A Survey of Text Mining Classification And Clustering Techniques

Sweety Bakyarani. E[1], Dr.Srimathi.H [2]
[1, 2] Dept of Computer Applications
[1, 2] SRM Institute of Science and Technology, Tamil Nadu, India.

Abstract- *Test Analytics, Text Data Mining, or Text Mining, are terms that are used to describe the process of extracting meaningful, quality information from text sources. With the advent of blogging and micro blogging sites, the amount of text available in the World Wide Web has grown exponentially over the past decade. And unlike other datasets, textual data is highly unstructured, making it difficult to analyze. Text mining usually involves first structuring the input text, deriving meaningful patterns from within the structured data, and finally evaluating and interpreting the output data.Effective and efficient algorithms for doing this is the need of the hour. The goal of this paper is to survey the algorithms involved in text classification and clustering available and are popularly used.*

Keywords- Classification, Clustering, K-Nearest Neighbor, Naive Bayes, Text Mining.

I. INTRODUCTION

Text Mining (TM), has garnered huge amount of attention in the past few years mainly because of the humongous growth of text data online. The main source of this text data being social networking sites, blogging, micro blogging sites, and news outlets to name a few. Text analytics focuses on extracting key pieces of information from this highly unstructured text. Text mining tries to find the “who,” “where,” “when”, “what” or the “buzz” of the conversation, “how” people are feeling and “why” the conversation is happening. It finds huge applications in business sector where it’s all about, what the customer wants and what the customer feels about a product. Computers are very good in understanding and processing structured data, but its significantly harder for them to understand unstructured data. Also understating natural language and inferring the context at which certain words are used is very complex. Hence, we need algorithms that can understand and process this huge volume of data that is available.

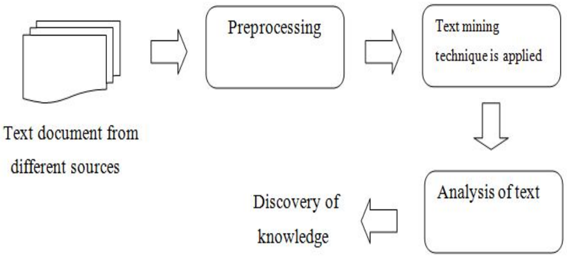
Text document from different sources
Preprocessing
Text mining technique is applied
Analysis of text
Discovery of knowledge

Figure1: Steps Involved in Text Mining

Text Mining can be grouped into seven areas they are:

1. **Information retrieval (IR):**It is nothing but the task of retrieving of data from unstructured sources. The primary focus is on data retrieval and not on analyzing it
2. **Document clustering:** Grouping and categorizing terms, snippets, paragraphs, or documents, using data mining clustering methods.
3. **Document classification:** Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods.
4. **Web mining:** Mining the Internet, specifically sites like Twitter, Facebook and other social networking sites to gather information. It makes use of the Interconnecting property of the World Wide Web.
5. **Information extraction (IE):** It is the process of identifying and extracting of relevant facts and relationships from unstructured text. It is usually the starting point of text mining.
6. **Natural language processing (NLP):**It is a field of study that encompasses AI and Linguistics, the main aim of NLP is to facilitate computers to understand natural language. Many Text Mining algorithms extensively use NLP.
7. **Concept extraction:**It is nothing but grouping or classifying text based on a specific concept.

II. CLASSIFICATION ALGORITHMS

The goal of text classification is to assign predefined classes to text documents. The problem of classification can be defined as follows. We have a training data set D = {d_1,d_2, .

. ., d_n } of documents, such that each document di is labeled with a label ℓ_i from the set L = {ℓ_1, ℓ_2, . . . , ℓ_k }. The task is to find a classifier c where c: D −→ L c(d) = ℓ, a classifier that can assign the correct class label to new document d. d is called as the test instance. The classification is called hard, if a label is explicitly assigned to the test instance. A classification is called soft, if a probability value is assigned to the labels of the test instance. There are also other types of classifications that allow multiple labels can be assigned to the test set. Many classification algorithms are implemented in tools like WEKA, Mallet, BOW etc.

2.1. Naïve Bayes Classifier

Classifies that use probability are very popular and they also perform well. It is a probabilistic approach that makes assumptions about how words in documents are generated. They propose a probabilistic model based on these assumptions. A set of training data is used to estimate the parameters of the model. Bayes rule is used to classify new examples and to select the class that is most likely to have generated the example. The Naive Bayes classifier is the simplest and the most widely used classifier. It models the distribution of documents in each class using a probabilistic model and assumes that the distributions of different terms are independent from each other. The Naïve Bayes classifier performs well on real world applications.

There are two main models commonly used for naive Bayes classifications. Both models goal is to find the posterior probability of a class, based on the distribution of the words in the document. One model takes into consideration the frequency of the words and the other does not.

(1) Multi-variate Bernoulli Model: In this model the document is represented as a vector of binary features that indicated the presence or absence of the words in the document. Thus, the frequency of words is ignored.
(2) Multinomial Model: The frequencies of words in a document are got by representing the document as a bag of words.

For a small vocabulary size, the Bernoulli model may perform better than Multinomial model. But Multinomial Model always performs better on large vocabulary size. For optimal vocabulary size, the Multinomial Model performs better than Bernoulli.

2.2. Nearest Neighbor Classifier

Nearest neighbor classifier is a proximity-based classifier. It uses distance-based measures to perform the classification. The main idea behind this classifier is that documents which belong to the same class are more “similar” or close to each other. The text document is classified, based on the class labels of similar documents in the training set. The approach is called k-nearest neighbor classification, if we consider the k-nearest neighbor in the training dataset. The most common class from these k neighbors is considered as the class label.

2.3. Decision Tree classifiers

Decision tree is a hierarchical tree of training instances. In this classifier, the condition on the attribute value is used to divide the data. Decision tree recursively partitions the training data set into smaller divisions at each node or branch, based on a set of tests called as‘d’. Each node in the tree is a test for some attributes of then training set. Each branch corresponds to one value of the attribute. An instance is classified by starting at the root node and testing the attribute of this node and moving down the tree branch corresponding to the value of the attribute. This process is repeated recursively. With respect to text data, the conditions on the decision tree nodes are defined in terms of the text documents. A node may be subdivided to its children based on the presence or absence of a term in the document. Decision trees are generally used in combination with boosting techniques. Boosting techniques helps to improve the accuracy of the decision tree classification.

2.4. Support Vector Machines

Support Vector Machines (SVM) is a supervised classification algorithms that has been extensively used in text classification problems.SVM are Linear Classifiers. In case of text documents, the classification is based on the linear combination values of the features of the document. SVM tries to establish a good linear separator between different classes. Single SVM separates two classes, a positive class, and a negative class. It also tries to find a hyper plane with maximum distance from both the positive and negative samples. The advantage of the SVM is that, it is very robust in high dimensionality. SVM rarely needs feature selection as it selects data points that are support vectors required for the classification.

III. CLUSTERING ALGORITHMS

Clustering is a popular data mining technique used widely in text analytics. Clustering finds applications in classification, visualization, and document organization.

Clustering is the task of finding groups of similar documents in a collection of documents. The similarity is computed using a similarity function. Text clustering can be done at different levels such as clusters can be documents, paragraphs, sentences or terms. Software tools like Lemur and BOW provide implementations of common clustering algorithms. There are many clustering algorithms that can be used in the context of text data.

Text document can be represented as a binary vector by considering the presence or absence of word in the document. Or we can use weighting methods such as TF-IDF. But both these methods do not work well as text data has some very unique characteristics.

Some of the unique characteristics of text representation are as follows:

i. Text representation has a very large dimensionality, but the underlying data is sparse.

ii. Words of the vocabulary of a given collection of documents are commonly correlated with each other. We need to design algorithms which take the word correlation into consideration during clustering.

iii. Since documents differ from one another in terms of the number of words they contain, normalizing document representations during the clustering process is important.

3.1. Hierarchical Clustering algorithms

The principle behind hierarchical clustering is that it begins with all documents being treated as a cluster. Clusters are then merged based on a criterion specific to the method chosen. In all methods we begin with as many clusters as there are cases and end up with just one cluster containing all cases. By inspecting the progression of cluster merging it is possible to isolate clusters of cases with high similarity. The hierarchy can be constructed top-down called divisive or bottom-up called agglomerative fashion. Hierarchical clustering algorithms are Distanced-based clustering algorithms. There are two types of measure: similarity coefficients and dissimilarity coefficients. In the top-down approach we begin with one cluster which includes all the documents, then we recursively split this cluster into sub-clusters. In the agglomerative approach, each document is initially considered as an individual cluster. Then successively the most similar clusters are merged together until all documents are form a single cluster.

There are four different merging methods for agglomerative Clustering Algorithms:

i) *Single Linkage Clustering (SLINK):*

Each case begins as a cluster. Initially find the two most similar clusters by looking at the similarity coefficients between pairs of cases. The clusters with the highest similarity are merged to form the nucleus of a larger cluster say A and B. The next cluster to be merged with this larger cluster is the one with the highest similarity coefficient to either A or B. The process is repeated for all cases.

ii) *Group-Average Linkage Clustering:*

This method is a variation on simple linkage. We begin by finding the two most similar cases. These two cases form the nucleus of the cluster. At this stage the average similarity within the cluster is calculated. To determine which case is added to the cluster we compare the similarity of each remaining cases to the average similarity of the cluster. The next case to be added to the cluster is the one with the highest similarity to the average similarity value for the cluster. Once this third case has been added, the average similarity within the cluster is re-calculated.

iii) *Ward’s Method:*

Ward’s method, is considerably more complex than the simple linkage method. The aim in Ward’s method is to join cases into clusters such that the variance within a cluster is minimized. To do this, each case begins as its own cluster. Clusters are then merged in such a way as to reduce the variability within a cluster. Two clusters are merged if this merger results in the minimum increase in the error sum of squares. At each stage the average similarity of the cluster is measured. The difference between each case within a cluster and that average similarity is calculated and squared. The sum of squared deviations is used as a measure of error within a cluster. A case is selected to enter the cluster if it is the case whose inclusion in the cluster produces the least increase in the error.

iv) *Complete Linkage Clustering(CLINK) or Furthest Neighbor*:

It is a variant of Simple Linkage it is also called the furthest neighbor. We begin the procedure by taking two cases A and B with the highest similarity in terms of their correlation or average Euclidean distance. These two cases form the nucleus of the cluster. Rather than look for a new case that is like either A or B we look for a case that has the highest similarity score to both A and B. The case with the highest similarity to both A and B is added to the cluster. This method reduces dissimilarity within a cluster because it is based on overall similarity to members of the cluster rather than similarity to a single member of a cluster.

3.2. k-means clustering

k-means clustering is one the partitioning algorithms which is widely used in the data mining. The k-means clustering, partitions n documents in the context of text data into k cluster representations around which the clusters are built. The main disadvantage of k-means clustering is that it is very sensitive to the initial choice of the number of k. There are some techniques used to determine the initial k, e.g. using another lightweight clustering algorithm such as agglomerative

ALGORITHM:k-means clustering algorithm

Input: Document set D, similarity measure S, number k ofcluster
Output: Set of k clusters

```
initialization
Select randomly k data points as starting centroids.
while not converged do
Assign documents to the centroids based on the closestsimilarity.
Calculate the the cluster centroids for all the clusters.
end
return k clusters
```

IV. CONCLUSION

In this paper we have provided an overview of some of the most fundamental algorithms and techniques that are used in the text analytics. Though it is not possible to describe all different methods and algorithms thoroughly, it gives a brief overview of current innovations in the field of text mining. Text mining is essential to scientific research given the very high volume of scientific literature being produced every year. Thus, processing and mining this massive amount of text is of great interest to researchers. Discovering hidden knowledge is an essential requirement for many corporate as well. In this short survey, the idea of text mining has been introduced and several techniques available have been presented. As this is a new area of research, there are many potential research areas in this field of Text mining. Some of the challenges that are faced by text miners include finding better intermediate forms for representing the outputs of information extraction. Mining texts in natural languages is a challenge. Many text mining tools have issues with multilingual documents. Understanding the concept of context sensitivity with words in a document is also a case to be considered.

REFERENCES

[1] Prof.S M.Inzalkar,Jai Sharma, A Survey on Text Mining-techniques and application,International Journal of Research In Science & Engineering e-ISSN: 2394-8299 Special Issue: Techno-Xtreme 16 p-ISSN: 2394-8280.

[2] Mehdi Allahyari, et All, A Brief Survey of Text Mining: Classification, Clustering andExtraction Techniques,KDD Bigdas, August 2017, Halifax, Canada.

[3] Divya Nasa, “Text Mining Techniques- A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering , ISSN: 2277 128X Volume 2, Issue 4, April 2012 pp 51-540, in IJARCSSE

[4] S.Niharika, V.Sneha Latha, D.R.Lavanya,” A SURVEY ON TEXT CATEGORIZATION”, International Journal of Computer Trends and Technology- volume3Issue1-2012.

[5] Charu C Aggarwal and ChengXiang Zhai. 2012. Mining text data. Springer.

[6] Mehdi Allahyari and Krys Kochut. 2016. Discovering Coherent Topics withEntity Topic Models. In Web Intelligence (WI), 2016 IEEE/WIC/ACM InternationalConference on. IEEE, 26–33

[7] S. Jusoh & H. Alfawareh,” Techniques, Applications and Challenging Issue in Text Mining”, International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN (Online): 1694-0814