Advances in Machine Learning for Text Classification: A Survey

Daniel Martinez
Department of Civil Engineering, University of Buenos Aires, Argentina

ABSTRACT

This research focuses on Text Classification. Text classification is the task of automatically sorting a set of documents into categories from a predefined set. The domain of this research is the combination of information retrieval (IR) technology, Data mining and machine learning (ML) technology. This research will outline the fundamental traits of the technologies involved. This research uses three text classification algorithms (Naive Bayes, VSM for text classification and the new technique -Use of Stanford Tagger for text classification) to classify documents into different categories, which is trained on two different datasets (20 Newsgroups and New news dataset for five categories).In regards to the above classification strategies, Naïve Bayes is potentially good at serving as a text classification model due to its simplicity.

KEYWORDS: Text Classification, Information Retrieval, Naive Bayes Classifier, Vector Space Model Text Classification, Part of Speech Tagging, Natural Language Processing.

INTRODUCTION

The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. Which include unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify the documents and discover patterns from different types of the documents .

Text classification (TC) is an important part of text mining, looked to be that of manually building automatic TC systems by means of knowledge-engineering techniques, i.e. manually defining a set of logical rules also called as training , that convert expert knowledge on how to classify documents under the given set of categories. For example would be to automatically label each incoming news with a topic like “sports”, “politics”, or “business”. A data mining classification task starts with a training set D = (d1….. dn) of documents that are already labeled with a class C1, C2 (e.g. sport, politics). The task is then to determine a classification model which is able to assign the correct class to a new document d of the domain.

Basically there are two stages involved in Text Classification. Training stage and testing stage. As explained in the above paragraph, in training stage documents are pre-processed and are trained by a learning algorithm to generate the classifier. In testing stage, a validation of classifier is performed. There are many traditional learning algorithms to train the data, such as Decision trees, Naïve-Bayes (NB**),** Support Vector Machines (SVM**),** k-Nearest Neighbor (kNN**),** Neural Network (NNet),etc.

In this research, we study the problem of text classification, that is classifying the news documents into different categories based on three different supervised algorithms namely Naive Bayes classifier, Vector Space Model for text classification and a new technique -Use of Stanford Tagger for text classification. We have tried to compare the efficiency and accuracy of the algorithms to analyze the effectiveness of each algorithm. The research has been carried out on two different datasets namely 20Newsgroup and New Dataset of news for five categories.
This paper is organized as follows. Related work on text classification and VSM in Section 2. Methodology of text classification in section 3 followed by the three text classification methods in section 4.Section 5 provides experimental setup and results followed by section 6 which concludes paper along with direction for future work followed by acknowledgement and references.

RELATED WORK

Our work is closely related to Vandana Korde and C Namrata Mahender[1] on text classification and classifiers. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization, Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents . They compare different text classifier for their efficiency.

One more related research paper to my research was of Y. H. LI and A. K. Jain [2] says that paper investigate four different methods for document classification: the naive Bayes classifier, the nearest neighbour classifier, decision trees and a subspace method. These were applied to seven-class Yahoo news groups (business, entertainment, health, international, politics, sports and technology) individually and in combination. Study of three classifier combination approaches: simple voting, dynamic classifier selection and adaptive classifier combination. Experimental results indicate that the naive Bayes classifier and the subspace method outperform the other two classifiers on data sets. Combinations of multiple classifiers did not always improve the classification accuracy compared to the best individual classifier. Among the three different combination approaches, adaptive classifier combination method introduced performed the best.

Mita K. Dalal and Mukesh A. Zaveri research paper [3] explains Automatic Text Classification is a semi-supervised machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and extracted features which is also a close related paper to my research. This paper explains the generic strategy for automatic text classification which includes steps such as pre-processing, feature selection using various statistical or semantic approaches, and modeling using appropriate machine learning techniques (Naïve Bayes, Decision Tree, Neural Network, Support Vector Machines, Hybrid techniques). This paper also discusses some of the major issues involved in automatic text classification such as dealing with unstructured text, handling large number of attributes, examining success of purely statistical pre-processing techniques for text classification v/s semantic and natural language processing based techniques, dealing with missing metadata and choice of a suitable machine learning technique for training a text classifier.

In one more research regarding the vector space model by Jitendra Nath Singh and Sanjay Kumar Dwivedi[4] states that different approaches of vector space model were used to compute similarity score of hits from search engine and more importantly, it is felt that this investigation will lead to a clearer understanding of the issues and problems in using the vector space model in information retrieval and our work intends to discuss the main aspects of Vector space models and provide a comprehensive comparison for Term- Count model, Tf-Idf model and Vector space model based on normalization.

METHODOLOGY

The methodology for the study of text classification is presented in figure 1. Documents, pre-processing, indexing, Feature extraction, classification algorithm and performance measure are detailed in this methodology, while text classification models which are used in this research are explained in the further section.

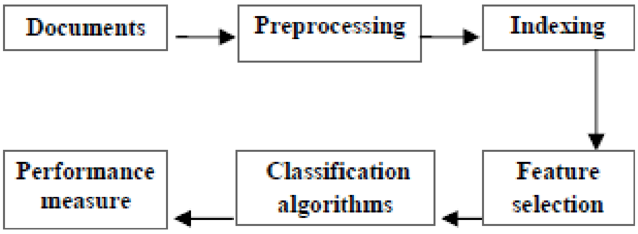

Fig.1. Document Classification Process

Documents Collection

This is first step of classification process in which we are collecting the different types (format) of document like html, .pdf, .doc, web content etc.

Pre-Processing

The first step of pre-processing which is used to presents the text documents into clear word format. The documents prepared for next step in text classification are represented by a great amount of features. Commonly the steps taken are:

Tokenization: A document is treated as a string, and then partitioned into a list of tokens.
Removing stop words: Stop words such as “the”, “a”, “and”, etc are frequently occurring, so the insignificant words need to be removed.

Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute

Indexing

The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector The Perhaps most commonly used document representation is called vector space model (SMART) vector space model, documents are represented by vectors of words. Usually, one has a collection of documents which is represented by word by word document Matrix.

Feature Selection

After pre-processing and indexing the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word.

Classification

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi supervised methods. From last few years, the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Rocchio’s.

Performance Evaluations

This is Last stage of Text classification, in which the evaluations of text classifiers is typically conducted experimentally, rather than analytically. The experimental evaluation of classifiers, rather than concentrating on issues of Efficiency, usually tries to evaluate the effectiveness of a classifier, i.e. its capability of taking the right categorization decisions. Many measures have been used, like Precision and recall, fallout, error and accuracy.

DATASET SOURCES

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It was originally collected by Ken Lang. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each, while others are highly unrelated.

The other new News data set is a collection of around 500 news documents from different news papers partitioned evenly across 5 different newsgroups namely Business, Nation, Sports, Technology and World news. It was originally collected for training and testing purpose of this project. Some of the news is very closely related to each other while others are highly unrelated as in the 20 Newsgroup.

This research has used around nine different newsgroups from 20 newsgroup dataset with all together more than one thousand two hundred documents for training and all news from new dataset of five categories. Testing on these classifiers is done using fifty random documents (news)which are chosen randomly from weband has no relation with the training data.

text classification methods

Naïve Bayes Classification Method

For some types of probability models, naive Bayes classifiers (NB) can be trained very efficiently in a supervised learning setting. Naive Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the colour, roundness and diameter features.

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so less training data is needed. And even if the NB assumption doesn't hold, a NB classifier still often performs surprisingly well in practice.

Algorithm

A) Training Phase

Step 1: Training System

a) Applying Pre-processing methods for the data present in each categories. i.e. stop word removal, Stemming.

b) Tokenizing the data and storing the words along with its category in the database.

Step 2: Probability Calculations

a) For each unique word in the categories, we try to find out the probability of each unique words belonging to that particular class.

b) Formula for the probability is as follows:
PrS[i] = Probability that word belongs to Document/class A(any category).
PrC[i] = Probability that word belongs to Document/class B(any category).
PrS[i] = [freq[i]/ freq[i]+freq2[j]].
PrC[i] = [freq2[j]/ freq[i]+freq2[j]].

c) Calculate probabilities for each category and store it in database.

B) Testing Phase

Step 1: Applying Pre-processing methods for the data present in test document. I.e. stop word removal, stemming.

Step 2: Tokenizing the data and storing the words along with its category in real time memory.

Step 3: Checking each unique word from test document with the word probability stored in database. If that word occurs in that particular category then probability of that word is added and this is repeated for all the words in that test document.

Step 4: Probabilities of each category is calculated and the one with the highest probability is the correct match.

VECTOR SPACE MODEL FOR TEXT CLASSIFICATION METHOD

Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System. The vector space model procedure can be divided in to three stages. The first stage is the document indexing where content bearing terms are extracted from the document text. The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure. Documents are also treated as a “bag” of words or terms. Each document is represented as a vector. However, the term weights are no longer 0 or 1. Each term weight is computed based on some variations of TF or TF-IDF scheme.

Algorithm

A) Training Phase

Step 1: Training System

a) Find the total number of documents present for processing which is depicted by N.

b) Apply pre-processing methods for the data present in categories .That is, tokenization, stop word removal and stemming.

c) Find the inverse document frequency (idf) with respect to each word of the category.
$idf_i = N/ df_i$
Where N=total number of document and df_i=In how many document the word occurs.

d) Find the term frequency (tf) with respect to each word of the category. Term frequency is the number of times a term occurring in that document.

e) Calculate the weight of the word by multiplying tf * idf values of that word with respect to the category and store it in the database.

B) Testing Phase

Step 1: Applying Pre-processing methods for the data present in test document. That is, stop word removal, Stemming and tokenization.

Step 2: Compare the words from the test document to the words in the database according to the category.

Step 3: Add the weights of those words which are present in the test document.

Step 4: Calculate the total of weights for each category and the category with the highest weight is the correct match.

USE OF STANFORD TAGGER FOR TEXT CLASSIFICATION METHOD

Vector A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. Once performed by hand, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags.

Algorithm

A) Training Phase

Step 1: Training System

a) For each sentence in the document we try to find out the noun , verb and adjective and store it in the Database.

b) This is done with the help of Stanford Tagger module.

c) Find the same for all the documents in each category and store appropriately under each category.

B) Testing Phase

Step 1: Find out the noun, verb and adjective for each sentence of the test file.
Step 2: This is done with the help of Stanford Tagger module.
Step 3: Compare this noun, verb and adjective with the one stored in database.
Step 4: Whichever category has maximum matches of this noun, verb, and adjective will be the correct match.

EXPERIMENTAL RESULTS

The performance of a classification algorithm is affected greatly by the quality of data source. Irrelevant and redundant features of data not only increase the cost of mining process, but also reduce the Quality of the result in some cases.

The first dataset used in our research is the 20Newsgroups dataset downloaded from 20Newsgroups site. We pre-process the HTML news items by (i) document parsing (remove headers and tags in the HTML files) and (ii) removing stop words and low-frequency words as mentioned earlier. We have used a total of more than 1200 documents belonging to nine different classes (athiesm (At), autos (Au), computer graphics (Cg), computer osms windows (Ms), computer IBM pc hardware (Ibm), computer mac hardware (Mac), computer windows x (x), forsale (Fs) and motorcycle (Mc)) for training and a test data set (randomly collected 50 news document). The second dataset used in our research is the New news dataset downloaded from different news paper sites such as jagran and herald for this research purpose. We have used a total of around 500 documents belonging to five different classes (business (B), Nation (N), world (W), sports (S) and technology (T)) for training and a test data set as mentioned above for testing .

Using the two sets of training documents, we compared the three classification algorithms (naive Bayes classifier (NB), Vector space model for text classification (VSM) and newly derived Use of Stanford Tagger for Text Classification (POSC)) on our test data sets. Here we keep human interpretation of the result (a human being classifying the documents into different categories after having a good knowledge of what exactly the dataset is) as a gold standard so that we can compare the results of classifying algorithm with it and see how it behaves. Table 1and 2 shows the number of documents present in the datasets. Table 3 shows a comparison using these three classification algorithms.

TABLE 1.Training data for 20Newsgroups.

	Categories	**At**	**Au**	**Cg**	**Ms**	**Ibm**	**Mac**	**x**	**Fs**	**Mc**
Training Data Of 20News groups	Number Of documents	133	131	141	159	157	148	132	136	144

TABLE 2.Training data for New news group.

	Categories	**B**	**N**	**S**	**T**	**W**
Training Data Of New newsgroup	Number Of documents	99	100	101	100	99

TABLE 3.Comparison of the three classification algorithms (NB, VSM, POS)

		NB	VSM	POSC
Test Data set(50 random documents)	1. Number of Miscalculations With respect to 20Newsgroup 2. Accuracy Rate (%) (with respect to golden standard)	5 90	6 88	12 75
		NB	VSM	POSC
Test Data set(50 random documents)	1. Number of Miscalculations. With respect to New newsgroup 2. Accuracy Rate (%) (with respect to golden standard)	3 94	2 95	15 69

LIMITATIONS

For Naive bayes classifier ,if your training set is small, high bias/low variance classifiers (e.g., Naive Bayes) have an advantage over low bias/high variance classifiers (e.g., kNN or logistic regression), since the latter will over fit. But low bias/high variance classifiers start to win out as your training set grows (they have lower asymptotic error), since high bias classifiers aren't powerful enough to provide accurate models. For vector space model of text classification ,long documents are poorly represented because they have poor similarity values (a small scalar product and a large dimensionality).Search keywords must precisely match document terms; word substrings might result in a "false positive match“. Semantic sensitivity; documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match". The order in which the terms appear in the document is lost in the vector space representation. Theoretically assumes terms are statistically independent. (e.g. ignores synonymy).Missing syntactic information (e.g. phrase structure, word order, proximity information).Some words occur frequently in more documents which has least importance in that context due to which unnecessary weight will increase and may lead to inconsistency. E.g. words like good.

CONCLUSION

Text Classification is an important application area in text mining why because classifying millions of text document manually is an expensive and time consuming task. Therefore, automatic text classifier is constructed using pre classified sample documents whose accuracy and time efficiency is much better than manual text classification. If the input to the classifier is having less noisy data, we obtain efficient results.

In our study, we applied three different text classifier models namely the vector space model for text classification (VSM), the Naive Bayes Classifier model (NB) and newly implemented Use of Stanford Tagger for text classification on two different datasets namely 20 Newsgroup and New dataset consisting of comparatively less data and evaluated the resultant scores on subsets of the datasets and also on 50 random news documents. We have also considered and evaluated the result given by human interpretation for all this three methods which here we consider as gold standards. Based on this evaluation, we found cases where one approach that is NB classifier worked significantly better than remaining two classifiers. It seems that Naïve Bayes is the best classifiers against several common classifiers in term of accuracy and computational efficiency. VSM approach works better with the New newsgroup dataset as the dataset is relatively small and has less irrelevant features.

FUTURE WORK

Future work in this area should be considered as study of more supervised text classification algorithm for different datasets. Comparing the efficiency of above algorithms. Also comparing supervised text classification algorithms with semi supervised and unsupervised algorithm. Try to find a most efficient algorithm by using the Modules of the earlier studied algorithm. More detail study of “Use of Stanford Tagger for Text Classification” with respect to more tags and how these techniques can be used on word sense disambiguation.

ACKNOWLEDGMENT

This work was performed as part of a Master of Engineering thesis in the project “Use of classification techniques for word sense disambiguation” and further evaluated.

REFERENCES

[1] Y. H. Li , A. K. Jain “Classification of Text Documents” The Computer Journal, Vol. 41, No. 8, (1998).

[2] Vandana Korde, C Namrata Mahender ”Text Classification and classifiers: A survey”.ijaia,2012.

[3] Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar "A review paper on algorithms used for Text Classification" International Journal of Application or Innovation in Engineering & Management (IJAIEM) 2013.

[4] Jitendra Nath Singh ,Sanjay Kumar Dwivedi " Analysis of Vector Space Model in Information Retrieval " CTNGC Proceedings published by International Journal of Computer Applications (IJCA), (2012) .

[5] Dik L. Lee, Huei Chuang, Kent Seamons "Document Ranking and the Vector-Space Model",IEEE 1997.

[6] 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).

[7] K. Naleeni, Dr.L.Jaba Sheela,"Survey on Text Classification", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 6 July (2014).

[8] Kratarth Goel, Raunaq Vohra, Ainesh Bakshi, "A Novel Feature Selection and Extraction Technique for Classification", IEEE International Conference on Systems, Man, and Cybernetics, October 5-8, (2014).

[9] S.L.Ting,W.H.Ip, Albert.H.C.Tsang ,"Is Naïve Bayes a Good Classifier for Classification?", International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, (2011)

[10] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naïve Bayes Text Classification", IEEE transactions on Knowledge and Data Engineering, VOL. 18, NO. 11, November (2006).

[11] Ioan Pop,"An Approach of the Naïve Bayes Classifier for the document classification", General Mathematics Vol. 14, No. 4 (2006).

[12] George Tsatsaronis ,Vicky Panagiotopoulou, "A Generalized Vector Space Model for Text Retrieval based on Semantic Relatedness", Association for Computational Linguistics, Athens, Greece, 2 April (2009).

[13] Y.H. Chen,Y.F. Zheng, J.F. Pan, N. Yang,"A hybrid text classification method based on K-congener-nearest-neighbors and hypersphere support vector machine", International Conference on Information Technology and Applications, (2013).

[14] Shuzlina Abdul-Rahman, Sofianita Mutalib, Nur Amira Khanafi, Azliza Mohd Ali,"Exploring Feature Selection and Support Vector Machine in Text Categorization", IEEE 16th International conference on computational science and engineering,(2013).