Document Text Classification Using Support Vector Machine

[1]Batoul Aljaddouh, [2]Nishith A. Kotak
[1]P.G Student of Information and Communication Technology Engineering,, [2]Assistant Professor
[1]Marwadi University, Rajkot, India,
[2]Marwadi University, Rajkot, India

***Abstract* - Document categorization is in trend nowadays due to large amount of data available in the internet. There are many different classification algorithms such as (Naïve Bayes, SVM, K-means, KNN, etc.) depending on the type of classifications and other features. In this paper, we have observed the best accuracy and efficiency using Support Vector Machine approach, which is been explored in the paper. The dataset used is combine between the two datasets. First, one is BBC news articles. Second is 20 news group. We have used more than 3500 distinct articles to build classifier into eight classes (Atheism, Business, Car, Entertainment, Politics, Space, Sport and Technology). The training accuracy for classifier is 96.40% resulting in about 130-misclassified cases. The overall accuracy tested over 20 articles of each class resulting in 160 testing articles obtained is 89.70% which is comparable to many existing algorithms. The paper further describes the factors affecting the classification technique and its implementations.**

***keywords* - Document Classification, Natural Language Processing, Text Mining, Text Classification, Support Vector machine, Artificial intelligence, Machine learning**

I. INTRODUCTION

Text Mining is the trending field which deals with known or unknown information about the huge amount of data. Different subfields like Natural Language Processing (NLP), Recommendation Systems, Information Retrieval (IR), etc. deals with the extraction of the hidden information in the form of text. Document Classification is considered as one of the most widespread fields in these days, due to the amount of texts available on the internet and social networking sites. Most social media companies understand user behavior by analyzing text, comments and behavior of the users. For an instance, Facebook determines the degree of customer satisfaction as well as its trends without relying on feedback for a particular post. Google recently made searches dependent on user words to find the space that we want search within. YouTube has also made customer comments a way to suggest or block similar videos based on analyzing users' comments and behaviors. Text categorization is a wide domain to make machine able to understand and organize the document (text, document), as a part of Natural Language Processing that focus to make machine understand and analysis human language.

Natural language processing[1] is a computer domain related to information engineering, computer engineering, machine learning and artificial intelligence focus on the interactions between human's languages and computer machines, by simplicity , how the computers will understand and analyze huge amounts of human language data. Text classification or document categorization is a problem related to library science, information science and computer science. The main aim is to assign any types of text files to one or more pre-defined classes according to their content. Library science uses the intellectual categorization of texts or documents files, while the automatically categorization of texts is mainly about computer science and engineering and information science. However, we cannot separate the two problems due to exist some of intersections between their solutions researches. Some documents have images, stickers, etc. Each type has specific classification systems or models[2]. Text classification also used for sentiment analysis like-wise used in Social Networking sites (like Facebook, twitter) also to motivation analysis that is very important domain used for marketing (like Amazon).

In this work, we used supervised learning method (Support Vector Machine algorithm) to classify the articles in various categories. This algorithm has special power in this field text classification due to high dimensional of text with high degree polynomial. The aim of this work is to classify and organize some of document on system. It can be used to give overview about articles or books before you launch to read.

II. LITERATURE SURVEY

Before we go into algorithms of classification texts, we will discuss some of the research papers that were researched in this field, along with mentioning the algorithms used and the results of each research. The first research paper discussed was [3] where the author used the probabilistic model of the Naïve Bayes algorithm that mainly relies on posterior information represented by the presence of certain words in a text with known class. It used Reuter's documents as dataset and compare with two types of Naïve Bayes approaches multinomial model and multivariate Bernoulli model. They founded that the multinomial model is better than the multivariate Bernoulli model. In empirical results on five real world corpora they find that the multinomial model decreases the error by approximately 27%, and sometimes by more than 50%. Next paper we have read Text Classification is [4] In this work he has taken about 1234 Arabic documents to classified it into two classes "Literature" or "Business". They used Naïve Bayes also to build a model with gained accuracy about 87.8%. Another paper has reviewed is [5] where the researcher used Random Forest algorithm. Paper they have many features for classification including the textual

description and multimedia accompany to the articles such as videos or images. Dataset used webpages from many famous News Web Sites, namely BBC, The Guardian and Reuter. Overall, 651, 556 and 360 web pages have been retrieved from each site. The accuracy gained is 84.4%. Deep learning algorithm is also used in this approach. Another paper [6] take into account many type of neural networks to achieve the mission of classification. The dataset used is collection of articles and news. The accuracy was different for each dataset and neural networks. In General, They found that RCNN achieves better accuracy than CNN. Where the average accuracy for CNN is 70.65 and for RCNN is 72%.

One of the common powerful algorithms used in this domain is SVM (Support Vector Machine) that use many hyperplanes to separate data within features space. Paper [7] that authors used Reuter's group as dataset to classify documents to ten classes. In this paper they make comparison between different degree of svm polynomial and SVM with RBF function. The results gained for polynomial SVM was 80% and for svm with RBF 80.4% they also compared with another model but mainly paper is about SVM. Last paper reviewed is [8] that tends to classify Arabic documents into seven classes using SVM and Naïve Bayes. The accuracy was about 77.8 %. The detailed explanation of the SVM classifier used for the purpose is mentioned in the next section.

Support Vector machine [9] is a supervised learning algorithm which is useful for regression, analysis and classification. The input data must be labeled with class name that data pattern belongs to this class (may belong to more than one class). The main goal of algorithm is to predict the class for new pattern depends on classifier model (Hypotheses) build based on training patterns. The Hypotheses maybe taken different degree (linear classifier, non-linear classifier). For an instance, if we have some given pattern of data with different classes. The pattern that can be represented as vector of features p. We have to know that separate these patterns with (p-1) dimensional hyperplane. There are many hyperplanes that can perform classification task the best choice that separate the pattern with largest margin among classes. The hyperplane that has high value of margin called Maximum margin hyperplane [9].

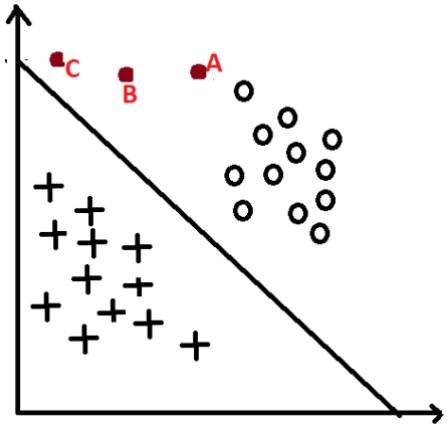

Figure 1: SVM with the training, testing data points and line of separation

SVM also deals with finding the maximum closest pattern to the line from various classes. Here for an example, given two class separated by a line of bifurcation, shown in fig. 1. In a case, given the three test points A, B and C. Here since the point A is far from the line of bifurcation, it can easily be classified under "**0 class".** But for point C, we have very little margin of difference from the line of bifurcation.

The pattern close to line called support vectors and the distance between support vectors and the middle line called margin [9]. SVM performs the task, best to identify the margin distance to optimize the hyperplane.

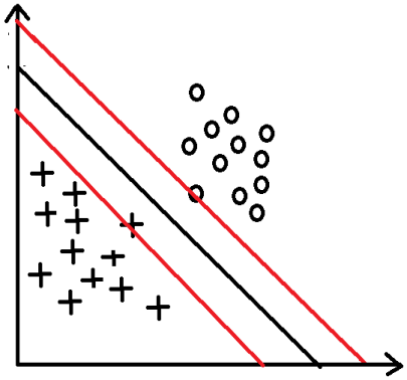

Figure 2: Optimal Hyperplane using the SVM algorithm

III. IMPLEMENTATION

The entire dataset used for training purpose comprises of the data from various news agencies like BBC, CNN, Sky News, etc. The document articles were collected from and categories into 8 distinct classes as mentioned in table 1. The total number of documents considered in the training set are more than 3500 distributed over all the 8 categories. Each document comprises of the article of more than 1000 words, overall averaging around 1800 words.

Table1: Document Categories and Training Dataset

Document Class (Category)	Number of Documents	Category Code
Atheism	**603**	0
Business	**511**	1
Car	**334**	2
Entertainment	**386**	3
Politics	**417**	4
Space	**436**	**5**
Sport	**511**	6
Technology	**401**	7

NLTK [12] is a platform for building Python applications to process human language data. It has many classes (programming class) each one has a lot of functions for process human language data and more than 50 lexical resources like "WordNet". In addition to many text processing libraries for tokenization, classification, stemming, parsing, tagging and semantic reasoning. The advent of NLTK has given a big chance to researchers to work in the field of text and sentences.

Feature engineering is considered as the main part of any recognition system. It specifies the success or failure of that system. Each phase is important to understand the nature of data and feature that distinguishes it. It collects only those data that are important to make bag of pattern, each pattern different from others according to feature space. This phase is carried out using NLTK library to deal with the text data.

Each article was passed through various phases like pre-processing, feature extraction, feature selection before they are stored in the database as shown in fig. 3. In the pre-processing phase, the stop words, punctuation marks, whitespaces and such non required entities were removed. Later on the article was converted to lowercase and the lemmatization of the words were done to stabilize all the words to its root form. These lemmatized sentences were then further tokenized to generate the tokens.

The feature Extraction [10] and feature selection phase was done using the Term Frequency (TF) and the Inverse Document Frequency (IDF). TF-IDF increases when the term exists in small number of documents that distinguish a particular document [11]. At same time it increases when the term appears many times in document. The feature selection deals with the same fashion as if we want to classify between two types of fishes like big fish and the small fish. Then the feature that would be useful is the length of the fish.

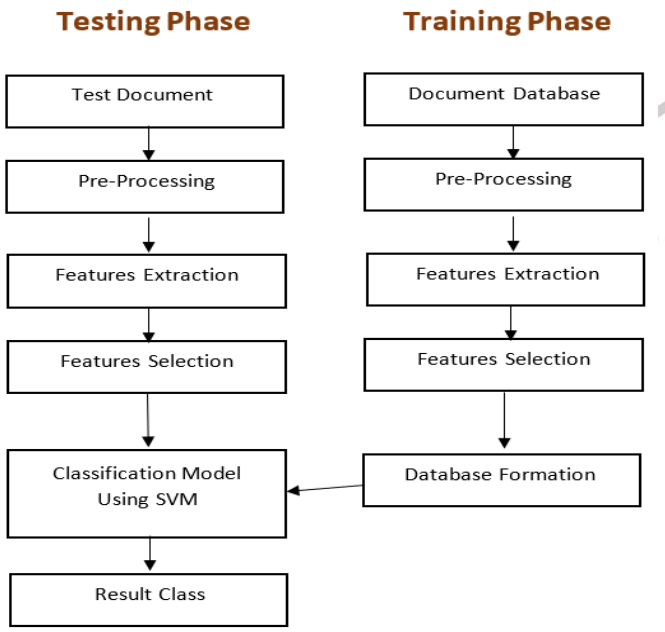

Figure 3: Document Classification Algorithm Flowchart

The same initial phase of pre-processing, feature extraction and feature selection were carried out for the testing phase. After the feature selection, the SVM model classification was applied with the trained database formed during the training phase, which results in the output class corresponding to the mentioned category.

IV. EXPERIMENTAL RESULTS

The database of the keywords corresponding to each class will be formed after the suitable feature selection. Using the trained model, accuracy of around 96.6% was observed over the training data set. Around 130 cases that were mis-classified was because of the intersection of the same word feature between the two or multiple classes. Just in a case, if in an article, is the mention about the business development due to the technological advancements, then the feature words that have been extracted from the document possess more frequency towards the technology class rather than the business class. Such intersecting document articles leads to the misclassification of the documents.

Table 2: Result of SVM Model for training phase

Accuracy score : 0.96604				
Report:				
	Precision	**Recall**	**F1-score**	**Support Number**
0	0.00	0.00	0.00	0
1	1.00	0.96	0.98	3598
Micro avg	0.96	0.96	0.96	3598
Macro avg	0.50	0.48	0.49	3598
Weighted avg	1.00	0.96	0.98	3598

Table 2 shows the statistical details of the scores of the model during the training phase. Recall is the metric that shows the ability of the model to find within a dataset every relevant case. The value of recall should be as maximum as possible. Precision is the parameter that is defined as the total count of true positives with respect to the total of true positives and false positives. Mixing Precision and Recall is called F1-score. F1-score is the combining average of recall and precision considering the two metrics into account.

Figure 4 shows the SVM classifier results for some 200 training patterns with training dataset number as X-axis and the accuracy mentioned along Y-axis for corresponding document.

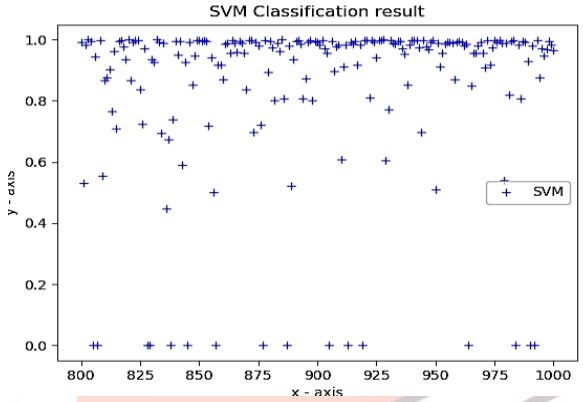

Figure 4: Graph of accuracy VS training dataset using SVM classifier

After the successful training phase, 20 unseen documents for each class were fed into the testing phase followed by the preprocessing, feature extraction and feature selection processes as shown in fig. 3. The testing accuracy obtained during the classification process was about 89.70%. Table 3 shows the statistical details of the scores of the model during the testing phase.

Table 3: Result of SVM Model for testing phase

Accuracy score : 0.89705				
Report:				
	Precision	**Recall**	**F1-score**	**Support Number**
0	0.00	0.00	0.00	0
1	1.00	0.89	0.94	160
Micro avg	0.89	0.89	0.89	160
Macro avg	0.50	0.44	0.47	160
Weighted avg	1.00	0.89	0.94	160

The confusion matrix for the testing dataset is as shown in fig. 5. This matrix provides good view for success and failure cases with accuracy 89.70%.

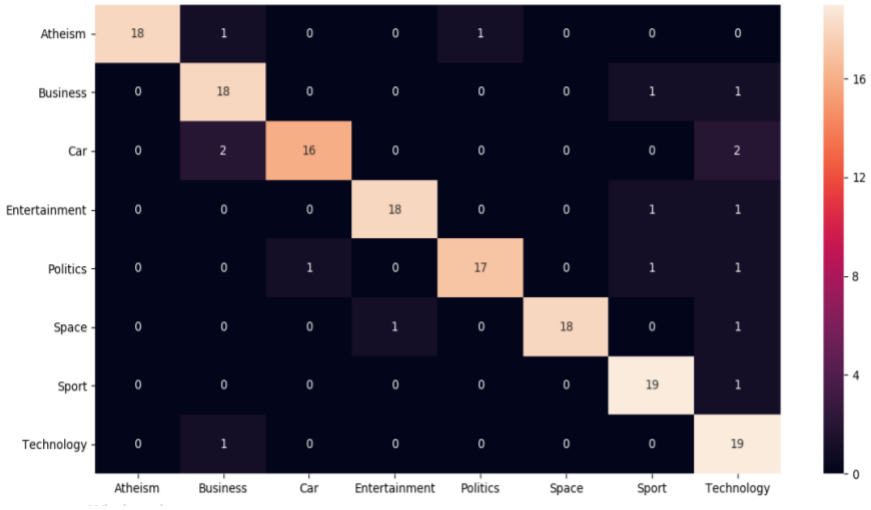

Figure 5: Confusion Matrix for testing patterns

V. Summary and Future Work

This proposed model provides a healthy accuracy of around 89.70% during the testing phase, which can be further improved by using the dependency parsing techniques. These techniques provides the better understanding of the dependent entity and thereby prevents from the confusion when there is the intersection of the features between the two or multiple classes. Also, an application or webpage can be developed as further extension to this concept, which can also provide the graphical view of the output class with their respective probabilities. Addition of the other categories for classification can also be done.

Also, this research can be extended for other languages to make document classifier more versatile. Some of research classify the documents based on the images within it. Furthermore, this work can also be integrated with image classifier to improve accuracy of classification.

REFERENCES

[1] Valencia-García, R. and García-Sánchez, F. (2013). NATURAL LANGUAGE PROCESSING and Human–Computer Interaction. Computer Standards & Interfaces, 35(5), pp.415-416

[2] Aggarwal, C. and Zhai, C. (2013). Mining text data. New York: Springer

[3] Mccallum, Andrew & Nigam, Kamal. (2001). A Comparison of Event Models for Naive Bayes Text Classification. Work Learn Text Categ. 752

[4] Deep ,Bassam "Text Classification Using Object Oriented Attributes" Syria 2016

[5] Dimitris Liparas1, Yaakov Hacohen-Kerner2, Anastasia Moumtzidou1, Stefanos Vrochidis1, Ioannis Kompatsiaris1 ," News Articles Classification Using Random Forests And Weighted Multimodal Features" Thessaloniki, Greece 2014

[6] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao , "Recurrent Convolutional Neural Networks For Text Classification" China 2015

[7] Thorsten Joachims, "Text Categorization With Support Vector Machines: Learning With Many Relevant Features" Germany 2003

[8] Saleh Alsaleem "Automated Arabic Text Categorization Using SVM and NB", Shaqra University, Saudi Arabia

[9] Nlp.stanford.edu. (2020). [online] Available at: https://nlp.stanford.edu/IR-book/pdf/15svm.pdf [Accessed 26 Jan. 2020].

[10] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, And Lofti Zadeh "Feature Extraction" Usa 2006

[11] Shahzad Qaiser , Ramsha Ali , "Text Mining: Use Of Tf-Idf To Examine The Relevance Of Words To Documents" , Malaysia ,2018

[12] Bird, S., Klein, E. and Loper, E. (2020). 1. Language Processing and Python. [online] Nltk.org. Available at: https://www.nltk.org/book/ch01.html [Accessed 26 Jan. 2020].