International Journal of Engineering & Technology, 7 (3.8) (2018) 11-15

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Review of Different Text Categorization Techniques

Anubhav Aggarwal[#1], Jasmeet Singh[*2], Dr. Kapil Gupta[#3]

Department Of Computer Applications,
National Institute Of Technology, Kurukshetra, India
[1]anubhavagg93@Gmail.Com,[2]jasmeet45.Js@Gmail.Com ,[3]kapil@Nitkkr.Ac.In

Abstract

In this paper, we focus on a major internet problem which is a huge amount of uncategorized text. We review existing techniques used for feature selection and categorization. After reviewing the existing literature, it was found that there exist some gaps in existing algorithms, one of which is a requirement of the labeled dataset for the training of the classifier.

Keywords— Bayesian; KNN; PCA; SVM; TF-IDF

1. Introduction

Digital documents are available in large number and these documents are continuously increasing. This web of digital documents contains emails, conference materials, journals, eBooks etc. People search for information in these digital documents rather than accessing just paper sources like books, newsgroups, magazines, etc. However, this web of digital documents lacks organization which makes it difficult to manage. Text categorization is recognized as one of the important techniques to manage these digital documents.

R. Jindal [1] defines text categorization as "Text categorization is a task of assigning one or more predefined categories to the analyzed document, based on its content". Text classification is the process of labeling a document under a defined set of labels. Let us assume,A_i is a document of the entire set of documents A and $\{C_1, C_2, C_3, C_4, \ldots., C_n\}$ is the set of all the labels to be assigned. The text classifier assigns one label $C_j$to the document $A_i$depending on the content of the document presented. The document depending upon their content can be labelled in one class or more than one class. "If a paper is labelled under one class, that paper will be considered as "single-label" paper and if a paper is labelled under more than one class, that paper is considered as "multi-label" paper"(Wang & Chiang, 2011).

The text categorization problem can be achieved using the machine learning process. The term Machine learning refers to the automated detection of the meaningful patterns in data. **Machine learning** is an extension of artificial intelligence (AI) that provides the ability to auto learn and improvises from past experiences without being explicitly programmed. **Machine learning** focuses on the learning of computer software that can get access to the data and use it to learn for identification of the problem that can come front and possible solutions that can be used to outlaw that problem.

Machine learning is commonly divided into two parts: Supervised and Unsupervised Machine Learning.

In the **predictive or supervised learning approach**, the primary point is to choose a steering from inputs x to yields y , given a named set of info yield sets $D = \{(X_i, Y_i)\}N_i = 1$ (where D is known as the preparation set, and N is the quantity of training cases). In the least complex setting, each preparation input $x_i$is a D-dimensional vector of numbers, say, first and second element of a set. These are known as highlights, qualities or covariates. Be that as it may, $x_i$could be a compound organized object, for example, a picture, a paper, mail content, an atomic shape, a chart, and so on.

So also, the form of the yield or reaction variable can on a fundamental level be anything, however, most techniques accept that y_i is a categorical variable from some limited set $y_i \in \{1, \ldots, C\}$, (for example, male or female), or that y_i is a genuine esteemed scalar, (for example, pay level). At the point when y_i is categorical, the issue is known as grouping or example acknowledgment, and when y_i is genuine esteemed, the issue is known as relapse. Another variation, known as ordinal relapse, happens where mark space Y has some characteristic ordering, for example, grades $A- F$.

The second fundamental sort of machine learning is the **descriptive or unsupervised learning approach**. Here we are just given information sources,$D = \{x_i\}N_i = 1$, and the objective is to discover "intriguing examples" in the information. This is some of the time called information revelation. This is a substantially less very much characterized issue, since we are not advised what sorts of examples to search for, and there is no undeniable error metric to utilize (dissimilar to supervised realizing, where we can look at our expectation of y for an offered x to the watched esteem).

The text classification process can be divided into five major components:

Acquiring the dataset

- Document representation
    1. Pre-processing
        (i) Removing Stop-words
        (ii) Stemming
    2. Bag of Words representation
- Feature Selection

- Training the text classifier
- Evaluation of the Text Classifier

2. Motivation

These days, a tremendous amount of information is being produced each one second making a colossal bunch of information on the web which is being utilized by each other individual getting to the web, as indicated by a report led by an IDC and EMC, 40 Zettabytes data would be created on the web by 2020, bringing about a 50-crease development from the earliest starting point of 2010 [2] which will, in the long run, increase the inquiry time radically. The Computer World magazine expresses that unstructured data may represent more than 70%– 80% of all information in associations. Unstructured data is normally text-based. What's more, there is a developing need to classify this content. To comprehend such an issue organized classified content information will help the client to get the information all the more precisely at a speedier pace. Text categorization has many applications as follows:

- The top-level official of any institute or any organization receives a lot of Emails. While most of the emails are not required to be seen by the official but are referred to some department heads, hence manual work is required to manually analyze those emails to send those to relevant departments of the Institute. The Automatic categorization of those emails can reduce manual work to great extent.
- Platforms such as E-commerce and news agencies can use Text categorization technology which could help them to categorize content and products.
- Text categorization can help us to find panic conversations on Facebook, Twitter and so on. The authorities can respond to those situations quickly.
- Any government, non-government organizations, researchers can use structured categorized text for faster information processing.

3. Literature Review

Three major components of the text categorization are feature selection, selecting a classifier for training, and finally evaluating its performance.

A. Feature Selection

Feature selection is the process of extracting importance words or terms from a text document to be considered as an Input to the classifier. Most of the time some words of text documents are not equally important as others. For example, "Most scientists think that butterflies use the position of the sun in the sky as a kind of compass that allows them to determine which way is north". This sentence contains a few important words (Butterflies, scientists, direction, compass) and a few unimportant words (Most, think, kind, sky, and determine). Working on only relevant features also decreases the computation time. The two most common methods used for feature selection are:

1) TF-IDF (Term Frequency – Inverse Document Frequency)*[3]:* This technique can be used to extract most relevant words or features from a document. The Term Frequency (TF) part is very similar to the bag of words. What this means is that each term or word is going to be up-weighted by how often it occurs in a document. For example, if we have a word that occurs ten times, it is going to have ten times as much weight as a word that occurs only once.

The idea of Inverse Document Frequency (IDF) is that the word also gets a weighting that is related to how often it occurs in the corpus as a whole, in all the documents put together.

The weight of a word q is calculated such as

- Weight is low if a word is having a very low frequency in the document.
- Weight is 0 if a word doesn't exist in the document.
- Weight is high if a word exits a high number of times in the document, and a low number of times in other documents in the training set.

$$tf - idf(q,d) = f(q,d) * \log\left(\frac{N}{1+cf(q)}\right) \quad (1)$$

where d is the document, q is a word in d, $f(q,d)$ is the frequency of q in d, N is the total number of documents in the training dataset, $cf(q)$ is cumulative frequency of q in all the training dataset. The weight of word q equals to its $tf - idf$ score. After calculating the $tf - idf$ score of all the words, we can select the words with highest weights as the features of the document to be classified.

2) ***PCA (Principal component analysis):*** PCA is a method for compressing a lot of data into something that captures the essence of data [4] [5]. PCA is a linear transformation algorithm. Like TD-IDF, this technique is also used extensively in text categorization for selecting important words from a document. PCA takes a dataset with a lot of dimensions or features and gives us the most relevant features.

For example, we have a dataset X containing m documents and each document A_i is represented as a vector of terms. Let A_1be a document such that $A_1 = [a_0, a_1, a_2, a_3, a_4, a_5, \dots\dots\dots, a_n]1 * n$, where $a_0, a_1, a_2, a_3, a_4, a_5, \dots\dots\dots, a_n$ are the number of occurrences of words or terms in a particular document A_i Similarly for $A_2, A_3, A_4, A_5, \dots\dots\dots, A_n$documents are represented by a vector.

The dimension of X is $m * n$. The next step is to calculate the eigen-decomposition of the covariance matrix X_t. The eigen decomposition finds us a set of eigen values and eigen vectors which can be used to describe our data. The features or words which have greater or lesser than 0 values in eigen vectors can be used as important words. The centroid of the graph is also shifted according to the set of features which are being extracted using the method. The shifting of the centroid in the process helps in determining the use of positive features during the classification process.

B. Text classification

After feature selection, different classifiers are available for training using the dataset. After the training phase is completed, we can use the trained classifier to predict the categories of documents. [1] [3]

1) ***Decision tree:*** Decision tree is a categorization technique which reconstructs the training dataset in the form of the tree-like structure where the new document which needs to be categorized is created as the root node and the child's of the corresponding nodes of the tree declares as the categories in which they need to be categorized[6]. The decision tree does the binary classification at each node of the constructed tree comprises of only two child nodes which can be further classified. It can be used as the multi-level classifier by classifying the document at multiple levels the decision tree. The main advantage of the of the Decision tree classifier is that the output tree that is generated by the decision tree is quite easily recognizable. The main disadvantage of the decision tree classifier refers to as the "over-fitting". 'Over-Fitting' develops when the learning constraints repeatedly construct a hypothesis that helps in reducing training dataset error and ignoring the

increase in the test set error. The two approaches that can be used for decreasing the "over-fitting" in the decision tree are

- Pre-pruning
- Post-pruning

Pre-pruning refers to the condition in the tree model that suggests stopping the growth of the tree classifier before, it perfectly classifies the training dataset provided to the classifier.
Post-pruning refers to the method in which a document is perfectly classified once and afterward the tree is post pruned.
The post-pruning method is being often used as it is easier to go back a step than guessing the ending of the characterization model in the tree.

***2) K-NN(K Nearest Neighbour)*:** KNN being a lethargic learning calculation, it forbids from being connected to ranges where the dynamic arrangement is required for substantial storehouse [7]. It is a case-based learning strategy. The algorithm assumes that it is possible to classify documents in the Euclidean space as points the distance between two points in the plane with coordinates $p = (x, y)$ and $q = (a, b)$ can be calculated

$$d(p,q) = d(q,p) = \sqrt{((x-a)^2 + (y-b)^2)} \quad (2)$$

Conceptually, every training document x called an instance is depicted as the vector length, the length of the directed words length

$$< w_1(x), w_2(x), w_3(x), w_4(x) \dots \dots \dots w_n(x) > \quad (3)$$

Where $w_j(x)$ is the weight of the j_{th} term, this weight is being used and adjusted according to the different aspects such as feature or a particular score assigned to a feature to help dividing the document into a certain set of class. The easiest way of assigning the weight to any word is by declaring the same weight to all valuable feature extracted from the document as 1 and others as 0 which can further contributes in the non-weighted feature approach.

3) *Bayesian*: The Naive Bayes text classifier is a simple and non-iterative text classifier which is based on the Bayes' theorem which concludes that a single extracted feature from the document is purely independent of any other feature that exists in the document. It assumes each feature independently irrespective of any connection of them with the other feature which sometimes gets used for the condition where a large amount of dataset is being used in order to train the classifier.

Naive Bayes classifier is not a single algorithm based but is a combination of a n number of algorithms clubbed together having the common principle for the classification of the document. They are majorly of four types: Gaussian naive bayes, multinomial naive bayes, Bernoulli naive bayes and semi supervised parameter estimation.
Naive Bayes classifier algorithm is based on the conditional probability model that is if a document is being provided for the classification are represented by a vector $A = (a_1, a_2, a_3, a_4, \dots\dots, a_n)$ representing n no. of features which are being independent from each other in the introduced document, it is assigned to this instance probabilities.[8][9]

$$p(C_k \parallel a_1, a_2, a_3 \dots \dots . a_n) \quad (4)$$

4) *Vector-based methods* -There are mainly two types of the vector-based classification technique: The centroid algorithm and the support vector machines [10][11]. In the two defined techniques, the centroid technique is the easier. Amid the learning stage, just the normal component vector for every classification is ascertained and set as centroid-vector for the class. This strategy is additionally unseemly if the quantity of classifications is extensive.

The second vector based method is Support Vector Machine which is a classification technique based on splitting the data in such a way that it intersects two or more feature into its classified model that fits the example which will attain the best-suited label for that corresponding document. The support vector technique split the training dataset known as the support vectors into two or more groups depending upon the classification method and the feature selection technique and also to the classifier on which it may rely the document to be classified among many available labels.
The support vectors [12] once obtained are plotted on a multidimensional graph in which each dimension specifies a single feature which is being used in order to obtain these support vectors. They are then segregated into different no of sets which are being used to label the test document. This splitting is performed by introducing a no. of hyperplanes among the support vectors which is being already extracted by the machine from the training dataset using any of the feature selection technique. The hyperplanes among the support vectors obtained from the training dataset are introduced in such a way that they are equidistant from the sets of same labeled support vectors.
For example in Fig 1, two dimensions are being used as there exist only two features which are being relied upon for the classification of the document, X_1 and X_2 being the two features which are being used to categorize the document. In this support vectors with circular notations are grouped under one label while the support vectors with square notation are grouped under the one label.

The hyperplane in the figure is introduced between the vectors so that it is equidistant from the circular as well as the square set. In this if a new document is being introduced for the classification then it will be measured by the point that the support vector lies on which side of the hyperplane and it can be classified accordingly.

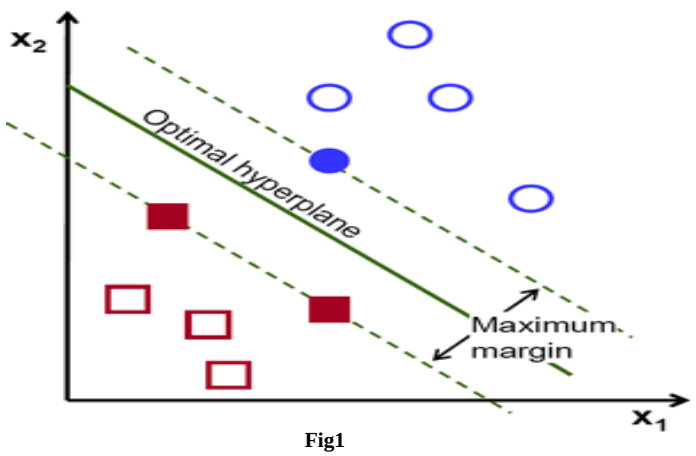

Fig1

Another example introduced that will help in understanding the importance of identifying the correct hyperplane for a test dataset. This is explained in Fig 2 below, the two features don't help in determining the correct position of the hyperplane as it requires an additional feature in order to introduce a hyperplane between the support vector which will help in categorizing the document under the correct label. As shown in the fig a single feature can also be useful in varying the plotted graph significantly. Therefore the increase in the extracted feature results in the better categorization of the document presented for the labeling

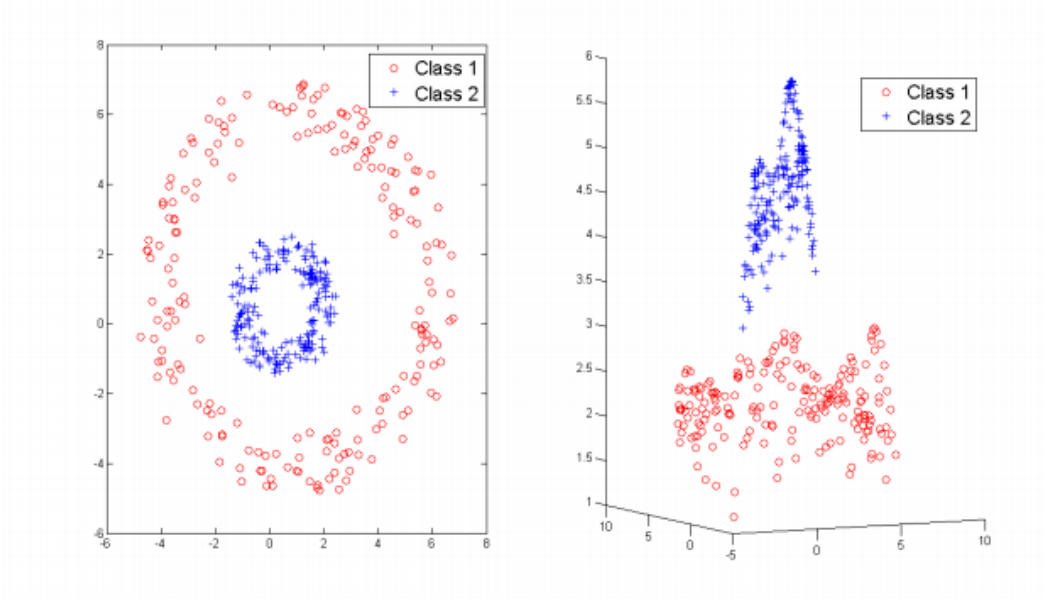

Fig 2

C. Evaluation

Another major process that needs to be performed for the text classification process is the evaluation of classifier [13] after the process is being performed in order to judge the accuracy and the reliability of the text classifier for different representation schemas and datasets.

The evaluation of the classifier is done on the basis of two methods:

- Recall
- Precision

Finally, retrieving the $F1$ score using the recall and precision determines the performance of the text classifier.

1) ***Recall:*** It is equal to the number of true positive results divided by total number of true positive and false negative results

$$Recall = \frac{T_p}{T_p + F_n} \quad (5)$$

2) ***Precision:*** It is equal to the number of true positive results divided by total number of true positive and false positive results

$$Recall = \frac{T_p}{T_p + F_p} \quad (6)$$

3). ***F1 score:*** The $F1$ score is the single measure of the classification procedure's usefulness. It considers the value of both recall and precision procedures in order to compute the score of the text classifier. The higher the $F1$ score means the higher the accuracy of the classifier, the 1 value in the $F1$ score represents the perfect score whereas the 0 represents the lowest possible score for $F1$.

$$F1 \text{ score} = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \quad (7)$$

4. Comparative Study

The comparative study in Table1 below shows that how different models of text classifier react with different types of inputs that it does also significantly depends on the representation schema which is being used for the feature selected using a number of different algorithms for the feature selection. This is also noticeable that classification technique used in the process also depends upon the schema of how the data is being represented, for example, the decision tree can only be used when the data is being represented in the tree or hierarchal graph structure as it can't be used in any other representation which makes it less dependable. Whereas KNN and SVM can easily perform under any representation technique which makes them more volatile and useful as compared by any other text classifier and also the most commonly performed text classifier in the modern day

Table 1

Results reported by	Dataset	Representation scheme	Classifier used	*F1*
[Ko et al.., 2004][14]	20 newsgroup	Vector representation with diff weights	Naive Bayes	0.8300
			KNN	0.8104
			SVM	0.8610
[Tan st al..,2005][15]	20 newsgroup	Vector representation	Naive bayes	0.8350
			KNN	0.8480
			SVM	0.8890

[Liang et al.,2006] [16]	Reuters 21578	Vector representation	KNN	0.7970
[Hao et al..,2007] [17]	Reuters 21578	Hierarchical graph structure	SVM	0.8620
			KNN	0.7888
			Decision tree	0.8790

The Table 2 [1][18] shows the value of precision and recall for SVM, KNN, and Naive Bayes

Table 2

Classifier	Recall	Precision
Naive Bayes	83.32%	83.72%
SVM	85.81%	85.81%
KNN	87.61%	90.52%

5. Gaps Identified

- SVM (Support Vector Machines), KNN (K Nearest Neighbor) and all other approaches are all being formulated for the labeled dataset. Not much of the work is being done on semi-supervised learning which makes it quite difficult to use the unlabelled data for the text categorization.
- Most of the time, collecting the labeled data set is unavailable for the training of the classifier. So a lot of initial labeling is required to gain access to the raw data to be formulated as a machine learning dataset for the further text categorization of upcoming data.

6. Conclusion and Future Scope of Work

Text classification is arising as one of the most distinguished technique to handle unstructured data. However, only supervised machine learning algorithms are used for text categorization. Semi-supervised approaches may reduce the burden to collect a large amount of labeled dataset. As the unlabelled data is available in huge amount, it could facilitate the text categorization process.

Decision tree method is not used extensively because it needs data being represented in just tree or hierarchal graph structure. Whereas KNN and SVM have been known to give greater accuracy and they can easily perform under any representation technique. In future, Ontology-based text categorization approaches can be incorporated into Machine learning approaches to get better results than ever.

REFERENCES

[1] R. Jindal, R. Malhotra, A. Jain (2015), "Techniques for text classification: Literature review and current trends", Webology, Volume 12, Number 2.

[2] John Gantz and David Reinsel. 2012. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Technical Report 1. IDC, 5 Speen Street, Framingham, MA 01701 USA.

[3] F. Sebastiani (2002), "Machine learning in automated text categorization", ACM Computing Surveys (CSUR)

[4] Y.X. Zhang, Artificial neural networks based on principal component analysis, Input selection for clinical pattern recognition analysis, Talanta73(2007)

[5] T. Jolliffe, Principal Component Analysis, ACM Computing Surveys, Springer-Verlag, 1986. pp. 1–47

[6] Mark A Friedl and Carla E Brodley. 1997. Decision tree classification of land cover from remotely sensed data. Remote sensing of environment 61, 3 (1997)

[7] Eui-Hong Sam Han, George Karypis, and Vipin Kumar. 2001. Text categorization using weight adjusted k-nearest neighbor classification. Springer

[8] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization, Vol. 752. Citeseer

[9] Tao Dong and Wenqian Shang, 2011, An Improved Algorithm of Bayesian Text Categorization, Journal of Software, vol. 6, no. 9

[10] Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20.3 (1995).

[11] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. Springer

[12] M. Allahyari (2017), "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques" , Arxiv.

[13] Aggarwal, Charu C., and ChengXiang Zhai, eds. *Mining text data*. Springer Science & Business Media, 2012

[14] KO, Y. J., Park, J., and Seo, J., "Improving text categorization using the importance of sentences", *International Journal Information, Processing and Management*, vol. 40, no. 1, January 2004, pp. 65-79.

[15] Songbo, T., Cheng, X., Ghanem, M. M., Wnag, B. and Xu, H., "A novel refinement approach for text categorization", *Proc. of 14th ACM International Conference on Information and Knowledge Management*, 2005, pp.469-476.

[16] Liang, C. Y., Guo, L., Xia, Z. H., Nie, F. G., Li, X. X., Su, L., and Yang, Z. Y. , "Dictionary-based text categorization of chemical web pages", *International Journal Information Processing and Management*, vol. 42, no. 4, July 2006, pp.1072 – 1029.

[17] Hao, P. Y., Chaing, J. H., and Tu, Y. K.., "Hierarchically SVM classification based on support vector clustering method and its application to document categorization", *International Journal Expert Systems with Applications*, vol. 33, no. 3, October 2007, pp. 1-5.

[18] CAO Jian-fang, WANG Hong-bin. 2010. Text categorization algorithms representations based on inductive learning, 2nd IEEE International Conference on Information Management and Engineering