RESEARCH ARTICLE OPEN ACCESS

Machine Learning for Real Estate Contracts – Automatic Categorization of Text

B.Gayathri[1], Prof.Kanchana Devi.V[2]
[1]PG Scholar, [2]Assistant professor
School of Computing Science and Engineering, VIT University-Chennai Campus, Chennai, India

-------------------------------------************************----------------------------------

Abstract:

This paper explains the generic strategy for automatic text classification and analyses existing solutions to major issues such as dealing with unstructured text, handling large number of features and selecting a machine learning technique appropriate to the text-classification application. This paper entitled as "machine learning for real estate contracts-Automatic text categorization" is used to categorize the lease documents automatically as per the real estate terminologies using the concept of Natural language Processing(NLP). NLP is a field of computer science, Artificial Intelligence and computational linguistic concerned with the interactions computers and human. NLP has various concepts and this project uses the concept of "Text classification". The Automatic categorization of text tool will consume an API which performs the text classification or categorization.

Keywords **—Artificial Intelligence, Machine Learning, Text classification,Automatic text categorization,Natural Language Processing(NLP).**

-------------------------------------************************----------------------------------

I. INTRODUCTION

categorization model consists the list of categories as well as the document text required to classify documents into the classes to which they are defined. For instance, a model may allow categorizing lease documents with respect to their genre. The model would include categories such as Alteration, Task and Subletting, Audit Rights, Default, Holdover, Repair and Maintenance, Insurance, Parking, Utilities, Surrender and restoration, late fee. These kinds of clauses to classified the lease document. The categorization process is based on a hybrid model that combines statistical methods with lingual rules to achieve the maximum categorization accuracy and control over the results. Thus, each category includes additional training documents and rules to categorize documents. Each category contain additional fields to provide training text and define four sets of manual rules, like relevant and irrelevant terms, positive and negative that determine the behaviour of the categorization.

II. OBJECTIVE

The aim of Automatic Text Classification System involves assigning a text document to a set of pre-defined classes, using a machine learning technique. The classification is mainly done on the basis of significant words or features extracted from the text document. This features are distinguish by using a machine learning .

III.RELATED WORK

A.Classification using Association Rule with Naïve Bayes Classifier

Text categorization using the concept of correlation rule of data mining where Naïve Bayes classifier was used to categorize text finally, showed the dependability of the Naïve Bayes classifier with association rules. But since this method disregards the concept of estimation of negative example for any specific class determination, the correctness may fall in some cases. For categorizing a text it just computes the probability of different class with the probability values of the matched set while disregarding the unmatched sets[5].

B.Classification using — Association Rule with Decision Tree

Text categorization using decision tree showed an satisfactory correctness using 76% training data of total data set , while it is probable to achieve good accurateness using only 40 to 50% of total data sets as training data[5].

C.Work on Genetic Algorithm

Text Categorization based on genetic algorithm showed acceptable performance using 69% training data, but this process requires the time consuming steps to categorize the texts.

IV PROPOSED WORK

Text categorization has recently become a full of life analysis topic within the space of knowledge retrieval. Normally text documents contain additional words. The generic strategy for text categorization is depicted. The main steps concerned are i)document pre-processing, ii) feature extraction / selection, iii) model selection, iv) training and testing the classifier.Data pre-processing decreases the size of the input text documents significantly. Ought to method those words, preprocessing is vital steps in text mining. It is used to avoid incomplete information. and determines frequent patterns After preprocessing the text documents, employing a pattern taxonomy model. Finally, to spot positive and negative documents victimization the naive Bayesian classifier.

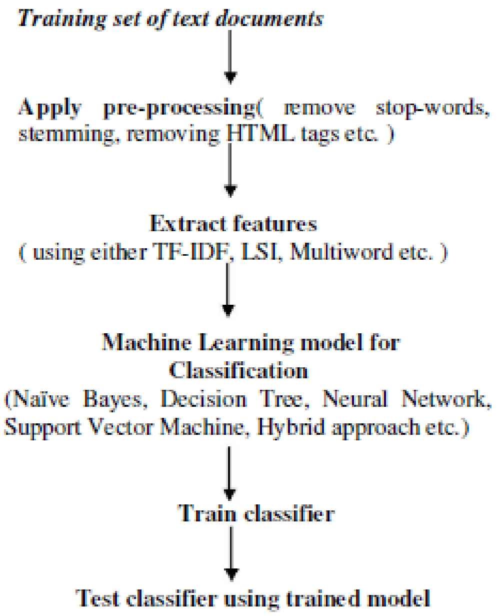

Training set of text documents
Apply pre-processing(remove stop-words, stemming, removing HTML tags etc.)
Extract features
(using either TF-IDF, LSI, Multiword etc.)
Machine Learning model for Classification
(Naïve Bayes, Decision Tree, Neural Network, Support Vector Machine, Hybrid approach etc.)
Train classifier
Test classifier using trained model

Fig.1 Generic strategy for Text classification

A. PPROCESSING

Lease documents consists of the large volume of information about the leased premises from client. As a result of this non uniformity, lease information tends to be inconsistent and noisy. The objective of this is often that it improves the standard of information about the lease and at constant time it reduces the issue of the mining method. Pre-processing could be a method of removing irregularity and incorrect information by information cleansing and information reduction techniques[7]. The preprocessing consists of two steps initial one is Stop Word removal, Tokenization, multiword grouping and Stemming method.

B. Stop word removal and Tokenization

The first common words in any text document doesn't offering any meaning in the documents, those area unit prepositions, articles ,and pro-nouns etc. These words in the lease documents are treated as stop words. Stop-words are structural words which occur often in the language of the text (for example: a“, “the“, “an“, “of“, “an“, “is“ etc. in English language), so that they are not useful for company. Stopping is the action of dropping words to their root or base procedure. As a result of each text document deals with these words that don't seem to be necessary for text mining applications, these words area unit eliminated. The Porter's stemmer is a popular algorithm, which is a suffix undressing sequence of methodical steps for stemming an English word, dumping the vocabulary of the training text by about one-third of its original size. . This method collectively reduces the information and improves the performance. Tokenization is splitting up of character flow of documents into tokens which can be used for further processing and perceptive. Tokens can be words, numbers, identifiers or punctuation depending on the use case.

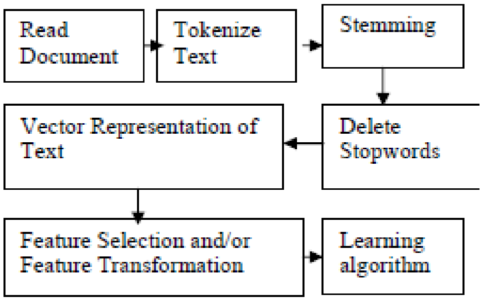

Read Document
Tokenize Text
Stemming
Delete Stopwords
Vector Representation of Text
Feature Selection and/or Feature Transformation
Learning algorithm

Fig:3 Steps involved in Preprocessing

C. Stemming Method

Stemming or lemmatization is the technique for the stepping down of words into their root. Several words within the West Germanic language are often reduced to their base kind or stem eg: agreed, agreeing, disagree, agreement and disagreement belong to agree. What is more, the names area unit reworked into the stem by removing the "s". The variation "Peter.s" during a sentence is reduced to "Peter" throughout the stemming method. The results of the removal could result in AN incorrect root.

D. Multiword grouping

multi-word methods are linguistics concerned with techniques which also try to overcome the two basic defect in categorization .polysemy (one word having many definite meanings) then .synonymy (different words having identical meaning). The LSI technique fundamentally tries to use the definition in a document building using SVD (Singular Value Decomposition) matrix operations[4]. A multi-word is a sequence of ordered words having a meaning (for example:" Information Technology ","VIT University", "Master of Computer Application").Multi-words are useful in categorization as well as disambiguation. Several methods can be used to extract multi-words from text such as the incidence approach, mutual information approach etc

E. Features Extraction

Features useful in text categorization are simple words from the language vocabulary, user nominal or selected keywords, multi-words or metadata. In text organization literature, the steps involved in feature reduction are mostly applying pre-processing such as stop-word removal , stopping etc. Text documents usually use words after a large vocabulary, nevertheless all words occurring in a document are not useful for categorization. So, Features Extraction consume proposed feature reduction techniques like TF-IDF, LSI, multi-word etc. or a combination of such techniques. The TF-IDF is a virtuously statistical method to measure the importance of a word based on its incidence of incidence in the lease document and in its relevant corpus.

F. TF-IDF

In information recovery, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is supposed to indicate how essential a word is to a document in a collection[7] .It is often used as a coefficient factor in searches of content recovery, text mining, and user modeling. The tf-idf value increases proportionally to the number of times a word looks in the document and is begin by the rate of the word in the document collection, which helps to adjust for the fact that some words appear more often.

$$\mathbf{TFIDF(t) = TF(d, t) \times IDF(t)}$$

To further differentiate them, we may count the number of times each term happens in each document; the number of times a term happen in a lease document is called its term frequency.

G. Inverse document frequency

Because of the term "a,an,or,the" is so communal, term frequency will tend to falsely expresses the documents which happens to use the word "the" more often, without giving enough weight to the more significant terms”premises”,”commencement” . The term "a,an,or,the" is not a great keyword to differentiate relevant and non-relevant documents and terms, unlike the less-common words like ”premises”,” commencement” so an inverse document frequency factor is integrated which reduces the importance of terms that occur very often in the document and increases the importance of terms that occur often[4]. Inverse Document Frequency (IDF) is used to measuring the particularity of terms in a set of documents. It assumes that a high-semantical term appears in only a few documents, while a low-semantical term is spread over many documents. The formula of IDF can be expressed by the following.

IDF(t) = log|D|/DF(t)

Where D is the set of documents in the collection and DF(t) is the document frequency, which is the number of documents where the term t appears at least once.

H. Modeling: Selection of appropriate machine learning technique for classification of text documents

Naïve Bayes is a concept in which the application of various Machine Learning Techniques to the text classification trouble like in the field of medicine, spam filtering, regard rule learning for knowledge base systems has been explored. Naïve Bayes Model works with the conditional probability which arise from well known statistical formulation “Bayes Theorem”, where as Naïve refers to “assumption” that all the attributes of the examples are independent of each other given the context of the category. Because of the independence assumption the parameters for each attribute can be learned separately and this greatly solves learning particularly when the number of attributes is large[2]. In this context of text categorization, the probability that a document d belongs to class c is calculated by the Bayes theorem as follows

P(c/d)=P(d/c) P(c)/P(d)

The estimation of P (d/c) is challenging since the number of manageable vectors d is too high. To overcome this difficulty, we use the naïve assumption that any two coordinates of the document is statistically autonomous. Using this assumption the most probable category ‘c’ can be estimated.

I. Multinominal naive bayes

The Multinomial Naive Bayes (MNB) has a number of prominent features for most text categorization model. It is straightforward and can be irrelevantly premeditated for large numbers of classes different critical classifiers. In general it is powerfull even when its statements are dishonored. MNB is a productive copy. A replica of the joint probability distribution p (w, c) of word count vectors w = [w1, ...,wN] and the class variables c : 1 <= c <= M, where N is the number of possible words and M the number of possible classes.Bayes classifiers utilize the Bayes theorem to solve the productive joint distribution dependent to a class prior p (c) and a class conditional p (w|c) models with separate parameters, so that p(wc) = p(c)p(w|c). Naive Bayes classifiers make use of the further supposition that the class conditional probabilities are self-governing. The most important strength of MNB is its scalability[1]. Preparation of a MNB is prepared by summing the counts wn establish for each pair (c, n) in training documents and normalizing these above n to acquire p(n|c).

K. MULTINOMIAL NAIVE BAYESIAN CLASSIFIER

A classic approach for text categorization is the Multinomial Naive Bayes(MNB) classifier, which models the delivery of words (features) in a documentas multinomial i.e. the possibility of a document given its class is the

multinomialdistribution . The estimated individual possibilities for
$P(X_t= x_t|C = c)$

$$\hat{P}(X_t = x_t|C = c) = \frac{N_{ct} + \alpha}{N_c + \alpha|V|}$$

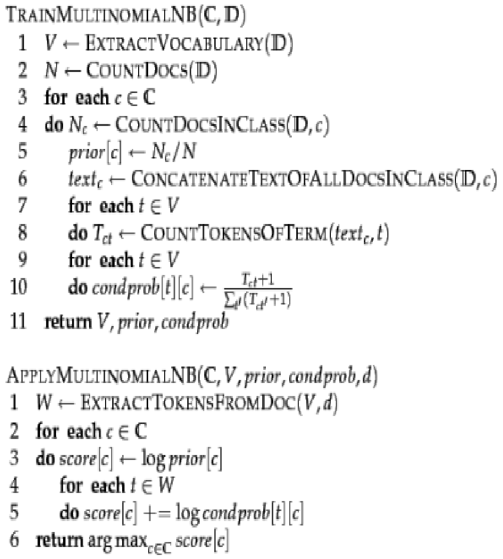
Where jV j is the size of thewords, $N_{ct}$is the number of times feature t emerge in class c in the training set, and $N_c$is the total sum of features ofclass c. The constant_ is a smoothing constant used to hold the difficulties ofoverfitting and the edge case where N_{ct}= 0. Setting _ = 1 is known as Laplacesmoothing . Applying this estimate to the decision rule we get

$$\begin{aligned} d(X_1, \ldots, X_n) &= \arg\max_c P(C = c) \prod_{i=1}^{n} \hat{P}(X_i = x_i|C = c) \\ &= \arg\max_c P(C = c) \prod_{i=1}^{n} \frac{N_{ci} + \alpha}{N_c + \alpha|V|} \end{aligned}$$

In the literature the minimum-error classification rule is more often used

$$\begin{aligned} d(X_1, \ldots, X_n) &= \arg\max_c \left[\log P(C = c) + \sum_{i=1}^{n} f_i \hat{P}(X_i = x_i|C = c) \right] \\ &= \arg\max_c \left[\log P(C = c) + \sum_{i=1}^{n} f_i \frac{N_{ci} + \alpha}{N_c + \alpha|V|} \right] \end{aligned}$$

where f_i is the frequency of word x_i.



```
TrainMultinomialNB($\mathbb{C}$, $\mathbb{D}$)
$V \leftarrow$ ExtractVocabulary($\mathbb{D}$)
$N \leftarrow$ CountDocs($\mathbb{D}$)
for each $c \in \mathbb{C}$
do $N_c \leftarrow$ CountDocsInClass($\mathbb{D}$, $c$)
    $prior[c] \leftarrow N_c/N$
    $text_c \leftarrow$ ConcatenateTextOfAllDocsInClass($\mathbb{D}$, $c$)
    for each $t \in V$
    do $T_{ct} \leftarrow$ CountTokensOfTerm($text_c$, $t$)
    for each $t \in V$
    do $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
return $V, prior, condprob$

ApplyMultinomialNB($\mathbb{C}$, $V$, $prior$, $condprob$, $d$)
$W \leftarrow$ ExtractTokensFromDoc($V$, $d$)
for each $c \in \mathbb{C}$
do $score[c] \leftarrow \log prior[c]$
    for each $t \in W$
    do $score[c]$ += $\log condprob[t][c]$
return $\arg\max_{c \in \mathbb{C}} score[c]$
```

Fig:3 Multinomial Naïve bayes Algorithm

V.SYSTEM ARCHITECTURE

The fig.4 shows the small print of planned system, this proposed system takes input as text documents. The particular step is to get free of uneven information from lease documents mistreatment stop words, tokenisation, multiwords grouping and stemming. The preprocessed documents are split into a collection of paragraphs. The prevailing terms are extracted mistreatment the TF-IDF model. This model follows two steps; initial it describes a way to extract the frequently occuring terms from the text documents. Second it describes a way to update the discovered patterns effectively for performing the information discovery from the lease documents. Text categorization is that the task of automatically sorting a collection of documents into classes from a predefined categories. This task comes below the class of knowledge retrieval (IR) and machine learning (ML). Text classification ways are naive Bayesian, support vector machine and call tree. Here the multinomial naive theorem is engaged to categorize the documents as per the specified real estate category. As a result, even once words are dependent, every word contributes proof singly. Therefore the importance of the weights for categories with robust word dependencies is larger than categories with fragile word dependencies ,thedistribution of words for the period of a collection as a multinomial. A corpus is delighted as a ordering of words and it is unspecified that every word position is produced independent of every one special. Whereas naive Bayes is easy would model a document because the existence and lack of explicit words however multinomial naive Bayes explicitly models the word counts.

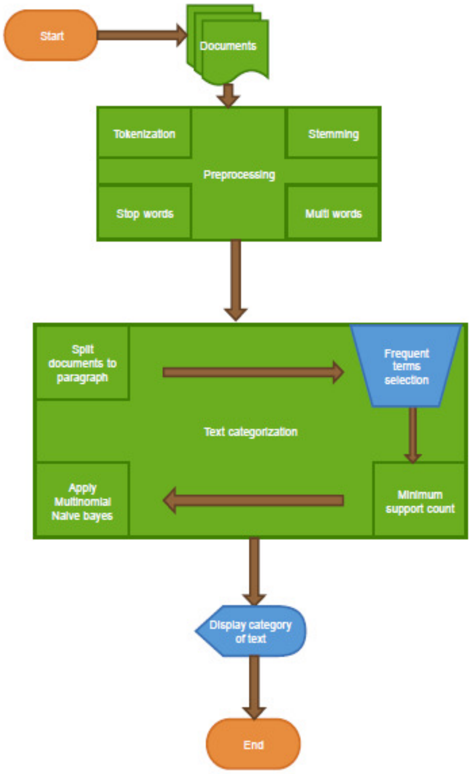


Fig:4 System Architechture diagram

VI.CONCLUSION

Document categorization could be a growing involvement within the analysis of text mining. Characteristic the documents decently into stated classes still present the challenge, traceable to the massive and vast quantity of options within the document. There are several words within the documents, thus several terms are captured from these documents and thousands of terms are found. However, there are some terms that are assistive and tedious to the results. It is necessary to find and construe with that that options are helpful and fundamental. Preprocessing and frequent pattern generation is two necessary steps to enhance the mining quality. Multinomial Naive Bayes event model is extra appropriate once the dataset is large. Whereas easy naive Bayes categorize the document supported the presence and absence of definitive words however multinomial naive Bayes expressly models the word counts.

REFERENCES

[1] Sumathi Subramanian, ”Document classification using Multinomial Naïve Bayesian Classifier” International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 5, May 2014

[2] Mr.C.Mani , J.Jayasudha , “Machine Learning For Real Estate Contracts Automatic Categorization of Text”,International Journal of Computer Techniques –– Volume 3 Issue 2, Mar-Apr 2016

[3] C. Navamani, J.Sindhuja, “Machine Learning For Real Estate Contract-Entity Recognition Using Search Engine”, International Journal of Computer Techniques -– Volume 3 Issue 2, Mar- Apr 2016

[4] Pritam C. Gaigole , L. H. Patil , P.M Chaudhari ,” . Preprocessing Techniques in Text Categorization”, National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2013) Proceedings published by International Journal of Computer Applications

[5] Mita K. Dalal,Mukesh A. Zaveri,” Automatic Text Classification: A Technical Review”, International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011

[6]VrushaU.Suryawanshi, PallaviBogawar, PallaviPatil, PriyaMeshram, KomalYadav, Prof. Nikhil S. Sakhare,” Automatic Text Classification System”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 2, February 2015 419 ISSN: 2278 – 1323

[7] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya,” Preprocessing Techniques for Text Mining - An Overview”,International Journal of Computer Science & Communication Networks,Vol 5(1),7-16.

[8] R.Mohana, S.Sumathi,”Document classification using Multinomial Naïve Bayesian Classifier”,International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 5, May 2014