TEXT CLASSIFICATION ON NEWS ARTICLES

NG KANG WEI

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

TEXT CLASSIFICATION ON NEWS ARTICLES

NG KANG WEI

RESEARCH PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY UNIVERSITY OF MALAYA, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF DATA SCIENCE

2019

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: (I.C./Passport No.:)

Registration/Matric No.:

Name of Degree:

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:
Designation:

TEXT CLASSIFICATION ON NEWS ARTICLES

ABSTRACT

An abstract must not exceed 500 words, typed in a single paragraph with double- spacing, and written in Bahasa Malaysia and English language. A maximum of five (5) keywords should also be listed below the abstract.

Keywords: Keyword, keyword, keyword, keyword.

ABSTRAK

Ini merupakan abstrak dalam Bahasa Melayu (satu perenggan).

ACKNOWLEDGEMENTS

Thanks guys. I owe you many.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Introduction

As optical character recognition technology advances, large number of physical documents are made electronically available and many more articles are created and available online. The information contain within these documents would need to be extracted, analyzed, stored and made searchable so that others can make use of it. Natural language processing (NLP) methods are needed to analyze the content or the sentiment in the text.

Document classification is one of the NLP method that categorize the text into different topics or categories. This classification would be helpful to future researchers who want to find some topic from the text. The researchers could just focus on the category they are interested in rather than skimming through all the documents to obtain the intended information.

The technology breakthrough in recent years, machine learning algorithms, processing power of processors have been a boost in NLP field. With the breakthroughs, there has been an advancement of the methods used in NLP with artificial intelligence without the need of intervention of domain experts.

There are several approaches to the document classification problem, multilabel where each document can belong to several categories or classification, where each document can only belong to one category. Within machine learning methods, there is clustering which is an unsupervised learning method or the supervised learning approach. This shall focus in the the direction of classification rather than multilabel and clustering.

In document classification most of the algorithms used vector space model to represent the unstrucutured textual data. (Ababneh, Almanmomani, Hadi, El-Omari, & Alibrahim, 2014). This vector space model represent the sequence of the textual features and their

weight, it is easy to implement and provide uniform representation for documents. However, it has a drawback, it represent all the words in the documents, the dimension of the vector would be huge. This huge vector space model would impact the performance of the machine learning tasks. (Moldagulova & Sulaiman, 2018). Therefore, this study would focus on the dimension reduction on vector space model on document classification.

1.1.1 Problem Statement

Term vectors is one of the most commonly used document representation algorithms, but dimension of the feature space can too large and the vectors can be too sparse. (Moldagulova & Sulaiman, 2018)

1.1.2 Research Objectives

1. To identify a document representation algorithm that is optimized to extract the features from news articles
2. To investigate the machine learning (ML) algorithm used in document classification and apply the most suitable algorithm.
3. To evaluate the performance of the document represenation and document classification algorithm.

1.1.3 Research Questions

1. Which dimension reduction algorithm is optimized for news article?
2. How complexities of the features influence the accuracy?
3. Which is the best machine learning algorithms in document classification?

1.1.4 Research Motivation

In the age of big data, the amount of data generated and collected is growing at an explosive rate. Much of the data generated and collected is in the form of unstructured

text. The value contained within the text cannot be extracted and be useful to us without natural language processing (NLP). Document classification is one of the pillars in natural language processing.

In document classification, the unstructured text would be given a label or multiple label dependent on the method used. These labels would make the data more meaningful and searchable. Users can search for a topic just by selecting text with the particular label rather than performing a manual word search on all the text document.

Bag of words is the most commonly used document representation method in document classification. Bag of words would produce a vector space model of the textual data representation. The dimension of this vector space model would be big because of the amount of words in the documents. This huge dimension of vector space model would decrease the performance of the document classification algorithms.

With dimension reduction algorithms, the dimension of the vector space would be decreased and the performance of the machine learning algorithms would be increased. However, the effect of the different dimension reduction algorithms might have a different effect on the performance of the machine learning algorithms.

This study would investigate the effect of the dimension reduction algorithms on the performance of the machine learning algorithms.

1.1.5 Research Significance

This ressearch would classify news articles into different categories. In order to do that, first the features have to be extracted from the documents. The features might need to be compressed. Then the features are used to train machine learning models. After that, the trained machine learning models are validated and tested to evaluate its performance.

The document representation method is the most commonly used bag of words approach, Term Frequency - Inverse Document Frequency (TF-IDF). In this approach, the documents

are converted into vector space models. Due to the large amount of words in the documents, the vector space would be large and sparse, this is known as the curse of dimensionality problem. This large and sparse vector space would be an obstacle to document classification, machine learning models accuracy would be impacted due to the large vector space.

By applying dimension reduction algorithms to the vector space model, the vector space and sparsity of the vector could be reduced. With the reduced vector, the accuracy of the machine learning models would be increased. However, which of the dimension reduction algorithms would perform best for document classification on news articles? This study would try to answer this question by studying which of the dimension reduction algorithm is most suitabled and applying it on a dataset.

1.1.6 Expected Outcome

1. A prototype document classification application with 80
2. A dimension reduction algorithm is applied on the extracted features of the documents
3. A best suited machine learning algorithm is applied on the document classification application
4. Evaluate the accuracy of the document classification application

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In text classification proposed in this study, there are 3 stages namely document representation, dimension reduction and classification. For document representation the chosen method is Term Frequency-Inverse Document Frequency (TF-IDF) which is based on the bag of words method. TF-IDF provide a measure of weight or importance to the words. The value of TF-IDF estimate the amount of information provided by each word in its document. (Arroyo-Fernández, Méndez-Cruz, Sierra, Torres-Moreno, & Sidorov, 2019) The value of TF-IDF increases proportionally to the number of times a word appears in a document but is offset by the frequency of the word in the corpus. This means that if a word appears for many times in a document but it also appears for many times in many other documents in the corpus, then it is not an important word. TF-IDFThe formula for TF-IDF is shown below:

$$TFIDF = tf \times \log_e \frac{N}{df} \tag{2.1}$$

where:

tf = the number of times a word appears in a document

N = the total number of documents in the corpus

df = the number of documents that contain the word

The remaining parts are dimension reduction and classification, the methods in these 2 parts would be evaluated and chosen.

2.2 Dimension Reduction

2.2.1 Single Value Decomposition (SVD)

Single value decomposition is one of the most commonly used dimension reduction algorithms. It generalizes a complex matrix with many dimensions into a matrix of lower dimension via an extension of the polar decomposition. SVD detects the part of the data that contains the maximum variance in a set of orthogonal basis vectors. The data with the maximum variance would be the most prominent features of the data. (Sweeney et al., 2014)

Latent Semantic Analysis (LSA) a technique applied in natural language processing that apply SVD in its process. SVD is applied in LSA to transform the features by dropping the least significant values in the matrix thus reducing the dimensions of the matrix. (Karami, 2017)

Even though SVD has been applied in LSA and in other fields, it has been found that SVD is not as efficient as principal component analysis (PCA) and has a few drawbacks. SVD can extract the prominent features of the matrix but it does not help in reducing the sparsity of the matrix. In some of text clustering algorithms that used SVD for dimension reduction only find SVD useful when the features are redundant. (Narhari & Shedge, 2017)

2.2.2 Nonnegative Matrix Factorization (NMF)

Nonnegative matrix factorization (NMF) is a multiplicative updating solution to decompose a nonnegative temporal-frequency data matrix into the sum of individual components. The sum of individual components are calculated from a nonnegative basis matrix. (Chien, 2019)

NMF and its variant have found to be applied in many fields such as feature extractions, segmentation and clustering, dimension reduction and others. However, the nonnegativity constraints of NMF proved to be problematic when the data matrix is not strictly non-

negative. The semi-NMF relaxes the non-negativity constraint of NMF so that the resulting matrix has mixed signs. (Allab, Labiod, & Nadif, 2017)

Researchers who performed experiments to compare the performance of SVD, NMF and PCA have found that NMF do not perform as well as PCA. SVD and NMF achieved a similar accuracy in the experiment. This might due to both SVD and NMF are dependent on matrix decomposition technique while PCA is dependent on eigenvalue decomposition. (Mohamed, 2019)

NMF might be more suitabled for clustering as it took the least amount of time in clustering the data compared with SVD and PCA. (Mohamed, 2019). This advantage of NMF over SVD and PCA is negligible in this study since this study would focus on classification rather than clustering.

2.2.3 Principal Component Analysis (PCA)

Principal component analysis is one the state of the art dimension reduction algorithm. The main purpose of PCA is to project data samples from high dimensional space into low dimensional space by linear transformation while preserving the original data features as much as possible. (Ma & Yuan, 2019) In other words, PCA is used to emphasize the variations in the data, bring out the important features in the data.

PCA is no stranger to the text classification field. It has been applied to different languages of text classification other than English. Label Induction Grouping (LINGO) a technique used in categorizing Indian Marathi language text document. PCA is applied in LINGO performed better than SVD as PCA extract the features better and has less loss of information. (Narhari & Shedge, 2017)

In another research on text classification on Arabic text and English text, it is also found that PCA outperformed SVD and NMF. The researchers found that PCA yields better result in terms of accuracy and normalized mutual information. The advantage of PCA is

that after transforming the matrix, the important features vector are orthogonal to each other. PCA also has a whiting transform that reduce noises in the data which in turn boost the performance and the accuracy of the machine learning algorithm.(Mohamed, 2019)

A variant of PCA which is in the form of a tree structure has been applied on the dimension reduction on sentiment analysis. The technique is called tree-structured multi-linear principal component analysis (TMPCA). TMPCA can retain the sentence structure and word correlations. (Su, Huang, & Kuo, 2018). However, this is a novel technique and PCA has been proven to be effective enough to handle the amount of data in this study.

2.2.4 Summary

The 3 dimension reduction or matrix decomposition algorithms above are considered blind source separation (BSS) methods, unsupervised learning algorithms. Its performance might not be as good as a deep learning algorithm such as Word2vec but deep learning algorithm's performance is dependent on the the scale of the data. Deep learning algorithm can only perform well when there is a lot of data. In our study, the data might not be sufficient to use a deep learning algorithm, thus the above methods are reviewed. Out of the 3 dimension reduction algorithms reviewed above, PCA seems to be the most promising in the field of text classification. Therefore, PCA would be chosen as the dimension reduction algorithm in this study.

2.3 Document Classification

2.3.1 Support Vector Machine

Support vector machine is a machine learning algorithm that construct a hyper plane to separate the examples into different classes. It has been proven to be very effective in dealing with high dimensional data. (Shinde, Joeg, & Vanjale, 2017). It is also proven to produce dramatically best results for topic modelling in experiments with the Reuters dataset.

(Dumais, Platt, Heckerman, & Sahami, 1998). Various issues need to be considered when applying SVM in document classification, the processing of the data, which kernel to use, and the parameters of SVM. A variant of SVM, called one-class SVM which is trained only with positive information has been used in document classification. (L. M. Manevitz & Yousef, 2002). The authors experimented with different kernels of SVM (linear, sigmoid, polynomial, and radial basis) with different type of document representation method (binary representation, frequency representation, TF-IDF representation, and Hadamard representation). The best result (F1 score of 0.507) is achieved with binary representation, feature length 10 and with linear kernel function.

In another research, the researchers apply SVM in the classification on web document instead of news or ordinary text document. The document representation method used in this research is vector space model, just the nouns term in the web pages. The researchers experimented with different SVM kernels and varying the size of the training sets. Expectedly, the precision, recall and accuracy increased as the size of the training set increase. Linear kernel achieved the best result out of the various SVM kernels, a classification accuracy of 80% is achieved. (Shinde et al., 2017).

SVM is relatively new compare to others algorithm in the field of document representation. It is not very efficient with large number of observations and it can be tricky to find an appropriate kernel for the problem.

2.3.2 k-Nearest Neighbours (kNN)

kNN is a classification machine learning algorithm that classify objects based on the closest training examples in the feature space based on a similarity measure. It is a simple and effective classification, as it only need 3 prerequisites. The 3 prerequisites are training dataset, similarity measure and the value of k which is the number of closest neighbours to be considered.

kNN needs minimal training, it only needs to plot the training examples into a feature space. kNN has been applied in document classification before, it is found that kNN take significant longer time to classify a document into a topic. This is because kNN uses all the features of the data to compute the distance. Since the authors are using term vector space document representation method, the dimension of the feature space is high, thus the more time is needed for kNN to compute all the distance between the test object with the training objects. Other than the time taken to compute the distance, the k value is another obstacle in kNN algorithm. In a high dimensionality feature space and the points are not evenly distributed, the k value is hard to be determined.

To overcome the problems mentioned above, the authors applied term vector space reduction method, divide the document feature matrix into parts. Term vector space reduction reduces the sparsity of the document term matrix by removing the features less appeared in the corpus. By reducing the term vector space, a slight deterioration in the classification accuracy but the time cost is dramatically reduced. kNN still achieved an accuracy of 92.7% but the time taken reduced from 53 minutes to 11 minutes. (Moldagulova & Sulaiman, 2018)

2.3.3 Neural Network

Neural network has a resurgence in recent years as there is a breakthrough in the neural network as Geofrey Hinton (et al.) discovered a technique called Contrastive Divergence that could quickly model inputs in a Restricted Boltzmann Machine (RBM). RBM is a 2-layer neural network that model the input by bouncing it through the network. This process is less computationally complex than backpropagation. (Hinton, 2002).

Currently, neural network is applied in deep learning to solve various problems, document representation is one of them. Ranjan (et al.) applied Lion Fuzzy Neural Network on document classification. The researchers used WordNet ontology to retrieve

the semantic of the words, and then added the context information onto it, thus the features obtained are semantic-context features. The classification part is performed by Lion Fuzzy Neural Network, which is a variant of Back Propagation Lion (BPLion) Neural Network that includes fuzzy bounding and Lion Algorithm. The neural network model used is trained incrementally. It achieves a higher accuracy than Naïve Bayes and some variant of the Lion Neural Network. (Ranjan & Prasad, 2018)

Other than the modified neural network shown above, a simple feed-forward neural network is also efficient in document classification. By using the Hadamard product as document representation method, a simple neural network also can achieve a good classification accuracy in document classification compare to Naïve Bayes, kNN, and SVM. (L. Manevitz & Yousef, 2007)

2.4 Conclusion

From the review of dimension reduction algorithms, PCA performed best compared to SVD and NMF. In addition, PCA performed well in text classification field.

In machine learning algorithms for text classification, all 3 machine learning algorithms reviewed above would be applied. One of the objectives of this study is to investigate the performance of different machine learning algorithms in text classification. The same dataset would be used to train 3 models and the performance would be evaluated.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This section would illustrate process flow to carry out the experiments to fulfill the aforementioned objectives of this research. The following flow charts would illustrates the steps taken in the several experiments.

3.2 Text preprocessing process flow

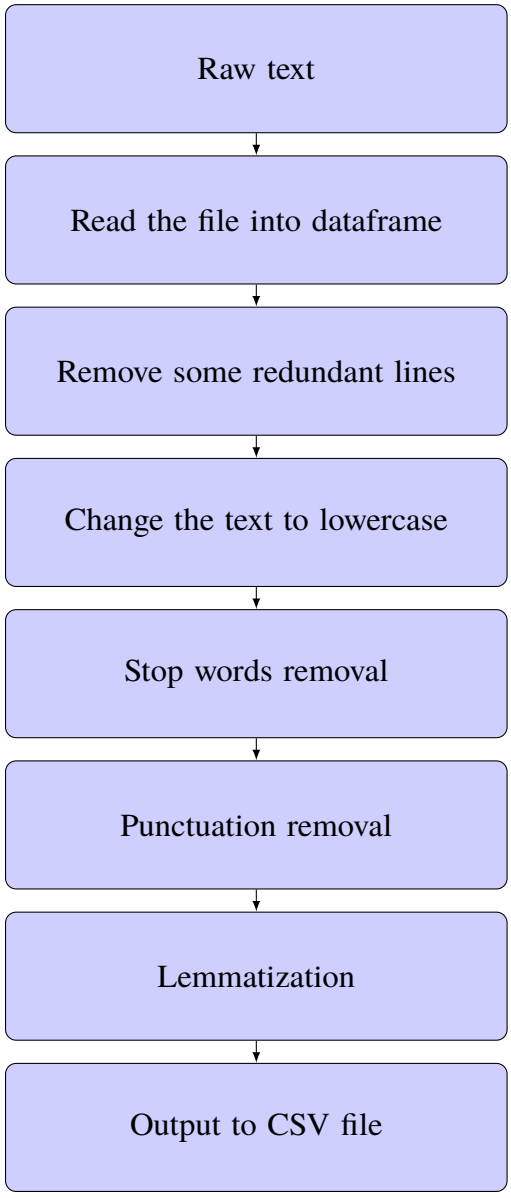

Figure 3.1: Overall process flow

This flow chart above illustrates the preprocessing process. The news articles articles

dataset would be in the form of raw text. There are some preprocessing to be done before features can be extracted from the text to build the text classification models. First, the raw text would be read into a dataframe or a data structure for ease of access and processing.

In the raw text files, there are some lines that are not relevant for text classification purpose, these lines would be removed to reduce the noise in the dataset. All the text in the dataset would be converted to lowercase for ease of processing. There are some words in the text that are useful in human speech but do not convey meanings in text analysis. These words are stop words, these words have to be removed as well. After stop words removal, the text should contain only the words that are vital to text analysis but there would be some punctuations and symbols in the text. These punctuations and symbols have to be removed as well to make the text cleaner and less noisy.

After that, there is an important step, lemmatization. Lemmatization would convert most of the words in the text to their root form which is known as a lemma. This would reduce the noise in the data by transforming similar words into a single word. The alternative to lemmatization would be stemming. Stemming would be faster than and need less processing power than lemmatization but stemming has a drawback against lemmatization. The result of stemming may not be a real word because stemming just chop off the ends of the words thus the resulting words may not be a real words. On the other hand, lemmatization uses vocabulary and morphological analysis of words to remove the inflectional endings only and resulting the root form of words. (reference needed)

3.3 Overall Process Flow

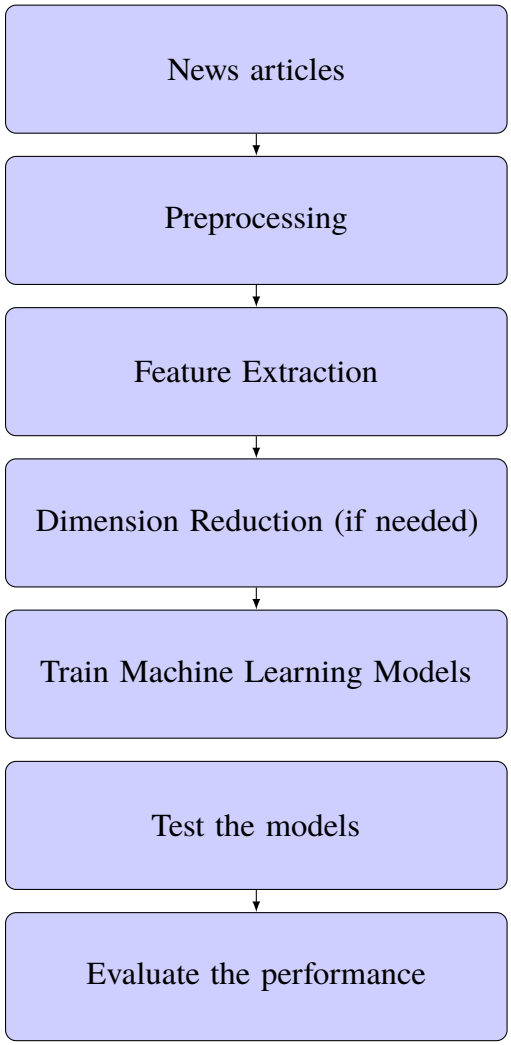

Figure 3.2: Process flow of preprocessing the text dataset

The process flow chart above is the overall process for the few experiments. As mentioned above, the news articles dataset has to be preprocessed before it can undergo feature extraction and train machine learning models.

After preprocessing, the resulting data would be used in several experiments with a few subtle differences. The differences in the experiments would be in the feature extraction stage and the dimension reduction stage.

3.4 The dataset

The dataset used in this research is the 20 news group dataset. It is freely available at `http://qwone.com/~jason/20Newsgroups/`. It is a collection of news articles that can be divided almost evenly to 20 groups. Some of the groups are closely related to one another and some are widely different. Categories with the same prefix for example comp, rec and talk would have high similarity with one another but also subtle differences. The categories with the different prefixes would be very different from one another.

This characteristic of the dataset make it a good dataset to test text classification. A good classification model should be able to differentiate the different groups of news article even the groups that are closely resembled one another.

The table below show the number of count for each category. There is a total of 18790 articles and most of the categories have almost 1000 articles in them.

Category	count
alt.atheism	798
comp.graphics	970
comp.os.ms-windows.misc	980
comp.sys.ibm.pc.hardware	978
comp.sys.mac.hardware	957
comp.windows.x	978
misc.forsale	962
rec.autos	988
rec.motorcycles	993
rec.sport.baseball	993
rec.sport.hockey	998
sci.crypt	991
sci.electronics	981
sci.med	988
sci.space	986
soc.religion.christian	997
talk.politics.guns	910
talk.politics.mideast	940
talk.politics.misc	774
talk.religion.misc	628

Table 3.1: The count of each category in the dataset

The bar chart below visualized the data displayed on the table for ease of viewing. Most of categories have almost 1000 articles except alt.atheism, talk.politics.misc and talk.religion.misc.

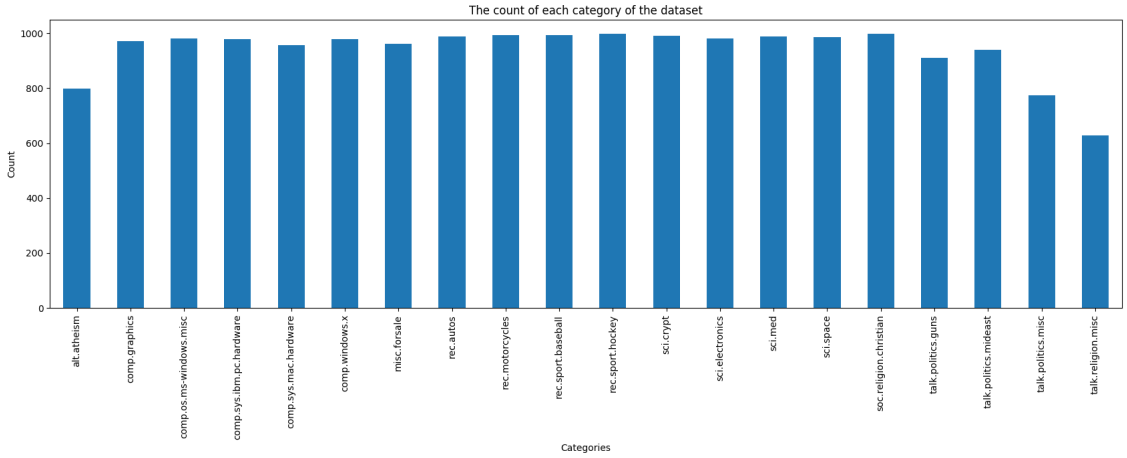

Figure 3.3: The count of each category in the dataset

As mentioned there are 18790 articles in the data, the articles would be split at random into 8:2 ratio. 80% of the data would be used as training data to train the classification model while 20% of the data would be use as the test data. The training data would consist of 15032 articles while the test data would consist of 3758 articles.

3.5 The experiments

The experiments that has been conducted in this research is as follows:

1. Term frequency
2. Term frequency with naive dimension reduction
3. Term frequency with SVD
4. TF-IDF
5. TF-IDF with naive dimension reduction
6. TF-IDF with SVD

Basically, there are 2 feature extraction method used in the experiments, namely term frequency and term frequency - inverse document frequency (TF-IDF). Each of the feature extraction method would be tested with and without dimension reduction algorithm.

There are 2 dimension algorithms involved, one is a naive method, which means that the features or columns lesser than a certain value would be removed. In other words, words or terms that do not appear much in the dataset would be removed. Another dimension reduction algorithm is single value decomposition (SVD). This method would retain the essence of the data, the part of the data with maximum variance.

After feature extraction and dimension reduction (if needed) are applied, the resulting features would be used to train machine learning models. There are 3 machine learning algorithms chosen in these experiments namely k-nearest neighbour (kNN), support vector machine (SVM), and neural network (NN). All 3 of the machine algorithms would be applied to all the different resulting features and the accuracy scores would be evaluated.

3.6 Conclusion

With the method mentioned above, the experiments would be carry out and the results would be recorded. The results would be compared and the differences would be analysed.

CHAPTER 4: RESULTS AND DISCUSSIONS

4.1 Introduction

The results of the experiments are shown in this section. The differences in the accuracy and performance of each of the methods would be discussed and analysed. Hopefully, the results and discussions would be able to answer the research questions posted in the beginning of this research.

4.2 Term frequency

ML	no of features	accuracy	time taken (s)
kNN	8000	0.31	3.74
SVM	8000	0.81	5.27
NN	8000	0.84	91.82

Table 4.1: Term frequency

Term frequency is one of most common feature extraction method in text. It convert all the words in the dataset into a matrix where each column is a word and the value in each of the columns are the number of times each word appear in the text. Each row in the matrix is a document. This vector space model representation of the words would result in a sparse matrix since each of the documents only contains a subset of all the words in the whole dataset.

In this experiment, the number of features of vector space model from term frequency extraction is limited to 8000, this is because of the memory constraint of the machine, if it is unlimited the resulting matrix would be of bigger size and would have the probability of running out of memory while processing the matrix.

In the results shown above, NN achieved an accuracy of 0.84 which is the best accuracy out of the 3 classification algorithm but it is also the one that takes the longest to process which is 91.82s. kNN took the least time to process, 3.74s but has the lowest accuracy

score, 0.31. Overall performance, SVM is the best, it only took 5s to process, which is 80s less than NN and has an accuracy of 0.81 which is comparable with NN.

reason of KNN accuracy low [reference] [https://arxiv.org/abs/1011.2807]

kNN accuracy is low in this scenario is possibly due to the sparsity of the feature matrix or vector space model. The data points are few and far in between. kNN is designed for low dimensional data therefore it doesn't perform well in large and sparse data. The underlying reason would be due to kNN classify a new data point based on the distance of the new data point with the nearby data points. Since the other data points are few and sparse, the classification of the new data points would be biased to the few data points that are near. This bias would reduce the accuracy of the classification.

NN took the longest time to process the features, this is due to the layers of neurons in NN. The vector space model even though sparse has high dimension, NN would need as much if not more neurons as the number of features to process the data. This processing would takes both time and computing power.

SVM depends on the prominent features of the dataset, the support vectors to classify the dataset so it is relatively faster and as accurate.

4.3 Term frequency with naive dimension reduction

ML	parameter value	no of features	accuracy	time taken (s)
kNN	100	2530	0.34	3.94
kNN	500	478	0.42	4.66
kNN	1000	173	0.41	5.17
SVM	10	12705	0.83	6.10
SVM	100	2530	0.76	5.67
SVM	500	478	0.63	20.20
NN	10	12705	0.87	116.78
NN	100	2530	0.80	48.08
NN	500	478	0.65	61.87

Table 4.2: Term frequency with naive dimension reduction

The feature extraction method used in this experiment is same as the experiment above but dimension reduction method is applied to the resulting matrix before training. The dimension reduction algorithm used in this experiment is a naive one which means that the columns with the least occurrence of word are removed assuming those words are unimportant and would not have much influence on the accuracy.

The parameter value is just an integer value passed to the dimension reduction function implemented. It has an inversely proportional relationship with the number of features. The larger the parameter value, the more columns would be removed and the number of features would decrease.

In this experiment, kNN still has the worst performance among the 3 which is at 0.34 with 2530 number of features. As the number of features decreases to 173, the accuracy increases slightly to 0.41. As the number of features decreases, the time taken for kNN to produce the result increases from 3.946s to 5.17s.

SVM and NN have the same trend in this experiment, the accuracy decreases as the number of features decreases. Accuracy of SVM decreases from 0.83 to 0.63 as the number of features decreases from 12705 to 478. The time taken by SVM fluctuates as the amount of features decreases, it decrease from 6s to 5s when the amount of features

decreases from 12705 to 2530, but the time taken increases from 5s to 20s as the number of features decreases further to 478. The decrement in accuracy in SVM is slight in relative to the decrement in the number of features. Accuracy decreased by 0.20 while the number features has decreased by 12227.

The time taken by SVM increases quite drastically when the number of features decreased. The sparsity of the vector space model decreases, resulting in a denser vector space model. SVM need to take into account more data points to calculate an effective hyperplane to separate the data points into different classes.

On the other hand, the accuracy of NN decreases from 0.87 to 0.65 as the number of features decreases from 12705 to 478. However, the time taken by NN fluctuates from 116.78s to 48.08s and then to 61.87s when the amount of features decreases from 12705 to 2530 then to 478. The overall decrement in time taken is possibly due to the reduction of features, less neurons are needed to process the data and thus the speedup.

In the scenarios where SVM and NN have more features than the first experiment with just term frequency, both of the classification algorithms have better accuracy. These 2 classification algorithms are efficient and able to process features in a large and sparse vector space.

4.4 Term frequency with SVD

ML	no of features	accuracy	time taken (s)
kNN	4000	0.31	482.66
kNN	2000	0.38	216.59
kNN	500	0.49	60.89
kNN	100	0.55	15.38
kNN	10	0.30	2.78
SVM	4000	0.80	370.43
SVM	2000	0.78	172.58
SVM	500	0.77	85.03
NN	4000	0.80	220.76
NN	2000	0.79	91.13
NN	500	0.77	40.86

Table 4.3: Term frequency with SVD

In this experiment, the feature extraction or document representation method used is still term frequency but the dimension reduction algorithm has been changed. Instead of reducing the dimension of the feature matrix naively by removing the terms that has the lowest frequency, SVD is used. SVD would reduce the dimension of the vector space model by retrieving the features with maximum variance in the data.

The number of features shown in the table above is the number of columns in the resulting matrix after applying SVD.

Similar with the trend in the previous experiment (term frequency with naive dimension reduction), the accuracy of kNN increases slightly, from 0.31 to 0.55 when the features decreases from 4000 to 100, but the resulting accuracy is still far from satisfactory. The time taken by kNN decreases from 482s to 15s as the number of features decreases. The trend of accuracy increment as the number of features decrease cease when the number of features is reduced to 10. kNN accuracy decrease from 0.55 to 0.30 at that point. The fluctuation of the accuracy would be due to kNN dependency on Euclidean distance between the data points to classify a new data. As the amount of features decrease, the feature matrix becomes less sparse, kNN would need to process less data points thus the

speedup. A less sparse matrix resulting from the decrement of features also make kNN classification more accurate and less biased but when the features decrease to an extent where there is no sufficient data to classify a data point correctly, the accuracy dropped.

The accuracy of SVM and NN also have the same trend with the previous experiment, the accuracy decreases slightly when the number of features decreases. Accuracy of SVM decreases from 0.80 to 0.78 then to 0.77 and the accuracy of NN decreases from 0.80 to 0.79 then further to 0.77 as the number of features decreases from 4000 to 2000 and 500.

As expected, the time taken by the classification model to reduce the dimension and predict the result decreases as the number of features decreases. However, when compare with previous experiments, the time taken in this experiment is astoundingly higher in the case where the features amount to 4000 and 2000. SVD would be the culprit, dimension reduction comes at a cost which is processing power. To calculate the maximum variance of the features and transform the feature matrix into a smaller dimension would require no small feats of calculation, this would consume both time and processing power.

However, the resulting accuracy is not as good as just with term frequency. This is expected because the number of features decreased, the classification would have lesser information to classify a new data point correctly. The reduction in features may save some memory space but in order to achieve that more processing power and time would be needed.

Comparing the performance of the classification models in this experiment with the second experiment, term frequency with naive dimension reduction, the performance of the classification models are almost similar when the amount of features are around 2000. If the scenarios where the amount of features is around 500 is taken into account, SVD has a better performance. The classification models with SVD achieved a higher accuracy than that of naive dimension reduction. At this stage, the advantage of SVD over naive

dimension reduction is shown. SVD transformed the feature matrix into smaller dimension but retaining the maximum variance or the most prominent feature of the data. Therefore, models trained with SVD should have a better performance than those that trained with naive dimension reduction.

The next experiments would apply different document representation algorithm to investigate further the effect of dimension reduction has on classification model performance and compare the performance of different document representation method.

4.5 TF-IDF

ML	no of features	accuracy	time taken (s)
kNN	8000	0.76	3.71
SVM	8000	0.87	2.39
NN	8000	0.88	55.82

Table 4.4: TF-IDF

The previous experiments found that SVD has a slight edge over naive dimension reduction when the number of features decreased to a certain extend. In this and the next experiments, the effect on dimension reduction is further explored. A different document representation algorithm is applied in this and the next 2 experiments.

The document representation algorithm applied is TF-IDF. In contrast with term frequency which only take the frequency of each word into account, TF-IDF takes both the frequency of each word and the rarity of it into account. If a term or word appear in high frequency but in many documents, this word may not be of importance and consequently is not a meaningful feature. If a word appear rarely and only in a few documents, this word would have high importance and would be meaningful feature of the few documents.

With TF-IDF, kNN can achieved a satisfactory accuracy score of 0.76 even though the number of features in the resulting matrix of TF-IDF is the same with term frequency which is 8000. The vector space model of TF-IDF would not be as sparse as that of term frequency which is kNN is more suitable for kNN.

NN is still provide the highest accuracy score of 0.88 but the time taken also the longest at 55.82s.

SVM achieved an accuracy of 0.87 which is just 0.01 shy of what achieved by NN and the time taken is the lowest among the 3 which is 2.39s.

In comparison with the first experiment that apply term frequency document representation without dimension reduction, the performance of the classification models significantly

improve. kNN accuracy has more than doubled from 0.31 to 0.76. SVM accuracy increases from 0.81 to 0.87 while accuracy of NN increases from 0.84 to 0.88. Besides accuracy, the time taken also improved, time taken by SVM reduces from 5.27s to 2.39s while time taken by NN reduces from 91.82s to 55.82s. All these improvements are achieved with the same number of features in the vector space, which is 8000.

The performance of the classification models in this experiment is also better than the those in the 2 experiments above with dimension reduction. However this may not be a fair comparison because the different number of features are used and dimension reductions are not applied in this experiment. Dimension reduction algorithms would be applied to the vector space model from TF-IDF in the following experiments in order to have a fair comparison.

The performance increment from term frequency to TF-IDF seems to prove that TF-IDF is a better document representation method.

4.6 TF-IDF with naive dimension reduction

ML	parameter value	no of features	accuracy	time taken (s)
kNN	50	4221	0.74	4.00
kNN	100	2530	0.71	4.11
kNN	500	478	0.49	5.15
SVM	50	4221	0.86	2.58
SVM	100	2503	0.83	2.63
SVM	500	478	0.66	2.94
NN	50	4221	0.85	38.80
NN	100	2530	0.83	34.28
NN	500	478	0.64	77.12

Table 4.5: TF-IDF with naive dimension reduction

Similar with the experiment with term frequency, naive dimension reduction is applied to the vector space model generated from TF-IDF. The trend over all the 3 machine learning models when the number of features decreases are similar. The accuracy of the classification models decreases and the time taken increases.

The accuracy achieved by kNN with TF-IDF plus naive dimension reduction is still passable at 0.74 when the features reduced from 8000 to 4221. kNN's accuracy dropped slightly to 0.71 when the number of features decreases to 2530. When the number of features decreases from 2530 to 478, as expectedly the accuracy of kNN dropped for quite a large margin, from 0.71 to 0.49. The time taken by kNN increases slightly from 4s to 5s as the features decreases.

SVM has the same behaviour with kNN in this experiment, its accuracy decreases from 0.86 to 0.83 and then to 0.66 when the number of features decreases from 4221 to 2503 to 478. The time taken increases slightly as well as the number of features decreases.

NN also has the similar trend in accuracy as the number of features decreases. NN's accuracy decreases from 0.85 to 0.64 when the number of features decreases from 4221 to 478. The differences in NN is that the time taken almost doubled when the number of features is low compare to when the number of features is high. The time taken is 38s

when there is 4221 features and when there is only 478 features, the time taken doubled to 77s.

In the scenarios where the number of features are almost halved, from 8000 to 4221, the performance of the classification models are still comparable with the performance achieved with just TF-IDF without any dimension reduction. The accuracy are almost the same, the time taken is similar except NN which has quite a speedup when the number of features is reduced to 4221.

It can be deduced that with naive dimension reduction, the number of features and memory needed to store the vector space model is reduced. With this reduced number of features, the classification models can still achieve comparable performance at a slightly decreased capacity.

This is mainly because of the reduction in features. As the amount of information became lesser, less information is available to train a comprehensive model. Therefore, the accuracy of the models decrease. Even though the dimension reduction applied is a naive one and less computing intensive, it still increases the time taken compared with just with TF-IDF. The more reduction is performed, the time taken would increase as well.

4.7 TF-IDF with SVD

ML	no of features	accuracy	time taken (s)
kNN	4000	0.77	457.16
kNN	2000	0.58	208.64
kNN	500	0.55	58.58
kNN	100	0.69	15.09
kNN	50	0.69	7.58
kNN	10	0.55	2.54
SVM	4000	0.87	189.32
SVM	2000	0.86	69.63
SVM	500	0.83	17.54
NN	4000	0.85	215.73
NN	2000	0.84	87.12
NN	500	0.81	39.93

Table 4.6: TF-IDF with SVD

In this last experiment, SVD dimension reduction is applied to the resulting vector space model from TF-IDF. Similar with experiment before that apply SVD, each of the classification models would be tested with 2 set of vector space model, each with different number of features. To put it in perspective, the number of features without reduction is 8000.

When the number of features are at 4000 which is halved, kNN still can produced an accuracy of 0.77 which is similar with what is achieved with TF-IDF without dimension reduction. Same goes to SVM and NN, at 4000 features, the accuracy are quite similar with TF-IDF without dimension reduction. However, when the dimension of the vector space model is reduced to 2000, the accuracy across 3 of the classification models dropped. kNN being the most drastic, its accuracy dropped to 0.58 while SVM and NN dropped to 0.86 and 0.84 respectively, which is slightly worse than before but it is still satisfactory.

Comparing with the results of the experiment with TF-IDF and naive dimension reduction, this performance of the classification models with SVD has a slight advantage. The accuracy is slightly better with SVD rather than with naive dimension reduction. This

would be due to the advantage of SVD obtaining the maximum variance of the features over naive dimension reduction.

The time taken in this experiment is much higher than the experiment of TF-IDF without dimension reduction and TF-IDF with naive dimension reduction which is expected. This would be due to SVD reduction takes more time as more calculation is needed to transform the data. However, the time taken decreased when the further reduction is done, kNN take 457s to reduce the vector space model from 8000 to 4000 but just 2s to reduce the vector space model from 8000 to 10. Keep in mind that the time recorded here includes the time taken for the classification model to predict the test dataset as well as the dimension reduction time. This trend appear in SVM and NN as well. SVM took 189s to reduce 8000 to 4000 but just 18s to reduce 8000 to 500 while NN took 215s to reduce 8000 to 4000 and 40s to reduce 8000 to 500.

The decrement in time could be explained by the reduction of features, as the amount of features decreases, the time taken to process them also reduced. Therefore there is a speedup. This speedup that comes with the reduction of features, comes at a cost which is accuracy. For SVM and NN, the reduction in accuracy is meagre thus it would be logical to trade the slight accuracy with the speedup but in kNN the trade off would be lopsided in the favour of time taken.

4.8 Conclusion

From the results of the 6 experiments above, it is found that TF-IDF is a better document representation algorithm than term frequency. The resulting vector space model from TF-IDF can achieve a higher accuracy than that of term frequency.

The effect of dimension reduction on the accuracy of the classification models is analysed. Dimension reduction, naive and otherwise, do reduce the dimension of the vector space model, reducing the memory needed to store the matrix. However, this reduction in features and information would result in a loss of accuracy. SVD would be a better dimension reduction algorithm compared to the naive method because the accuracy achieved with SVD is higher than that of naive method.

Out of the 3 classification models tested in the experiments, SVM is the most efficient and versatile. SVM can achieve high accuracy (> 0.80) in most scenarios. NN can also achieve high accuracy in most of the cases tested but NN is more time consuming. SVM has an advantage over NN with the processing time. Therefore, SVM would be the most efficient text classification model among the 3 classification models.

CHAPTER 5: CONCLUSION

It is known that TF-IDF should be better than term frequency but Moldagulova, Aiman and Sulaiman, Rosnafisah proposed that with term frequency and naive dimension reduction, kNN classification model would be able to achieve a satisfactory result. (Moldagulova & Sulaiman, 2018). Therefore, this research compare term frequency and TF-IDF and apply different dimension reduction methods to both. The implementation of the kNN algorithm in this research might differ from that proposed by the authors. They divided the larger feature matrix into parts and train by part.

From the results of the experiments it is shown that TF-IDF is a better document representation method than term frequency. Classification models can achieved a better result with TF-IDF than term frequency. With term frequency as the document representation method, SVM and NN are able to achieve satisfactory accuracy but the accuracy achieved with TF-IDF is slightly higher. The advantage of TF-IDF is more obvious, with kNN, with term frequency the accuracy is a meagre 0.31 but with TF-IDF, the accuracy is 0.76 which is close to 0.80.

By applying the 2 types of dimension reduction to both the document representation method, the effect of dimension reduction on the performance of the classification models is analysed. The classification models accuracy from both naive dimension reduction and SVD are quite similar in most cases. SVD has a slight edge that can provide classification models with more prominent features and subsequently higher accuracy.

As the dimension of the vector space model is reduced, the accuracy of the classification models would decreased due to the lack of features or information. This reduction of features would reduce the memory used to store the vector space model. However, the reduction process would need intensive computing power and time, especially in the case

of SVD. This the trade off between dimension reduction and classification accuracy. Less memory to store the matrix would cause an increase in computing power consumed and time taken.

Out of the 3 classification models tested in this research, SVM has the best overall performance. SVM can achieve a satisfactory accuracy in most scenarios and the time taken for SVM to predict the test data is among the shortest. Thus SVM would be the best classification models in text classification.

For future works, ways to improve the accuracy of the classification models with TF-IDF should be explored. More document representation method, such as n-gram could be apply alongside TF-IDF and different dimension reduction could be applied. TF-IDF does not take the semantic of the term into account just the frequency and rarity, n-gram would be able to rectify part of the problem.

REFERENCES

Ababneh, J., Almanmomani, O., Hadi, W., El-Omari, N., Alibrahim, A. (2014, 02). Vector space models to classify arabic text. *International Journal of Computer Trends and Technology (IJCTT*, *7*, 219-223. doi: 10.14445/22312803/IJCTT-V7P109

Allab, K., Labiod, L., Nadif, M. (2017, Jan). A semi-nmf-pca unified framework for data clustering. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), 2-16. doi: 10.1109/TKDE.2016.2606098

Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting tf–idf. *Computer Speech and Language*, *56*, 107-129. Retrieved from http://www.sciencedirect.com/science/article/pii/S0885230817302887 doi: https://doi.org/10.1016/j.csl.2019.01.005

Chien, J.-T. (2019). Chapter 5 - nonnegative matrix factorization. In J.-T. Chien (Ed.), *Source separation and machine learning* (p. 161 - 229). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780128045664000176 doi: https://doi.org/10.1016/B978-0-12-804566-4.00017-6

Dumais, S., Platt, J., Heckerman, D., Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on information and knowledge management* (pp. 148–155). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/288627.288651 doi: 10.1145/288627.288651

Hinton, G. E. (2002, August). Training products of experts by minimizing contrastive divergence. *Neural Comput.*, *14*(8), 1771–1800. Retrieved from http://dx.doi.org/10.1162/089976602760128018 doi: 10.1162/089976602760128018

Karami, A. (2017, Nov). Taming wild high dimensional text data with a fuzzy lash. In *2017 ieee international conference on data mining workshops (icdmw)* (p. 518-522). doi: 10.1109/ICDMW.2017.73

Ma, J., Yuan, Y. (2019). Dimension reduction of image deep feature using pca. *Journal of Visual Communication and Image Representation*, *63*, 102578. Retrieved from http://www.sciencedirect.com/science/article/pii/S1047320319301932 doi: https://doi.org/10.1016/j.jvcir.2019.102578

Manevitz, L., Yousef, M. (2007). One-class document classification via neural networks. *Neurocomputing*, *70*(7), 1466 - 1481. Retrieved from http://www.sciencedirect.com/science/article/pii/S092523120600261X (Advances in Computational Intelligence and Learning) doi: https://doi.org/10.1016/j.neucom.2006.05.013

Manevitz, L. M., Yousef, M. (2002, March). One-class svms for document classification. *J. Mach. Learn. Res.*, *2*, 139–154. Retrieved from http://dl.acm.org/citation.cfm?id=944790.944808

Mohamed, A. (2019). An effective dimension reduction algorithm for clustering arabic text. *Egyptian Informatics Journal*. Retrieved from http://www.sciencedirect.com/science/article/pii/S1110866518301579 doi: https://doi.org/10.1016/j.eij.2019.05.002

Moldagulova, A., Sulaiman, R. B. (2018, Oct). Document classification based on knn algorithm by term vector space reduction. In *2018 18th international conference on control, automation and systems (iccas)* (p. 387-391).

Narhari, S. A., Shedge, R. (2017, Dec). Text categorization of marathi documents using modified lingo. In *2017 international conference on advances in computing, communication and control (icac3)* (p. 1-5). doi: 10.1109/ICAC3.2017.8318771

Ranjan, N. M., Prasad, R. S. (2018). Lfnn: Lion fuzzy neural network-based evolutionary model for text classification using context and sense based features. *Applied Soft Computing*, *71*, 994 - 1008. Retrieved from http://www.sciencedirect.com/science/article/pii/S1568494618304046 doi: https://doi.org/10.1016/j.asoc.2018.07.016

Shinde, S., Joeg, P., Vanjale, S. (2017, 09). Web document classification using support vector machine. In *2017 international conference on current trends in computer, electrical, electronics and communication (ctceec)* (p. 688-691). doi: 10.1109/CTCEEC.2017.8455102

Su, Y., Huang, Y., Kuo, C. . J. (2018, Aug). Efficient text classification using tree-structured multi-linear principal component analysis. In *2018 24th international conference on pattern recognition (icpr)* (p. 585-590). doi: 10.1109/ICPR.2018.8545832

Sweeney, E., Vogelstein, J., Cuzzocreo, J., A Calabresi, P., Reich, D., M Crainiceanu, C., T Shinohara, R. (2014, 04). A comparison of supervised machine learning algorithms and feature vectors for ms lesion segmentation using multimodal structural mri. *PloS one*, *9*, e95753. doi: 10.1371/journal.pone.0095753

APPENDIX A: DATA PREPROCESSING IMPLEMENTATION

The code in this section show how the 20 news group dataset is processed before it is used to train classification models.

```
"""
Read the raw dataset downloaded from uci repository into a csv
    file in 2 columns, namely text and category
"""
import os
import pandas as pd
import logging
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer
from nltk.corpus import wordnet
import time
import re

nltk.download('stopwords')
nltk.download('wordnet')
ps = PorterStemmer()
lemmatizer = WordNetLemmatizer()
logging.basicConfig(level=logging.INFO, format='%(asctime)s - 
    %(message)s')

output_loc = '../output/20newsGroup18828.csv'


def get_files():
    path = '/home/dante/development/datasets/20news-18828/'
    # path = '/home/db/development/datasets/20news-18828/'

    # Getting all the files path
    files = []
    for r, d, f in os.walk(path):
        for file in f:
            files.append(os.path.join(r, file))

    # Read the content of the file
    text = []
    text_class = []
    for f in files:
        logging.info('reading file: %s', f)
        text_class.append(get_category(f))
```

```
        with open(f, 'r', encoding='utf-8',
            errors='backslashreplace') as reader:
            text_str = reader.read()
            text_str = text_str.encode('utf-8').decode('utf-8',
                'replace')
            # text_str = text_str.replace('\n', ' ')
            text.append(pre_process_text(text_str).strip())

    # Create the dataframe from the data read
    list_of_tuple = list(zip(text, text_class))
    df = pd.DataFrame(list_of_tuple, columns=['text', 'category'])
    # Remove rows with empty text cell
    df = df[df.text != '']
    df.to_csv(output_loc, index=False, encoding='utf8')


def get_category(path):
    """
    Obtain the category of the text by getting it's parent
        directory name
    :param path: the file path
    :return: the category
    """
    split_filepath = path.split('/')
    return split_filepath[-2]


def pre_process_text(string):
    prefixes = ['Xref', 'Path', 'From', 'Newsgroup', 'Subject',
        'Summary', 'Keywords', 'Message-ID', 'Date',
                'Expires', 'Followup-to', 'Distribution',
                    'Organization','Approved', 'Supercedes', 'Lines',
                'X-Newsreader', 'References', 'NNTP-Posting-Host',
                    'In-reply-to', 'Sender', 'News-Software',
                'Article-I.D.', 'Article␣I␣D']
    string_list = string.split('\n')
    new_line = []
    for line in string_list:
        if line.startswith(tuple(prefixes)):
            continue
        else:
            # Remove email addresses
            tmp_line = re.sub('\S*@\S*\s?', '', line)
            # Lower the case
            tmp_line = tmp_line.lower()
            # Remove stopwords
            tmp_line = stopwords_removal(tmp_line)
            # Remove all symbols, retain only alphabets and numbers
            # tmp_line = re.sub('[^A-Za-z0-9]+', ' ', tmp_line)
            tmp_line = re.sub('[^A-Za-z]+', '␣', tmp_line)
```

```
            # Remove all the single letter
            tmp_line = re.sub(r"\s+[a-z]{1}[\s+]", "␣", tmp_line)
            tmp_line = re.sub(r"\s+[a-z]{1}$", "", tmp_line)
            tmp_line = re.sub(r"^[a-z]{1}\s+", "", tmp_line)
            # Remove all extra whitespace
            tmp_line = re.sub(r"\n", "␣", tmp_line)
            tmp_line = re.sub(r"\s+", "␣", tmp_line)
            # Stemming or lemmatization
            # tmp_line = stemming(tmp_line)
            tmp_line = lemmatization(tmp_line)
            # Strip the leading and trailing whitespace
            tmp_line = tmp_line.strip()
            new_line.append(tmp_line)
    new_text = '␣'.join(new_line)
    return new_text


def stopwords_removal(words: str) -> str:
    stop_words = set(stopwords.words('english'))
    more_stop_words = ['article', 'writes', 'write', 'say',
        'nntp', 'posting', 'host', 'berkeley', 'edu', 'hmm', 'll',
                       'wo', 't', 'reply', 'think', 'u', 'go', 've',
                          'repost', 'e', 'mail', 're', 'r', 'o',
                          'hey',
                       'hi', 'n', 'dear', 'reader']
    stop_words.update(more_stop_words)
    tokens = word_tokenize(words)
    filtered_words = [w for w in tokens if not w in stop_words]
    return '␣'.join(filtered_words)


def stemming(words: str) -> str:
    tokens = word_tokenize(words)
    stemmed = [ps.stem(word=w) for w in tokens]
    return '␣'.join(stemmed)


def lemmatization(words: str) -> str:
    tokens = word_tokenize(words)
    pos = nltk.pos_tag(tokens)
    lemmatized = [lemmatizer.lemmatize(word=item[0],
       pos=get_wordnet_pos(item[1])) for item in pos]
    return '␣'.join(lemmatized)


def get_wordnet_pos(treebank_tag):
    if treebank_tag.startswith('J'):
        return wordnet.ADJ
    elif treebank_tag.startswith('V'):
        return wordnet.VERB
```

```
    elif treebank_tag.startswith('N'):
        return wordnet.NOUN
    elif treebank_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN


def main():
    start_time = time.time()
    get_files()
    logging.info("Time␣taken:␣%.2fs", (time.time() - start_time))


if __name__ == "__main__":
    main()
```

APPENDIX B: DATA EXPLORATORY IMPLEMENTATION

The code below is used to check the distribution of the categories in the dataset.

```
import pandas as pd
import matplotlib.pyplot as plt

input_df = pd.read_csv('../output/20newsGroup18828.csv')

print(input_df.shape)

fig, ax = plt.subplots(figsize=(17, 7))
agg_df =
   input_df.groupby('category').count().sort_values('category')
print(agg_df)
fig_plot = agg_df.plot(kind='bar', ax=ax, title='The␣count␣of␣
   each␣category␣of␣the␣dataset',
           legend=False)
fig_plot.set_xlabel('Categories')
fig_plot.set_ylabel('Count')

plt.tight_layout()
# plt.autoscale()
# plt.show()
plt.savefig('../output/count.png', format='png')
```

APPENDIX C: FEATURE EXTRACTION AND DIMENSION REDUCTION IMPLEMENTATION

The code shown in this section is the code that used to generate the feature matrix and apply the dimension reduction to the feature matrix.

```
"""
Move the feature extraction part out of every script for ease of
    maintenance
"""
from sklearn.feature_extraction.text import CountVectorizer,
    TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.decomposition import NMF, TruncatedSVD,
    LatentDirichletAllocation
from sklearn import preprocessing
from sklearn import pipeline
from sklearn.base import BaseEstimator, TransformerMixin
from numpy import ravel
import pandas as pd
import numpy as np
import scipy
from scipy.sparse import csr_matrix, csc_matrix
import logging


logging.basicConfig(level=logging.INFO, format='%(asctime)s - %
    (message)s')
pd.set_option('display.width', 320)
np.set_printoptions(linewidth=320)
input_file = "../output/20newsGroup18828.csv"


def generate_tfidf():
    df = pd.read_csv(input_file)
    label = preprocessing.LabelEncoder()
    df['category_id'] = label.fit_transform(df.category)

    # Some details about the data
    # print(df.groupby(['category',
        'category_id']).count().sort_values('category_id'))

    # Limit the features
    tfidf = TfidfVectorizer(analyzer='word', stop_words='english',
        token_pattern=r'\w+', max_features=8000, lowercase=True,
                            use_idf=True, smooth_idf=True)
    x = tfidf.fit_transform(df.text)
```

```
    y = df['category_id']

    x_train, x_test, y_train, y_test = train_test_split(x, y,
        test_size=0.2, random_state=42)
    # To make space in memory
    del df
    return x_train, y_train, x_test, y_test


def generate_tfidf_reduced(factor: int):
    df = pd.read_csv(input_file)
    label = preprocessing.LabelEncoder()
    df['category_id'] = label.fit_transform(df.category)

    # Some details about the data
    # print(df.groupby(['category',
        'category_id']).count().sort_values('category_id'))

    # Limit the features
    tfidf = TfidfVectorizer(analyzer='word', stop_words='english',
        max_features=8000, lowercase=True,
                            use_idf=True, smooth_idf=True)
    # tokenizer=r'\w+'
    x = tfidf.fit_transform(df.text)
    y = df['category_id']

    print(x.shape)
    mat = x.tocsc()
    greaterThanOne_cols = np.diff(mat.indptr) > factor
    new_indptr = mat.indptr[np.append(True, greaterThanOne_cols)]
    new_shape = (mat.shape[0],
        np.count_nonzero(greaterThanOne_cols))
    x2 = csc_matrix((mat.data, mat.indices, new_indptr),
        shape=new_shape)
    x2 = x2.tocsr()
    print(x2.shape)

    x_train, x_test, y_train, y_test = train_test_split(x2, y,
        test_size=0.2, random_state=42)
    # To make space in memory
    del df
    return x_train, y_train, x_test, y_test


def generate_tfidf_svd(no_of_features: int):
    df = pd.read_csv(input_file)
    label = preprocessing.LabelEncoder()
    df['category_id'] = label.fit_transform(df.category)

    # Limit the features
```

```
    tfidf = TfidfVectorizer(analyzer='word', stop_words='english',
       token_pattern=r'\w+', max_features=8000,
                           lowercase=True,
                           use_idf=True, smooth_idf=True)
    x = tfidf.fit_transform(df.text)
    y = df['category_id']

    svd = TruncatedSVD(n_components=no_of_features, n_iter=7,
       random_state=42, tol=0.0)
    x_reduced = svd.fit_transform(x)

    x_train, x_test, y_train, y_test = train_test_split(x_reduced,
       y, test_size=0.2, random_state=42)

    return x_train, y_train, x_test, y_test


def generate_tf():
    df = pd.read_csv(input_file)
    label = preprocessing.LabelEncoder()
    df['category_id'] = label.fit_transform(df.category)
    count_vect = CountVectorizer(analyzer='word',
       stop_words='english', max_features=8000, lowercase=True)
    x = count_vect.fit_transform(df.text)
    y = df['category_id']

    x_train, x_test, y_train, y_test = train_test_split(x, y,
       test_size=0.2, random_state=42)

    return x_train, y_train, x_test, y_test


def generate_tf_reduced(factor: int):
    df = pd.read_csv(input_file)
    label = preprocessing.LabelEncoder()
    df['category_id'] = label.fit_transform(df.category)
    count_vect = CountVectorizer(analyzer='word',
       stop_words='english', lowercase=True)
    x = count_vect.fit_transform(df.text)
    y = df['category_id']

    print(x.shape)
    mat = x.tocsc()
    greaterThanOne_cols = np.diff(mat.indptr) > factor
    new_indptr = mat.indptr[np.append(True, greaterThanOne_cols)]
    new_shape = (mat.shape[0],
       np.count_nonzero(greaterThanOne_cols))
    x2 = csc_matrix((mat.data, mat.indices, new_indptr),
       shape=new_shape)
    x2 = x2.tocsr()
```

```
    print(x2.shape)

    x_train, x_test, y_train, y_test = train_test_split(x2, y,
        test_size=0.2, random_state=42)

    return x_train, y_train, x_test, y_test


def generate_tf_svd(no_of_features: int):
    df = pd.read_csv(input_file)
    label = preprocessing.LabelEncoder()
    df['category_id'] = label.fit_transform(df.category)
    count_vect = CountVectorizer(analyzer='word',
        stop_words='english', max_features=8000, lowercase=True)
    x = count_vect.fit_transform(df.text)
    y = df['category_id']

    svd = TruncatedSVD(n_components=no_of_features, n_iter=7,
        random_state=42, tol=0.0)
    x_reduced = svd.fit_transform(x)

    x_train, x_test, y_train, y_test = train_test_split(x_reduced,
        y, test_size=0.2, random_state=42)

    return x_train, y_train, x_test, y_test
```

APPENDIX D: KNN CLASSIFICATION MODEL IMPLEMENTATION

The code for kNN classification model is shown below.

```
from GenTfIdf import generate_tfidf, generate_tf,
    generate_tf_reduced, generate_tfidf_reduced,
    generate_tfidf_svd, \
     generate_tf_svd
from sklearn.metrics import confusion_matrix, accuracy_score,
    classification_report
from sklearn.neighbors import KNeighborsClassifier
import pandas as pd
import numpy as np
import time
import math

pd.set_option('display.width', 320)
np.set_printoptions(linewidth=320)

start_time = time.time()
x_train, y_train, x_test, y_test = generate_tfidf()
# x_train, y_train, x_test, y_test = generate_tfidf_svd(50)
# x_train, y_train, x_test, y_test = generate_tfidf_reduced(50)
# x_train, y_train, x_test, y_test = generate_tf()
# x_train, y_train, x_test, y_test = generate_tf_svd(10)
# x_train, y_train, x_test, y_test = generate_tf_reduced(1000)
print(x_train.shape)

k = int(math.sqrt(x_train.shape[0]))
# Rule of thumb for k is sqrt(sample size)
classifier = KNeighborsClassifier(n_neighbors=k)
classifier.fit(x_train, y_train)

predicted = classifier.predict(x_test)

result = confusion_matrix(y_test, predicted)
print(result)
accuracy = accuracy_score(y_test, predicted)
print("accuracy score: " + accuracy.astype(str))
report = classification_report(y_test, predicted)
print(report)
print('Total time taken: {0:.2f}s'.format(time.time() -
    start_time))
```

APPENDIX E: SVM CLASSIFICATION MODEL IMPLEMENTATION

SVM classification model implementation is shown as follows.

```
from sklearn.metrics import confusion_matrix, accuracy_score,
    classification_report
from sklearn.svm import LinearSVC
import time
from GenTfIdf import generate_tfidf, generate_tf,
    generate_tf_reduced, generate_tfidf_reduced,
    generate_tfidf_svd, \
     generate_tf_svd
import pandas as pd
import numpy as np

pd.set_option('display.width', 320)
np.set_printoptions(linewidth=320)

start_time = time.time()
x_train, y_train, x_test, y_test = generate_tfidf()
# x_train, y_train, x_test, y_test = generate_tfidf_svd(500)
# x_train, y_train, x_test, y_test = generate_tfidf_reduced(500)
# x_train, y_train, x_test, y_test = generate_tf()
# x_train, y_train, x_test, y_test = generate_tf_svd(500)
# x_train, y_train, x_test, y_test = generate_tf_reduced(500)
print(x_train.shape)

clf = LinearSVC(random_state=0, tol=1e-5)
clf.fit(x_train, y_train)

predicted = clf.predict(x_test)
result = confusion_matrix(y_test, predicted)
print(result)
accuracy = accuracy_score(y_test, predicted)
print("accuracy score: " + accuracy.astype(str))
report = classification_report(y_test, predicted)
print(report)
print('Total time taken: {0:.2f}s'.format(time.time() -
    start_time))
```

APPENDIX F: NN CLASSIFICATION MODEL IMPLEMENTATION

NN classification model implementation is shown as follows.

```
from GenTfIdf import generate_tfidf, generate_tf,
    generate_tf_reduced, generate_tfidf_reduced,
    generate_tfidf_svd, \
     generate_tf_svd
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import confusion_matrix, accuracy_score,
    classification_report
import time
import pandas as pd
import numpy as np

pd.set_option('display.width', 320)
np.set_printoptions(linewidth=320)

start_time = time.time()
x_train, y_train, x_test, y_test = generate_tfidf()
# x_train, y_train, x_test, y_test = generate_tfidf_svd(500)
# x_train, y_train, x_test, y_test = generate_tfidf_reduced(500)
# x_train, y_train, x_test, y_test = generate_tf()
# x_train, y_train, x_test, y_test = generate_tf_svd(500)
# x_train, y_train, x_test, y_test = generate_tf_reduced(500)
print(x_train.shape)

# Have to manually interrupt it to produce result
clf = MLPClassifier(tol=1e-3)
clf.fit(x_train, y_train)

predicted = clf.predict(x_test)

result = confusion_matrix(y_test, predicted)
print(result)
accuracy = accuracy_score(y_test, predicted)
print("accuracy score: " + accuracy.astype(str))
report = classification_report(y_test, predicted)
print(report)
print('Total time taken: {0:.2f}s'.format(time.time() -
    start_time))
```