluster is defined as the ratio of the number of correctly clustered tweets to the total number of tweets calculated using the below equation(1) or equation(2).

Accuracy=(TP+TN)/(TP+TN+FP+FN) ……(1)

$$\text{Accuracy \%} = \frac{\text{Correctly Clusterd Tweets}}{\text{Total Number of Tweets}} * 100 \ldots. \ (2)$$

Number of Records	Percentage of Correctness K-MEANS	Percentage of Correctness DBSCAN
10	100	100
20	98.23	99
50	96.586	98.72
86	95.5	96.5
300	86	95

Table1: Comparison Results of K-means and DBSCAN

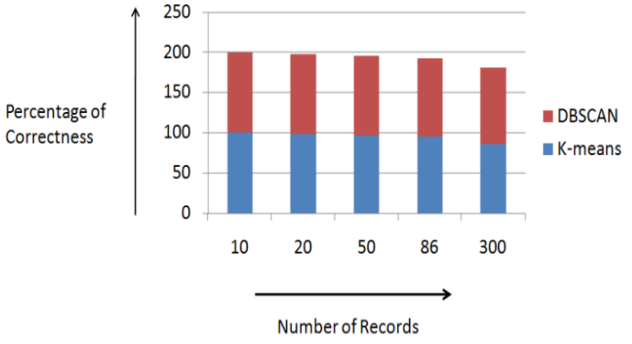


Figure: Graph to conclude Best Algorithm

After human inspection, the data set has a total of 86 tweets. Once the clustering is processed, 3 clusters are taken. Cluster 1 has 31 tweets, cluster 2 has 35 tweets, and cluster 3 has 20 tweets. The clusters are added into the sentiment analysis system in order to compute the score. Table 1 shows the result of this computation. A few people are asked to manually judge if this content is positive or negative. After that,

classifier evaluation metrics and confusion matrix are used to check the score from this project and the judgment from the people who review the content. Table1shows the evaluation report of evolution. True positives (TP) means human's check and system output are both positive. True negative (FP) means human's check and system output are both negative. TP and FP mean the system output has correct determine. False negative (FN) means human's check is positive, but system output is negative. False positive (FP) means human's check is negative, but system output is positive. FN and FP means the system output has wrong determine. ~FN and ~FP means the tweets are not about positive and negative.

5 CONCLUSION

Twitter base social network provides the great platform in measuring the public opinion with the reasonable accuracy with machine learning algorithms for sentiment analysis. In this, a new opinion mining of twitter data using unsupervised learning technique is proposed that can solve the problem of domain dependency and reduce the need of annotated training data. Unsupervised machine learning techniques have shown better performance than supervised learning. The main goal is to overcome the problem of clustering multiple files with unlabeled data and perform sentiment classification.

In this we used K-means is a typical clustering algorithm and it is widely used for clustering large sets of data. This project elaborates k-means algorithm and analyses the shortcomings of the standard k-means clustering algorithm. Because the computational complexity of the standard k-means algorithm is objectionably high owing to the need to reassign the data points a number of times during every iteration, which makes the efficiency of standard k-means clustering is not high. This project presents a simple and efficient way for assigning data points to clusters using DBSCAN algorithm

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] International Journal of Computer Applications (0975 – 8887) Volume 148 – No.12, August 12th, 2016 "Opinion Mining on Twitter Data using Unsupervised Learning Technique" TS Syed Raziuddin, PhD Professor & HOD of CSE Dept Deccan College of Engineering and Technology Darussalam Hyderabad TS.

[3] Emotion Analysis of Twitter using Opinion Mining by Akshi Kumar, Prakhar, Dogra, and Vikrant Dabas Dept. of Computer Engineering, Delhi Technological University New Delhi, India,2015.

[4] "Using Twitter for tapping public minds, predict trends and generate value", based on Fifth International Conference on Advanced Computing & Communication Technologies(2015) by Sanchita Kadambari, Kalpana Jaswar, Praveen Kumar.

[5] "Semantic Properties of Customer Sentiment in Tweets", 2014 based on 28th International Conference on Advanced Information Networking and Applications Workshops.