Alexandria Engineering Journal (2021) 60, 3401–3409

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

Efficient English text classification using selected Machine Learning Techniques

Xiaoyu Luo *

Hunan University of Technology and Business, China

Received 30 December 2020; revised 30 January 2021; accepted 6 February 2021
Available online 21 February 2021

KEYWORDS

Text classification;
English language;
Machine Learning;
Text mining;
Support Vector Machines

Abstract Text classification (TC) is an approach used for the classification of any kind of documents for the target category or out. In this paper, we implemented the Support Vector Machines (SVM) model in classifying English text and documents. Here we did two analytical experiments to check the selected classifiers using English documents. Experimental results performed on a set of 1033 text document present that the Rocchio classifier provides the best performance results when the size of the feature set is small while SVM outperforms the other classifiers. From the experimental analysis, we observed that the classification rate exceeds 90% when using more than 4000 features.

1. Introduction

Text Classification using machine learning consists of providing input to a text document to a set of pre-defined classes, using a machine learning technique. The classification is normally carried out on the basis of selected documents and features using text documents [1]. However, the classes selected before the experiment analysis and it's called supervised machine learning operation. Although, organization and companies maintained their records using text documents and email for industrial service and government. Also, individual communication and records existed in text and document conversations. The reason text classification demands increase due to the existence of a large amount of text information that existed randomly. In order to classify this information need a machine learning approach [3]. This research provides investigations on an enhanced approach for text classification which includes steps such as pre-processing approach which is entirely based on eliminating stop-words and stemming [2,4]. In-text classification statistical analysis is used to provide analytical and comparative research for feature selection using Naïve Bayes, Decision Tree, Neural Network, Support Vector Machines, Hybrid techniques. This paper also discusses some of the major issues involved in English text classification which includes interacting with misclassified text, treatment of a wide number of text features. This paper also deals with the selection of efficient machine learning techniques by providing a comparative analysis. Through comparative analysis, we select the efficient machine learning algorithm for our text classification. We have used an efficient technique for our English text classification. Pre-processing techniques for text classification and natural language processing-based approaches are applied

* Address: No. 569, Yuelu Road, Hunan University of Technology and Business, Changsha, Hunan 411104, China.
E-mail address: cathyluo1981@gmail.com.

Peer review under responsibility of Faculty of Engineering, Alexandria University.

https://doi.org/10.1016/j.aej.2021.02.009

in order to train the datasets. Machine learning-based text classification is considered to be more useful for applications that include the classification of text documents available in digital format [1,5]. These applications and importance can be identified from the use of spam filtering in email, web services, fake currency identification, fake news identification, and opinion mining techniques. As trends of blogging have taken popularity in online services using the internet, text and content classification play a vital role in this field as well. As new dictionaries and key terms are used for blogging by the bloggers through online systems. Massive amounts of data are usually transferred on daily basis from one organization to another organization which leads to unstructured data. In order to categorize such an amount of data researchers normally used ML techniques for categorization. This also helps to produce structured and well-managed content in the form of classes or clusters. However to implement the text mining or classification approach for opinion or sentiment analysis. This approach is adapted to keep track of newly added slang words or dictionary terms. It's observed that the nature of blog contents are likely to be added and delivered through web services on a daily schedule. Moreover, text posts on a blog do not strictly adhere to the blog topic. This leads to the introduction of a more enhanced, incremental, and multi-topic text classification approach. Based on such scenarios it's quite an emergence to develop machine learning tools and approaches for some of the popular languages such as English, Chinese, Arabic, and Hindi or Urdu. Our paper is structured as below: In Section 2 we have provided the generic strategy for text classification and the selection of a better text classification approach. Section 3 provides the main issues and problems related to text classification and the existing state of the art. The remaining sections consist of methodology, results, and simulations. Text Mining (TM) is considered an important approach for the phases of the Knowledge Discovery in Text (KDT) and classification of similar information or documents [2]. DM is mainly related to finding the related data and significant correlations, clusters, and drifts by filtering a large number of well-organized data, stored in the databases or data warehouses. Unlike the traditional DM which can be abbreviated as data mining techniques is the subdomain of AI. TM approaches are used for mining the unstructured data into well-structured data [3]. The applications of TM techniques are in use since 2010 to classify scientific journals and documents [1]. In 2015, Shrihari and Desai [12] proposed a method for the comparative analysis of classification approaches used in TM. From the comparative analysis, it was obvious that the approach with the highest accuracy rate is the Support Vector Machines (SVM) and Naïve Bayes (NB). These classification techniques are called the supervised machine learning techniques. Using a supervised machine learning approach the researchers develop a model before starting the experimental work. Once the analysis is done then a statistical approach is used to evaluate the model attributes. These machine learning classification can be used to create structured classes and clusters of similar data. In machine learning text data can be classified into predefined classes [5]. SVM is an approach that is used to predict and define how to classify new categories for a sample of datasets [6]. During the classification method that based on the NB algorithm, training examples are also used, but the document belonging to a given class is calculated based on the formula for finding the probability, that is mainly closely related to Bayes theorem and referencing the document to a class. In NB, each document subject to classification is considered as a vector of independent words [7]. According to Shrihari and Desai [12], SVM's accuracy is 90.21% as compared to the NB's accuracy is 79.83%. Researchers have used the Naïve Bayes algorithm to classify journal Research Documents [8]. According to the results of their research, the accuracy of the approach is high enough. In order to achieve greater accuracy in the classification of scientific papers, in this research paper we have applied two algorithms analyzed algorithms such as SVM and NB.

2. Literature

Normally, text classification can be categorized into different types such as supervised and unsupervised. In the case of text mining, the process of predefined class to train an "unknown" natural language processing text was proposed by Moreno and Redondo, it was represented as the single-label TC function. Most of the research presented on classification includes SVM, Naïve Bayes, and KNN algorithms. Each class ck $\in$ C during the testing phase of text classification. In order to define each and every document which is represented as dj where dj is the document's number. D represents the total count of the documents [9]. However, in order to justify multiple predefined clusters to an "unknown" text as presented by Feng et al. (2005). However, it is very common to present it as the multi-label TC task, whereas any number $0 < nj \leq |C|$ of classes may be predicted to each document dj $\in$ D. "TC represent text classification and its used for single-label" (Allahyari et al., 2017 which is a separate classifies neither a predefined class nor its extra to an "un-predicted" class (Sebastiani, 2006). In the past, several types of research were studied [10].

Naïve Bayes algorithm is used for pattern recognition and classification that falls under different variations in pattern classifiers for the basic probability and likelihood. The likelihood of such techniques was proposed by Lausch et al [11]. The simplest way for text classification is based on providing the well know Bayes formula. Naïve Bayes techniques are based on the self-determining suppositions. These hypotheses are clearly largely implemented in NLP with different variations provide the text outline of document, text, semantic, syntactic, and rational [12]. Afterward, the provisional independence declaration is implied for the occurrences are independence identified the category and therefore the required classification prediction is provided which is shown in Fig. 2 as below:

NB classification considers word frequency by keeping it simple keeps as shown in Fig. 1.

In Fig. 2 is shown the working of the proposed framework for English language document classification. In the first step data collection was performed from an open-source library such as the UCI library [13]. We selected the datasets for the UCI library based on English language documents. These datasets were fathered about English books. After selecting and gathering the data collection the next step is the preprocessing which has provided complete details in the methodology section [14–17]. Once pre-processing is done then we performed feature extraction for our experiments. The feature that we have selected mainly includes the word frequency, questions marks, full stop, initial word, and the final word of

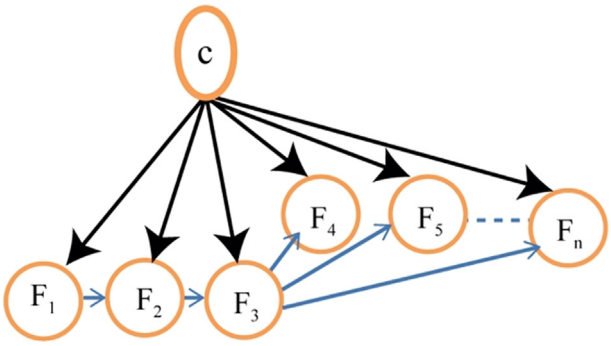

Fig. 1 Naïve Bayes classification.

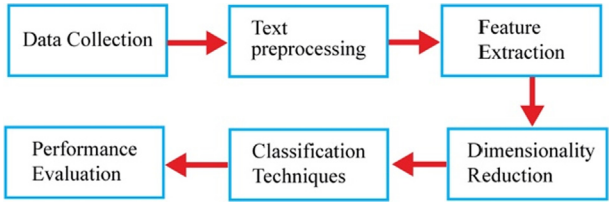

Fig. 2 Proposed Text Classification framework.

the documents [18–22]. We then reduce the dimension of the datasets by filtering and selecting the best feature for our experimental analysis. We choose two classification methods for our English text mining i.e SVM and KNN respectively.

3. Support vector machine

Support Vector Machine (SVM) approach is used to classify related documents using the vectors approach. This technique was explored by Khamar et al. to provide a two-class research problem that depends on the splitting among hyperplanes presented by classes of data [6]. In this approach based on training the sample such as {(xi, ci)}Ni is taken into account. For the input topic, the i^{th} case position of the word is required for the output which is considered as an output [26–30]. Although, this approach based on some of the conditions such as that the groups comprehensive by the subsection and the theme chosen by the subsection are "linearly divisible". In Fig. 3 it's shown in detail the working of the SVM method that how the topic separation between hyperplanes is provided. From this figure, it's very clear that the SVM algorithm may be used even for enormous frequency datasets as it's clear that the main objective of SVM is to scale the boundaries of the split of the data, which utilized rather than fits on features. The SVM techniques can be trained using predefined categorized text.

In SVM normally two basics parameters are used which are weight vector w and related bias b. These parameters are used to provide a separation between the hyperplane and the neighbouring spacial point is signified by ρ that it is signified boundary of parting. The aim of the SVM is to achieve the uppermost of the precise hyperplane for the side-line of separation ρ. However, the conclusion area is represented as the finest hyperplane. Fig. 3 represents the regular construction of an optimal hyperplane for two types of datasets. Each dataset is represented by a different shape such as a rectangle and triangle.

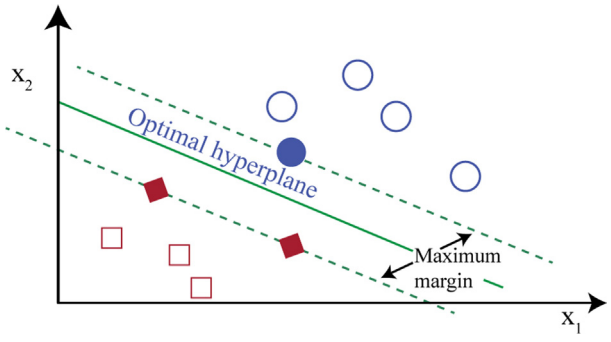

Fig. 3 SVM using Hyperplane.

4. K-nearest neighbours

In machine learning, the role of K-Nearest Neighbours (KNN) is considered an important approach. The aim of using the KNN concept is to predict the set of nearest similar datasets. It is widely used for the class of a quantified request based not only on the text but also on the nearest to it in the text region. K represents the number of neighbours in the available datasets. The KNN approach is based on the similarity learning approach which is applied for many data analytics and text classification approaches and domains. A test document is used to predict the category is to while the KNN classifier resides the closest neighbours within the learning texts then we uses the classes of the k neighbours to provide the value to the respected class. In Fig. 4 we have provided the concept of the KNN techniques and the clusters concept using different colours and shapes.

Basically, there exist many supervised ML approaches that are based on document classifications. Text mining is considered a vital approach in AI [31]. Text mining approaches can be categorized as unsupervised, semi-supervised, and supervised respectively. Orthodoxly, this technique can be solved automatically. Although, these manual classifications are con-

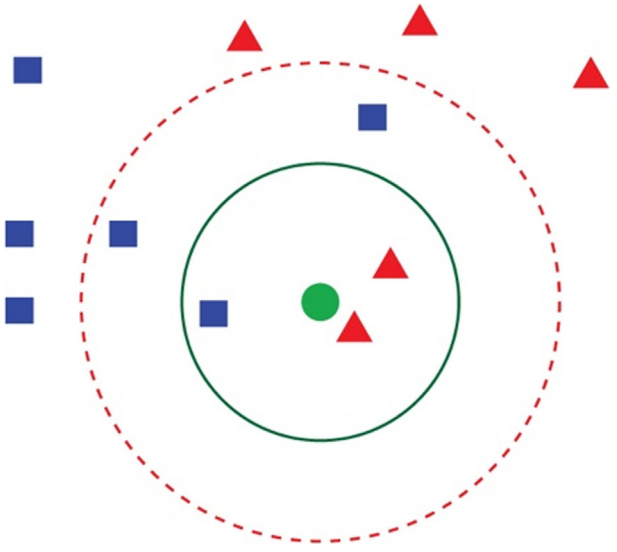

Fig. 4 Classification using KNN approach.

sidered very costly to use as well as required more time for mining and the communication cost is also too exhaustive [32–35]. Therefore, in the literature, most of the researchers have provided the analysis of the applications of supervised machine learning techniques to unsupervised text classification. The technique based on supervised ML that the basic I/O relative is trained using a minor amount of training values and hence it provides the output values for anonymous input data that have been expected [36].

Text labeling is used to assign a predefined class according to the requirements. Therefore, classification using ML provides both a supervised ML approach and an unsupervised approach. In the supervised ML approach, the training process is considered to be "supervised" if and only if the information accessible for the categorization as well as for the learning themes whichever is often related and similar. In the case of text mining, the main aim of supervised machine learning is to provide a learning set of samples in order that it can be used to train more predefined class labels to new available datasets text. In the literature, the researcher has also provided an enhanced KNN technique for text labeling and categorization in order to provide pattern matching [37]. The well-known method for converting bias for larger is provided through the KNN method. The method that's used to stratify on Chinese text is measured as very simple therefore this method can be conceptually possible to provide a solution for classifying the datasets. However, similar research problems are investigated by the researcher in the literature [38–43].

In the same way, Rane et al. explored the more improved version of the performance evaluation using the KNN technique. For the first time the researcher has provided and feature selection techniques based on Markup language and syntactical similarity score. Based on these approaches many large web-based contents are demanded large run in order to classify the demanded whole Web-based contents. However, the KNN approach using mixing KNN and genetic techniques in order to enhance the classification and labeling accuracy and efficiency. Therefore the enhancement in the correctness for classification might decrease the difficulties in the KNN. Although, the simulations result for the performance are compared by the authors with the normal KNN and benchmarking. An enhanced version of this technique was investigated by the Vapnik in their research. Later on, the same context was improved more through the application of the AKNN text technique and kMdd integration. In this method, the local weigh assignment for TF values for feature collection and the cosine resemblance values are accomplished by the researcher in AkNN. The classical method based on KNN used Reuters-21578 datasets, whereas the comparative analysis is also provided with the benchmarking. The investigator has reached a solution that the traditional KNN provides better performance as compared to other classification techniques in both categorizations as well as clustering. This approach was investigated by Sharmila et al. [27]. KNC classification method was used for the integration of KNN algorithm and very important categorization such as C4.5 technique, NB and SVM. The integration and features of these techniques were observed on the classification specification of different types of samples. The researchers investigated that the KNC algorithm is specially used to improve the accuracies of the selected technique. Naive Bayes classifier, C4.5, and SVM. The concept of KNC provides more accuracy as compared to the performances with AdaBoost. Bijalwan et al. Proposed a novel M1 technique by selecting the features of Naive Bayes and an SVM classifier. The author provided four machine learning approaches which include Language Model (LM), KNN, NB, and TF-ID. However, in this classification, the author used trending topics and training datasets. Han et al. proposed a novel supervised learning technique based on k-NN. The main concept behind the enhanced approach was to provide an enhanced efficiency to train the pattern as well as to provide the related texts. This technique of mixing the NNN classifier, TF-IDF, and framework used for text categorization. However, this model provides classification conferring to different factors, slashes, and investigation of the investigation. In literature, the researcher has also provided four classifiers which include DT, SVM, NB, and KNN. These four classifiers were based on ontology-based multiple for text categorization. Most of the authors found that the best classifier is NB coupled with the binary relevance transformation approach for single-label, while the best value is the HOMER-based categorization approach for multiple with in order to provide performance measurements investigated by Suguna and Thanushkodi (2010). However, the NB technique is a private of modest probabilistic technique based on a public assumption that all these attributes are autonomous for every value respectively. Furthermore, the authors proposed mainly three Bayesian equivalents, which provide a classical NB approach using the Bernoulli event framework in relation to the Bayesian counterpart based on 20 newsgroups, and the WebKB dataset was studied by Kurada and Pavan (2013). Text labeling based on a novel approach that requires the least texts for learning. NB approach based on stemmed frequently repeating the word and added genetic algorithm for the last categorization is used. The author justified that as it increases for both data learning and testing then the computation time that impacts the performance was improved quietly. Numerous research on document categorization has been only provided in a few areas. From the experiment, it's very obvious that SVM performs well as compared to the NB approach for text enrichment through Wikitology was explored by Xu (2018). They investigate that the NB technique provides more efficiency when they outperformed while using an external knowledge base. This approach is totally based on Naive Bayes with the integration of learning support vector machine. However, it's observed that the NB algorithm specifically applied to train the SVM whereas SVM is applied for new text categorization. Some of the researchers in the literature investigated that the technique for text mining is not only dependent on some value; however, it also often used to improve the precision of categorization accordingly. It has also been compared with the existing SVM algorithm was. The existing SVM technique was proposed by Kamruzzaman and Haider (2010). In a similar way, research has been applied by Shathi et al. (2016) using three dissimilar text categorizations framework such as the vector space model. This research used the NB and a novel implementation based on two different datasets such as 30 Newsgroup and Novel dataset containing fewer data.

The researchers investigated that the NB based approach worked significantly enough than the remaining two classification approaches. The SVM and NB approaches are used to classify every article based on ANT quantity its correctness is predefined in a group. The researcher explored that the SVM technique provides more efficiency than the NB for both

title and text parts was applied by Hassan et al. (2011). Alternatively, the utilization of meta-features in most of the document classification has provided important improvements in the efficiency of classification techniques. However, the meta-feature extraction approach is widely dependent on intensive applications of the KNN techniques to explore the nearest information related to the neighbourhood of the trained documents. This approach was proposed by Canuto et al. (2014) with an improvement in efficiency. A new method to hierarchical document classification like HDLTex that was used to integrate the multiple deep learning methods to produce the hierarchical classifications was investigated by (Kowsari et al., 2017). In another research detection of fake reviews via sentiment analysis based on ML techniques was provided by Elmurngi and Gherbi (2017). The author concluded that the SVM approach provides more efficiency than the K* and KNN–IBK. Huang et al. (2013) explored short text categorization based on sentiment analysis. Text classification issues were highlighted by Huang et al. (2013). Based on these techniques the relationship among the tweets can be identified easily. Aytekin (2013) proposed the idea of opinion mining using text classification and NLP.

5. Methodology

5.1. Datasets

In Fig. 6 we used the same text classification approach using three types of datasets. Data 1 comprised of four categories which consist of women (40), sports (120), and literature (30) last but not least the campus which counts 18 respectively. For the next categories, we selected test2 data. This tets2 data also comprised of four sub-categories which includes sports 21, constellation 22, game (23) as well as entertainment (20). For the third category, we have selected test3 data it consists of three sub three sub-categories. We named these three categories as Science and Technology consist of a total count of 16, fashion consist of a total number of count 11, current event consist of 18 respectively. The samples were collected from the news websites. Most of them are English, but some of them are mixed with special symbols and URLs. The datasets were mixed but through text mining, we have provided the classification using SVM as shown in Fig. 6.

5.2. Evaluation metrics

In order to evaluate the analysis matrix, we have divided the evaluation process into four categories. These four categories include True positive (TP), false negative (FN), True negative (TN), false positive (FP). We did the calculation by using the accuracy and precision equation as provided through equation (1) and equation (2) alternatively. In our experiment, the models of text classification we use are Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and Logistic Regression CV (LRCV).

$$Accuracy = \frac{Ps}{Ts} \tag{1}$$

We use Eq. (1) in order to find the accuracy of our simulation results. In this equation, Ps represented the predicted sample and Ts represents the total number of samples used in our experiments.

For precision we use the following equation:

$$Precision = \frac{TP}{TP + FN} \tag{2}$$

In Eq. (2) TP is for true positive, FN is for false negative. By inserting the values we can get the precision.

6. Results

From the results in Table 1, we have calculated the precision, recall, and F1 value based on our simulation. The equations are provided in Eqs. (1) and (2) respectively. From Table 1 it's very obvious that SVM techniques provide more efficiency as compared to Naive Bayes and Logistic Regression. We simulated it in Weka using datasets consist of English text documents. The datasets are freely available on the UCI library. Our proposed technique can be also applied for BBC English text classification. In order to run the BBC dataset require more hardware specification and also it requires tensors flow or python.

From Table 1 it's very clear that SVM has more precision as compared to NB and LR for the datasets that we have selected. We tested SVM, NB, and LR for three types of datasets available from the English news website. In Fig. 1 text classification for our datasets is provided. In training data 46 values were true and 56 were false. While in testing we received 56 true positive (TP) and 46 false negative (FN) respectively.

In these results, Fig. 1 represents the simulation of our English language documents using Weka.

Figs. 5, 6 and 7 represents the simulation achieved during the training of datasets using Weka. The evaluation mode we selected for 10-fold cross-validation. The number of seeds we set was only 1 seed. We have chosen ten attributes during our testing experiment. For each, we selected different attributes. After training our algorithm on a sample of data we test our new data set for English text documents. The results are shown in Fig. 5 with the evaluation are provided.

In Fig. 5 the same datasets are evaluated using the cluster method for the same number of attributes. We evaluate

Table 1 Comparative analysis of the machine learning algorithm.

	Data1			Data 2			Data 3		
	Precision	**Recall**	**F1 Value**	**Precision**	**Recall**	**F1 Value**	**Precision**	**Recall**	**F1 Value**
SVM	0.88	0.87	0.86	0.76	0.71	0.71	0.63	0.54	0.51
NB	0.40	0.28	0.24	0.27	0.29	0.18	0.12	0.33	0.18
LR	0.83	0.85	0.81	0.67	0.57	0.49	0.64	0.63	0.63

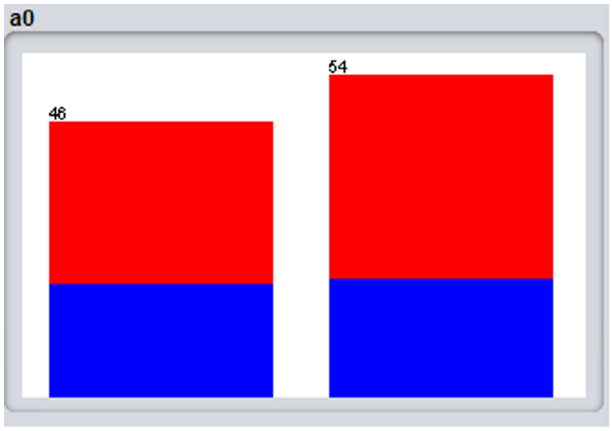

Fig. 5 Text Classification for document 1.

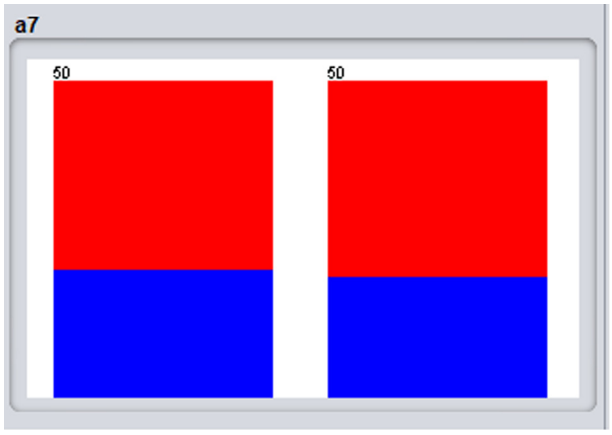

Fig. 6 Classification of documents 2.

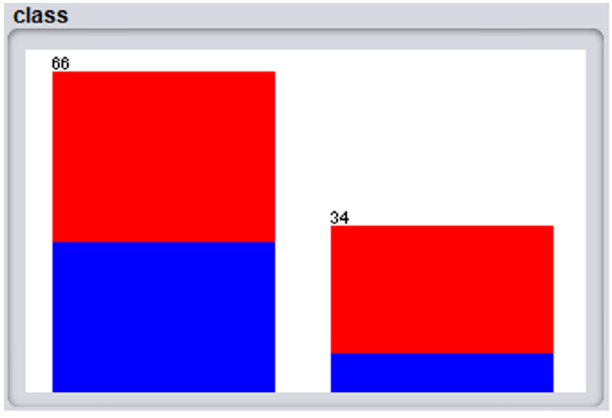

Fig. 7 Classification of the datasets using Weka.

through the cluster method by providing two clusters. Cluster 0 and cluster 1 respectively (see Figs. 9 and 10).

In Fig. 11 we used the same text classification approach using three types of datasets. Data1 comprised of four categories which consist of women (40), sports (120), and literature (30) last but not least the campus which counts 18 respectively. For the next categories, we selected test2 data. This tets2 data

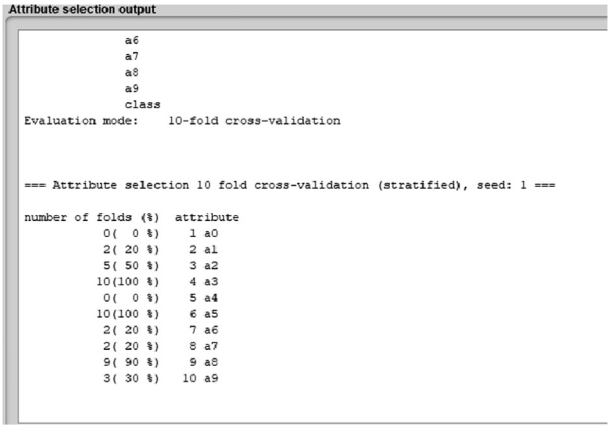

Fig. 8 Attribute selection and cross-validation for English Text Mining Using SVM.

also comprised of four sub-categories which includes sports 21, constellation 22, game (23) as well as entertainment (20). For the third category, we have selected test3 data it consists of three sub-categories. We named these three categories as Science and Technology consist of the total count of 16, fashion consist of a total number of count 11, current event consist of 18 respectively. The samples were collected from the news websites. Most of them are English, but some of them are mixed with special symbols and URLs. The datasets were mixed but through text mining, we have provided the classification using SVM as shown in Fig. 11.

7. Discussion

There exist several text classification issues while performing text classification that is related to the text mining approach. It was experiential that the data gathering procedure greatly leads to the succeeding stages of pre-processing, attribute abstraction, and eventually text mining. In Text pre-processing the type of data moulded that are collected in dissimilar approaches of pre-processing with dissimilar investigation results. The dataset must be organized via tokenization, stop words removal, stemming, and vector space document in order to be easy to use it. In order to extract features, there are some parameters that we have considered for our experiments. These parameters include the number of words which is called frequency in machine learning. Feature selection, dimension, and extraction. We have selected these parameters for text classification in order to improve the text classification process. From the literature, we found that if during experiments and simulations the learning text is increased then the evaluation can be increased. These attributes are used to increase the number of the terms that it is automatically composed to a class generated a good categorization of sample texts. Inadequate amount of samples of testing text documents, Features for the dataset depends on the selected machine learning approach and the size of the datasets. We have also provided a comparison of the different algorithms used for text classification. The performance has been evaluated using the English documents based dataset, the dataset we have used for our experimental work are available at the UCI repository.

```
                a6
                a7
                a8
                a9
                class
Evaluation mode:    10-fold cross-validation


=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0      +- 0        1    +- 0       11 class
 0      +- 0        2    +- 0        3 a2
 0      +- 0        3    +- 0        2 a1
 0      +- 0        4    +- 0       10 a9
 0      +- 0        5    +- 0        5 a4
 0      +- 0        6    +- 0        6 a5
 0      +- 0        7    +- 0        7 a6
 0      +- 0        8    +- 0        8 a7
 0      +- 0        9    +- 0        9 a8
 0      +- 0       10    +- 0        1 a0
```

Fig. 9 Evaluation of attributes for 10-fold cross-validation.

```
               Cluster
Attribute            0        1
                  (0.66)   (0.34)
============================
a0
  false          30.2327  17.7673
  true           37.8843  18.1157
  [total]         68.117   35.883
a1
  false          32.2914  20.7086
  true           35.8256  15.1744
  [total]         68.117   35.883
a2
  false          42.1239  17.8761
  true            25.993   18.007
  [total]         68.117   35.883
a3
  false           31.781    9.219
  true           36.3359  26.6641
  [total]         68.117   35.883
a4
  false          36.1452  19.8548
  true           31.9718  16.0282
  [total]         68.117   35.883
a5
  false          44.4417   1.5583
  true           23.6753  34.3247
```

Fig. 10 Text mining techniques using Cluster Method.

8. Conclusion and future work

From the literature, we concluded that the applications of text classification in the field of IT are growing rapidly. Our paper mainly aims to provide text classification, feature selection, and performance evaluation using Weka as an experimental tool. Our proposed approach can also be implemented in R, Tensors flow, Python, or Matlab simulation tool. In the initial stage, we performed pre-process in order to select the best feature such as frequency, initial letter, paragraph, question mark, and full stop consequently. We extract the text features from the English language-based documents using a supervised machine learning approach. We compared also provide the comparative analysis of different machine learning approaches which includes NB, SVM, and LR. From the simulations, it's very clear that the SVM outperforms the rest of the machine learning techniques for the datasets that we have used. The evaluation and comparison were achieved using some selected parameters such as precision, recall, and F1 value. Finally, we provide a discussion of the performance of our selected machine learning techniques. From our discussion, it is very obvious that each and every machine learning classification

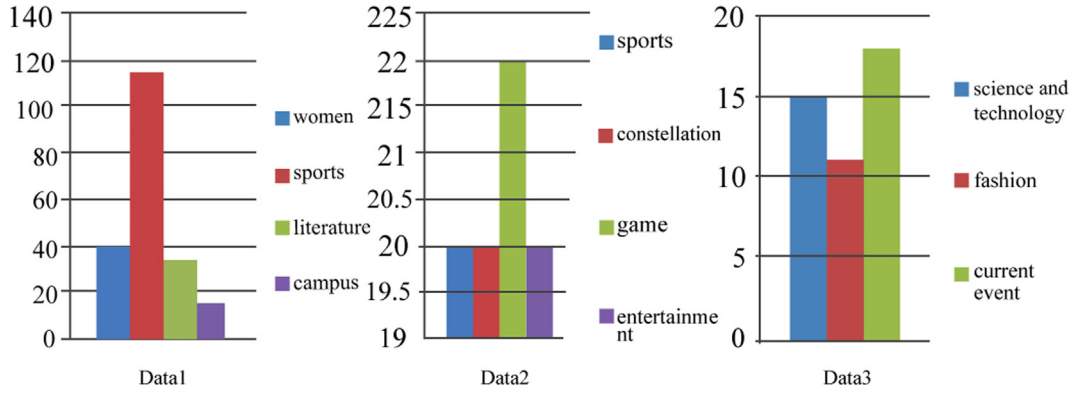

Fig. 11 Simulation results of the selected datasets.

techniques have advantage and disadvantage depends on the size of the datasets.

Funding

This work was supported by the Hunan Province Degree and Graduate Education Reform Research Project Grant (2020JGYB28), the Hunan Province Education Scientific Research Project Grant (20B152), and the Hunan Province Social Science Achievements Review Project (XSP21YBZ106).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Azaria et al, Medrec: Using blockchain for medical data access and permission management, 2016 2nd International Conference on Open and Big Data (OBD), IEEE, 2016.
[2] S. Sulova et al, Using text mining to classify research papers, Int. Multidiscip. Sci. GeoConference Survey. Geol. Mining Ecol. Manage. SGEM 17 (21) (2017) 647–654.
[3] A.H. Mohammad, T. Alwada'n, O. Al-Momani, Arabic text categorization using support vector machine, Naïve Bayes and neural network, GSTF J. Comput. (JoC) 5 (1) (2016) 108.
[4] M.S. Sabuj, Z. Afrin, K.A. Hasan, Opinion mining using support vector machine with web based diverse data, International Conference on Pattern Recognition and Machine Intelligence, Springer, 2017.
[5] B. Ghaddar, J. Naoum-Sawaya, High dimensional data classification and feature selection using support vector machines, Eur. J. Oper. Res. 265 (3) (2018) 993–1004.
[6] Y. Ke, M. Hagiwara, An English neural network that learns texts, finds hidden knowledge, and answers questions, J. Artif. Intell. Soft Comput. Res. 7 (4) (2017) 229–242.
[7] L. Al-Horaibi, M.B. Khan, Sentiment analysis of Arabic tweets using text mining techniques, First International Workshop on Pattern Recognition, International Society for Optics and Photonics., 2016.
[8] M. Bilal et al, Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques, J. King Saud Univ.-Computer Informat. Sci. 28 (3) (2016) 330–344.
[9] P.V. Ngoc et al, A C4. 5 algorithm for english emotional classification, Evolving Syst. 10 (3) (2019) 425–451.
[10] A.I. Kadhim, Y.-N. Cheah, N.H. Ahamed, Text document preprocessing and dimension reduction techniques for text document clustering, 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, IEEE, 2014.
[11] E. Vellingiriraj, M. Balamurugan, P. Balasubramanie, Information extraction and text mining of Ancient Vattezhuthu characters in historical documents using image zoning, 2016 International Conference on Asian Language Processing (IALP), IEEE, 2016.
[12] S. Vijayarani, M.J. Ilamathi, M. Nithya, Preprocessing techniques for text mining-an overview, Int. J. Comput. Sci. Commun. Networks 5 (1) (2015) 7–16.
[13] Z. Yao, C. Ze-wen, Research on the construction and filter method of stop-word list in text preprocessing, 2011 Fourth International Conference on Intelligent Computation Technology and Automation, IEEE, 2011.
[14] S. Vijayarani, R. Janani, Text mining: open source tokenization tools-an analysis, Adv. Comput. Intell.: Int. J. (ACII) 3 (1) (2016) 37–47.
[15] A.K. Uysal, S. Gunal, The impact of preprocessing on text classification, Inf. Process. Manage. 50 (1) (2014) 104–112.
[16] M. Eder, J. Rybicki, M. Kestemont, Stylometry with R: A package for computational text analysis, The R Journal 8 (1) (2016).
[17] G. Miner et al, Practical text mining and statistical analysis for non-structured text data applications, Academic Press, 2012.
[18] Z. Ju, J. Wang, F. Zhu, Named entity recognition from biomedical text using SVM, 2011 5th international Conference on Bioinformatics and Biomedical Engineering, IEEE, 2011.
[19] V.N. Phu, V.T.N. Chau, V.T.N. Tran, SVM for English semantic classification in parallel environment, Int. J. Speech Technol. 20 (3) (2017) 487–508.
[20] N. Li, D.D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decis. Support Syst. 48 (2) (2010) 354–368.
[21] C. Silva, B. Ribeiro, On text-based mining with active learning and background knowledge using svm, Soft. Comput. 11 (6) (2007) 519–530.
[22] B. Brahimi, M. Touahria, A. Tari, Data and text mining techniques for classifying Arabic tweet polarity, J. Digital Informat. Manage 14 (1) (2016).
[26] F. Peng, D. Schuurmans, S. Wang, Augmenting naive bayes classifiers with statistical language models, Inf. Retrieval 7 (3–4) (2004) 317–345.
[27] B.Y. Pratama, R. Sarno, Personality classification based on Twitter text using Naive Bayes, KNN and SVM, 2015 International Conference on Data and Software Engineering (ICoDSE), IEEE, 2015.
[28] M. Panda, Developing an efficient text pre-processing method with sparse generative naive Bayes for text mining, Int. J. Modern Educat. Comput. Sci. 10 (9) (2018).
[29] Y.E. Soelistio, M.R.S. Surendra, Simple text mining for sentiment analysis of political figure using naive bayes classifier method. arXiv preprint arXiv:1508.05163, 2015.
[30] Z. Gong, T. Yu, Chinese web text classification system model based on Naive Bayes, 2010 International Conference on E-Product E-Service and E-Entertainment, IEEE, 2010.
[31] A.C. Tantug, Document categorization with modified statistical language models for agglutinative languages, Int. J. Computat. Intell. Syst. 3 (5) (2010) 632–645.
[32] F. Peng, D. Schuurmans, S. Wang, Language and task independent text categorization with simple language models, Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003.
[33] C. Chelba, A. Acero, M. Mahajan, Discriminative training of language models for text and speech classification. 2008, Google Patents.
[34] W. Chen et al, A novel ensemble approach of bivariate statistical-based logistic model tree classifier for landslide susceptibility assessment, Geocarto International 33 (12) (2018) 1398–1420.
[35] A.J. Dobson, A.G. Barnett, An Introduction to Generalized Linear Models, CRC Press, 2018.
[36] B. Kalantar et al, Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN), Geomatics, Natural Hazards Risk 9 (1) (2018) 49–69.
[37] Y. Yang, M. Loog, A benchmark and comparison of active learning for logistic regression, Pattern Recogn. 83 (2018) 401–415.

[38] T. Fischer, C. Krauss, Deep learning with long short-term memory networks for financial market predictions, Eur. J. Oper. Res. 270 (2) (2018) 654–669.
[39] L.R. Lloyd-Jones et al, Transformation of summary statistics from linear mixed model association on all-or-none traits to odds ratio, Genetics 208 (4) (2018) 1397–1408.
[40] K. Jaidka, N. Chhaya, L. Ungar, Diachronic degradation of language models: Insights from social media, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018.
[41] A. Hasan et al, Machine learning-based sentiment analysis for twitter accounts, Mathe. Computat. Appl. 23 (1) (2018) 11.
[42] S.A. Salloum et al, A survey of Arabic text mining, in: Intelligent Natural Language Processing: Trends and Applications, Springer, 2018, pp. 417–431.
[43] M.H. Abd El-Jawad, R. Hodhod, Y.M. Omar, Sentiment analysis of social media networks using machine learning, 2018 14th international computer engineering conference (ICENCO), IEEE, 2018.

Further Reading

[23] M. Rushdi-Saleh et al, Bilingual experiments with an arabic-english corpus for opinion mining, Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, 2011.
[24] K.A. Hasan, M.S. Sabuj, Z. Afrin, Opinion mining using naive bayes, 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), IEEE, 2015.
[25] Q. Jiang et al, Deep feature weighting in Naive Bayes for Chinese text classification, 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS), IEEE, 2016.