Classification of Multiple Labeled Text Article with Reuter Dataset using Support Vector Machine

Jaivik Vaghani
Master of Engineering(Computer Engineering))
University of Guelph
Guelph,Canada
jvaghani@uoguelph.ca

***Abstract*—Text categorization is a typical job in text mining difficulties. In real-world document categorization contexts, uneven text data is not commonplace. The goal of this research is to look at the performance of SVM classifiers on data of textual news. For training & Ttesting the model, we create a collection of features using the C-V and T F-I D F techniques. Here, notably Reuters, we get the performance of the Support Vector Machine method utilizing both kernels(Linear and polynomial). The findings show that the SVM with a linear kernel performs better, attaining an accuracy of 86.10 percent on the benchmark dataset, while the SVM with a polynomial kernel obtains an accuracy of 78.28 percent.**

***Index Terms*—Reuters(data set) , Classification of Text, SVM, CV, TF - IDF Vectorizer.**

I. Introduction

The Internet contains an enormous volume of text-based content, which makes it very challenging to manually categorise. This makes it difficult to categorise various forms of content, such as separating safe from malicious emails, classifying patient reports according to medical records, classifying academic papers according to technical specifications, and categorising news stories. As a result, an automated method is required to make the classification of textual content simpler and more efficient.

The way of classifying and labeling language-based object into specific semantic categories or labels is known as text categorization, also known as text classification. This process entails identifying characteristics, establishing papers, putting algorithms into practise, and ensuring quality. With the prevalence of online activities and the availability of huge number of texts doc, it is important for getting pertinent details through categorization. An increasingly important practise is the management and organisation of textual information. The frequency of each phrase was employed as a feature to train the model in this instance of the intelligent document classification process, one of many ways used. Popular supervised learning methods for text categorization include the SVM.

The Support - Vector Machine technique aids in creating the optimal decision boundary between vectors that are a part of a group and those that are not. It is a good option for the task at hand—identifying multi-category news—because of its versatility, which enables it to be used to any vector. Due to technical advancements in the world of online news articles, where an abundance of material is frequently faced, people are more focused on their busy life. People consequently decide to read news stories that are pertinent to their specific interests. Finding news that is pertinent to one's interests can be challenging, as many valuable items may be irrelevant. This experiment is organised as : II(part) gives back ground data about area I'm researching. I discuss our suggested system flow and processing methodologies of algorithm in III(Section). The experimental findings and several performance assessment methods are presented in Section IV. Finally, I had added more debates and highlighted potential future research topics in Section V, where I had drawn a conclusion to the study.

II. Work Related to Paper

A. A Review of Methods and Techniques

It has been extensively investigated how to automatically categorise text content in many languages. According to the literature, researchers have concentrated on text classification for languages like English and Arabic.

The article [3] reviews numerous conventional approaches to categorising online articles of news & suggests a automatic framework for categorisation. Several classifiers were examined, and a Bayesian classifier reached 86 percent accuracy with confusion measurements. Another paper [4] presents a feature extraction method that effectively captures ambiguity by combining training materials and suggested labelling sets. Utilising probabilistic and membership-based metrics, uncertainty was assessed using engineering feature and segmentation of fuzzy C-means . The effectiveness of C fuzzy mean Segement in this multi label was evaluated by non-parametric assumption tests. The detected attributes were used to offer a technique for evaluating the network in multi-label categorization. Comprehensive experimental results show the effectiveness and important advantages of this idea over conventional methods.

Another study used two distinct news categorization datasets from NLPCC2014 and Reuters to test a system using attention-based recurrent neural networks. They outperformed all conventional methods, with F values of 76.5 percent and 41.7 %, respectively. To address problems with data analytics and

text processing, a different group of writers suggested a novel TF- IDF method coupled of the Louvain approachs. In terms of execution speed and accuracy , their technique performed better than cutting-edge methods. Last but not least, the authors of a publication discovered that a straightforward BiL-STM design with sufficient regularisation produced F1 score and good accuracy s on 4 benchmark datasets either matching or else outperforming the most recent world.

III. METHODOLOGY

SVM classifiers were used in this study to categorise text articles. The datasets from Reuters-21578 were used by the classification model1. Fig 1 shows the process diagram of system for this classifier, and the subsequent phases is outlined here.

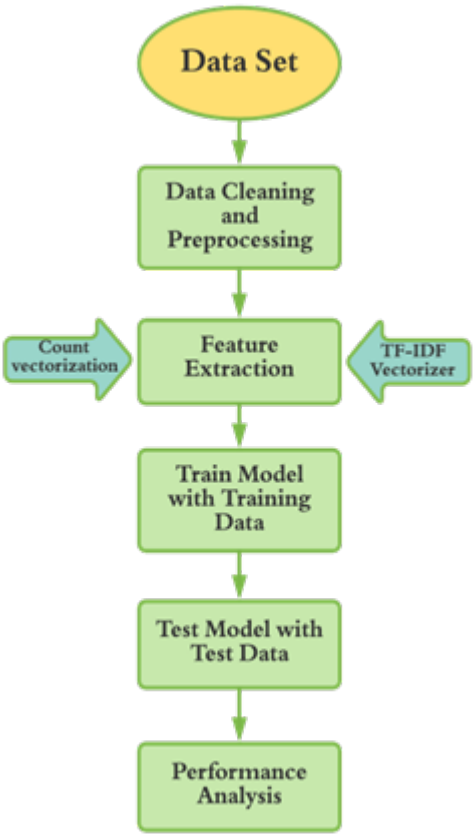

Fig. 1. System Flow Diagram.

A. Data Describtion

The Reuters-21578 dataset is largely regarded as a solid reference for document categorization tasks. It is a large dataset with a variety of classifications and labels. With a total of 90 classes and 10788 papers, we chose the top ten for our expermintel report. Tab 1 lists first 10 class and matching count of dataset in taken document.

TABLE I
TOTAL DOCUMENT COUNT (TEN)

Count	*Category*	*Subhead Total Doc Number*
1	earnn	3926
2	mzq	2369
3	crudes	55
4	taste	453
5	money	362
6	trades	329
7	graisn	295
8	corns	295
9	dlir	169
10	supplys	154

B. Data Preprocessing & Cleaning

Prior to feeding raw text data into the classifier, numerous data cleaning and preparation operations must be completed. These jobs are critical because they assist in the removal of undesired symbols, URLs, and other useless information from raw text documents. Figure 2 in this paper explains the stages required in data cleaning and preparation.

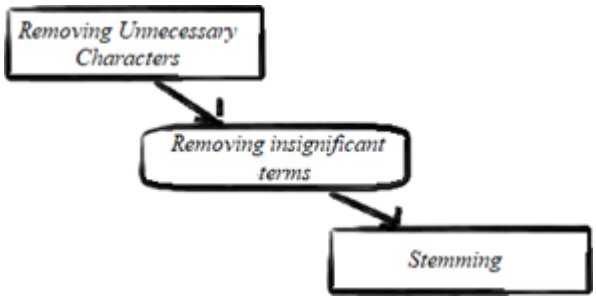

Fig. 2. Data Preprocessing & Cleaning Steps.

To tidy up our data, we transformed all uppercase letters to lowercase and deleted Unicode characters and URLs. In addition, we removed special characters such as ‘”, @,!, and so on from the raw text. These components have the potential to reduce the model’s performance. To clean up the text even further, we deleted English stop-words and used stemming on our dataset. After eliminating puntuation marks, numeric numbers, specialze symbols, hold -words, and stemming, the dataset was ready for classifier methods. Table II provides detailed information about the removed letters or words during preprocessing.

TABLE II
WORDS WHICH SHOULD BE REMOVED IN PREPROCESSING

Categories	*words & Charcter*
Special Characters	’!’, ’’, ’$’, ’#’, ’:’, ’,’,
Numbers	0,1, 2, 3, 4,5,6,7,8,9
Feeling	:)(, :S,(:(, :o,etc ..
hold-word	a,to, an, if, or,else, etc...

C. Feature Extracting

The selection of the proper features has a significant influence by the performance of Classification algorithm in

machine learning. To extract characteristics from text data, the C-V and TF-ID Frequency Vecto rizer are often employed. The Count Vectorizer generates a vector with dimensions according to the individual words in the corpus. Each word is represented by a dimension, with a value of 1 in that dimension and 0 in others, indicating its frequency. On the other hand, the TF-IDF vectorizer assigns numerical values to each word depending on its occurrence and frequency. Words are maped using their frequency multiplication is done by the inverse doc frequency. In basic term, regularly occurring texts like articles, prepositions, and conjunctions are given less weight, while less frequent ones are given greater weight.This weighting improves classifier performance. This method is used to identify the meaning of a keyword included inside a text. If the keyword is 't' and the Doc is 'd,' the complete document is represented as 'D', after that the equation is,

$$tf - idf(t, d, D) = tf(t, d) * idf(t, D) \quad (1)$$

Here,

$tf(t, d)$ is the frequency of 't' in 'd'
$idf(t, D)$ is how 't' is common or rare across 'D'

Fig. 3. Equations.

For this report, we have utilized the top 1500 features..

D. Classifie Used

For taken dataset was separated into two sets to feed the classification algorithms: the testing set and the train set. The training set comprises 60% of the dataset, with the remaining 40% in the test set. Among several classifiers, we employed a commonly used SVMC(Support Vector Machine Classifier) in this paper. The svm should be used to solve regression or classification issues. It processes the input using the kernel approach and computes a suitable boundary between potential outputs based on these modifications. It is especially useful for text datasets since it can deal with overfitting difficulties in high-dimensional domains.Because most text datasets can be separated linearly, the linear kernel is preferred for text classification. We examined the outcomes of Support Vector Machine Classifiers with both kernels (linear and polynomial)in this study.

IV. RESULTS

This section presents an overview of the performance of Support Vector Machine Classifiers on the Reuters-21578 dataset. Sklearn's train test split module was used to randomly partition the dataset into two pieces. The first half of the data, which accounts for 60% of the total, was used to train the classifiers, and the second half, which accounts for 40% of the total, was used to test their efficacy. We compared the results of two distinct kernels - using SVM Classifiers in this study. Table (III) shows the overall efficiency of SVM utilising linear and polynomial kernels.

This section describes how Support Vector Machine Classifiers fared on the Reuters-21578 dataset. Sklearn's train test split module was used to divide the dataset into two pieces at random. The first 50 percent of the data was used to train the classifiers, while the other 50 percentf was utilised to test their effectiveness. In this paper, we compared the results of two alternative kernels, with Svm Classifiers. Table3 shows the total effeciency of SVM when both of these kernels are used.

TABLE III
ACCURACY FOR SUPPORT VECTOR MACHINE MODEL

Model	***Linear k***	***Polynomial k***
SVM model	O.86	O.78

TABLE IV
RESULTS OF LINEAR KERNEL USING SVM CLASSIFIER

Classes	**Precision**	**Re call**	**F 1 - Score**	**Support Score**
earnn	0.86	0.86	0.86	1712
mzq	0.76	0.75	0.56	1521
crudse	0.56	0.82	0.85	1332
taste	0.86	0.54	0.64	1287
money	0.66	0.67	0.72	1098
trades	0.72	0.79	0.85	897
graisn	0.63	0.72	0.77	756
corns	0.73	0.69	0.63	421
dlir	0.86	0.81	0.56	251
supplys	0.85	0.80	0.62	138

TABLE V
RESULTS OF POLYNOMIAL KERNEL USING SVM CLASSIFIER

Classes	**Precision**	**Re call**	**F1 - Score**	**Support Score**
earnn	0.78	0.77	0.75	1712
mzq	0.76	0.78	0.56	1521
crudse	0.56	0.77	0.65	1332
taste	0.78	0.45	0.74	1287
money	0.69	0.76	0.53	1098
trades	0.52	0.78	0.57	897
graisn	0.59	0.72	0.86	756
corn	0.66	0.78	0.68	421
dlir	0.68	0.81	0.56	251
supplys	0.84	0.79	0.65	138

Table VI compares the findings of the old approach to the outcomes of the suggested method.

TABLE VI
COMPARING THIS REPORT WITH PREVIOUS SIMILAR RESEARCH

System	***Algorithm Model Used***	***Obtained Accuracy***
This report	SVM with line ar Kernel	0.86
Chang shun	ATTENT ION LS TM BI - R NN	0.466
Pooj a et al. [17]	LS-T WSVM	0.981
Chang shun	TF - IDF and SVM	0.323
Jing zhou	Bow - CNN	0.81
Johann es	CMLPC	0.90

ACKNOWLEDGMENT

According to the paper, the SVM classifier is effective in addressing the News Categorization challenge. Specifically,

when using the svm model classifier with a best(linear), an amazing accuracy of 86.10% is obtained. Further research could look into potential enhancements to SVM-based multi-category techniques. Additionally, new SVM-based classification methods could be developed and evaluated in related application domains. This will reinforce the dominance of SVM-based algorithms in text categorization challenges.

REFERENCES

[1] [1] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman, "Comparing dimension reduction techniques for arabic text classification using bpnn algorithm," in 2010 First International Conference on Integrated Intelligent Computing, 2010, pp. 6–11.

[2] [2] B. Pendharkar, P. Ambekar, P. Godbole, S. Joshi, and S. Abhyankar, "Topic categorization of rss news feeds," Group, vol. 4, p. 1, 2007.

[3] [3] J. Ahmed and M. Ahmed, "Online news classification using machine learning techniques," IIUM Engineering Journal, vol. 22, no. 2, p. 210225, Jul. 2021. [Online]. Available: https://journals.iium.edu.my/ejournal/index.php/iiumej/article/view/1662

[4] [4] J. R. Quinlan, C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[5] [6] C. Du and L. Huang, "Text classification research with attention-based recurrent neural networks," International Journal of Computers Communications& Control, vol. 13, no. 1, pp. 50–61, 2018.

[6] [7] C. Iwendi, S. Ponnan, R. Munirathinam, K. Srinivasan, and C.-Y. Chang, "An efficient and unique tf/idf algorithmic model-based data analysis for handling applications with big data streaming," Electronics, vol. 8, no. 11, p. 1331, 2019.

[7] [10] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, p. 150, 2019.

[8] [11] P. Saigal and V. Khanna, "Multi-category news classification using support vector machine based classifiers," SN Applied Sciences, vol. 2, no. 3, pp. 1–12, 2020.