Academic Registration Text Classification Using Machine Learning

Mohammed S Alhawas and Tariq S Almurayziq

University of Ha'il, Ha'il, KSA

Summary
Natural language processing (NLP) is utilized to understand a natural text. Text analysis systems use natural language algorithms to find the meaning of large amounts of text. Text classification represents a basic task of NLP with a wide range of applications such as topic labeling, sentiment analysis, spam detection, and intent detection. The algorithm can transform user's unstructured thoughts into more structured data. In this work, a text classifier has been developed that uses academic admission and registration texts as input, analyzes its content, and then automatically assigns relevant tags such as admission, graduate school, and registration. In this work, the well-known algorithms support vector machine SVM and K-nearest neighbor (kNN) algorithms are used to develop the above-mentioned classifier. The obtained results showed that the SVM classifier outperformed the kNN classifier with an overall accuracy of 98.9%. in addition, the mean absolute error of SVM was 0.0064 while it was 0.0098 for kNN classifier. Based on the obtained results, the SVM is used to implement the academic text classification in this work.
Keywords: *NLP, Deep Learning, Text Classification, ML, Tags.*

1. Introduction

Natural language processing (NLP) systems can analyze an unlimited amount of text data consistently and fairly. You can recognize complex contextual concepts, decipher the language, and extract the most important facts and contexts [1]. Understanding natural language helps machines to understand natural language and simulate the ability of a person to discover and record emotions. Intelligence technology allows you to analyze algorithms with scientific analysis.
Text classification can be defined as a machine learning (ML) technique that assigns some predefined categories to open text. Text classifiers are used to organize, structure, and classify almost any type of text from documents, medical research, files, and the entire web. For example, new articles can be sorted by subject. Support tickets can be organized according to urgency. Chat conversations can be organized by language. Brand mentions can be organized by emotions and more. Text classification is one of the basic tasks of NLP with a wide range of applications such as sentiment analysis, topic labeling, spam detection, and intent detection. This task implements a text classifier that processes, analyzes, and classifies a specified dataset using a Support Vector Machine (SVM) and K-nearest neighbor (KNN) approach. Support Vector Machines (SVMs) are another powerful ML algorithm for text classification because they don't require a lot of training data to produce accurate results. The SVM draws a line or "hyperplane" that divides the room into two sub-rooms.
Therefore, classifying the input text into the correct class improves the text classification process and saves time and effort. In this task, you will use a text classification approach to develop a text classifier that can take academic text input, analyze its content, and automatically assign relevant tags such as admissions, graduates, and registrations.

1.1 Problem Statement

The main objective of this work is to develop a text classifier that can take such scholarly text input, analyze its content, and then automatically assign relevant tags such as: admissions, graduates or registration. To achieve the stated goal, the following objectives must be met:

- Pre-process the dataset and feed the training and testing dataset.
- Implement the text classifier.
- Check and validate the classifier.
- Evaluate the developed classifier.

2. Literature Reviews

ML is the study of computational approaches that create systems with the ability to automatically learn and improve from experience. It is widely seen as a field within artificial intelligence. ML algorithms allow systems to make autonomous decisions without external support. Such decisions are made by finding valuable basic patterns in complex data. Based on your approach to learning, the types of data you input and output, and the types of problems you solve, there are several main types of ML algorithms: supervised learning, unsupervised learning, and reinforcement learning. There are several hybrid approaches and other common methods that provide natural extrapolation from ML problem formats. As mentioned earlier, learning styles can be supervised or unsupervised.

Manuscript received December 5, 2021
Manuscript revised December 20, 2021
https://doi.org/10.22937/IJCSNS.2022.22.1.14

Automatic classification of user-generated orders to automate the allocation of tickets to the correct team of maintenance providers. The classified text is a short, natural linguistic expression. The alternative used so far is to manually select one of the many preconfigured classes, which can be time consuming. It describes related work related to text classification and also considers how to select features for text classification. There is little research on the evaluation of feature selection methods for text classification, but we will investigate evaluation methods from other areas of text classification [2]. Classification can be defined as a method of supervised learning in ML. This is also related to the problem of predictive modeling, where class labels are predicted for specific data [3]. Mathematically, it maps the function (f) from the input variable (X) to the output variable (Y) as a target, label, or category. You can do this on structured or unstructured data to predict the class of a particular data point. For example, spam detection, such as "spam" or "no spam," can cause classification issues for email service providers. However, classification performance identification measurements play a decisive role in the design of classifiers. Evaluation methods and measurements are at least as important as algorithms and are the first important step in successful data mining. There are various content-related research strategies aimed at revealing unstructured data from text. This paper examines the use of objective confidence in literary arrangements that treat text as word vectors [4]. Learning algorithms can be divided into four main types in this area: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [5]. These learning approaches are becoming more and more popular every day. However, different types of ML techniques can play an important role in creating effective models in different application areas, depending on learning ability, data type, and desired results [6].

2.1 kNearest Neighbor

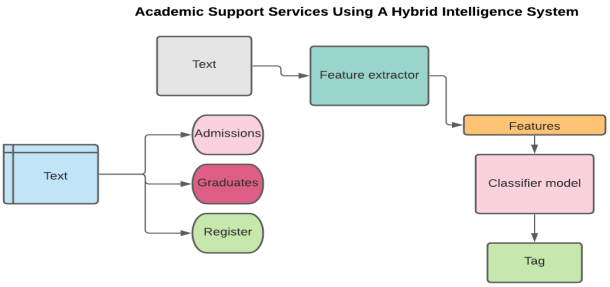
kNN is a non-parametric method for classification. The basic idea of the kNN classifier is to find the k-nearest neighbors of the candidate document from the training document and evaluate the k-nearest neighbor categories [7-8]. Therefore, the candidate document is assigned to the class with the highest score. The kNN algorithm compares the test document with each training sample [9].

2.2 Support Vector Machine

SVM is a monitored learning algorithm for classification and regression. The main goal of SVM is to find the maximum separation hyperplane between the vectors that belong to the category and the vectors that do not. The maximum distance between data points helps to classify future data points more reliably.

A simple text classification system can be deconstructed into the following four phases.

- Text pre-processing
- Dimensionality Reduction



- Classification
- Evaluation

3. Methodology

Text classifiers can be used to organize, structure, and classify almost any type of text, from documents, medical research, files to the entire web. Preprocessing is a very important step for text classification applications. Since text data cannot be manipulated by ML, it must be converted to a numerical vector representation. Word preprocessing involves cleaning up the text data to remove unwanted noise and encoding the text numerically to enable useful feature extraction. Text cleaning in NLP begins with tokenization. Tokenization identifies atomic tokens that do not require any further processing [10]. Text classification requires a natural language parser that determines the syntactic structure of the text by analyzing the constituent words based on the underlying grammar. Figure 1 shows the used methodology.

Figure 1. The used methodology for text classification

1.2 Syntactic word expressions

Text cleanup is usually followed by syntactic word expressions. This is a technique for extracting textual features to resolve the loss of syntactic and semantic relationships between words. You can use the Ngram technique to solve these syntactic problems. Ngram is a combination of adjacent words or letters of length n found in text or document sentences [11]. Ngram uses a probabilistic model to understand the structure of a language and predict the next word and the entire sequence.

1.3 Weighted Words

After a word is syntactically represented, the next step in text preprocessing is to translate the document into a vector containing the frequency of the words [12-14]. This methodology can be simply represented as in figure 2.

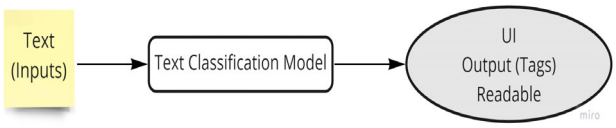


Figure 2. Simplified text classification methodology

4. Results and Discussions

This partition outlines the results obtained, the data used, and the test procedure for processing the survey questions raised. The experimental area focuses on the structure of the dataset, and the experimental and obtained results are displayed in the results area. This error is assumed for all data sets, and then all errors are converted to positive. This is achieved by using the absolute value of each error as follows:

The mean absolute error (MAE) is generally defined as (Eq. 1).

$$\text{Absolute Error} \rightarrow |\text{Prediction Error}| \quad (1)$$

However, mathematically the MAE can be illustrated as (Eq. 2).

$$mae = \frac{\sum_{i=1}^{n} abs(y_i - x_i)}{n} \ldots (2)$$

In this work, the dataset is divided into two subsets, the testing data set and the training dataset. Testing data set consists of 14765 samples in which it structured as 14765 rows X 3 columns. The columns are labeld as category, text, and full text. Category represents the first column that holds register, graduates, or admissions.

For linear SVM, the accuracy score is summarized in figure 3 below. The confusion matrix is a 3X3 matrix, in which its first row represents the admissions category. This classifier predicted 5260 case correctly as admissions (TP), it incorrectly predicted 15 cases as graduates, and it incorrectly predicted 40 cases as register. Looking at column 1, we can find that the total number of admissions sample is 5260 samples. If we look at the second category which is graduates, we notice that the SVM classifier correctly predicted 4705 cases as true positive graduates while it predicted no incorrectly cases. In addition, the last category which is register, the classifier correctly predicted 4745 cases while it incorrectly predicted 0 cases as admissions or as graduates in which they are register. Hence, the accuracy of this model equals to the total number of correctly predicted instances divided by the total number of incorrectly misclassified instances. The total number of correctly predicted can be found through the diagonal of the confusion matrix which are (5260 + 4705 + 4745) = 14710. The misclassified instances are (15+40) = 55. Table 1 shows the confusion matrix of the developed SVM classifier.

Table 1: SVM confusion matrix

Admissions	**5260**	**15**	**40**
Graduates	0	4705	0
Register	0	0	4745
	Admissions	Graduates	Register

For kNN classifier, the first row represents the admissions category. This classifier predicted 5245 case correctly as admissions (TP), it incorrectly predicted 15 cases as graduates, and it incorrectly predicted 55 cases as register. Looking at column 1, we can find that the total number is 5255 admissions samples. If we look at the second category which is graduates, we notice that the kNN classifier correctly predicted 4705 cases as true positive graduates while it predicted no incorrectly cases. In addition, the last category which is register, the classifier correctly predicted 4735 cases while it incorrectly predicted 10 cases as admissions in which they are register.

Hence, the accuracy of this model equals to the total number of correctly predicted instances divided by the total number of incorrectly misclassified instances. The total number of correctly predicted can be found through the diagonal of the confusion matrix which are (5245 + 4705 + 4735) = 14685. Table 2 shows the confusion matrix of the developed kNN classifier.

Table 2: kNN confusion matrix

Admissions	**5245**	**15**	**55**
Graduates	0	4705	0
Register	10	0	4735
	Admissions	Graduates	Register

5. Conclusions

In this work, we have presented a novel approach for text classification based on two different ML methods which are SVM and kNN. Two classification algorithms have been developed and implemented in the open-source web application, the Jupyter notebook. The different results show a marked improvement in the classification rate of texts. The developed models showed an excellent classification accuracy using both SVM and kNN

classification methods. However, the SVM_based classifier outperformed the kNN_based classifier in which they achieved 99.6%, and 99.4% accuracy percentages respectively. In addition, the SVM classifier has a very small mean absolute error (mae) which was very low (0.0064), while the kNN classifier has a bigger mean absolute error (mae) which was (0.0098). The achieved results showed that the SVM classifier is better than the kNN classifier. For future work, we will try to improve the classification performance and try to develop and implement new techniques to overcome the developed ones. In addition, the future work will take in consideration the development of mobile application for the implemented classifiers.

Acknowledgments

I want to thank my university, "University of Ha'il" for providing us with all the needed facilities to complete this master's degree. A special thanks to Gharbi Alshammari for his continuous guide and support. I want to acknowledge and thank my department for allowing me to conduct my research and providing any assistance requested. Finally, I would like to thank all my friends and colleagues who have helped me on this project. Their enthusiasm and willingness to provide feedback made completing this study an enjoyable experience.

References

[1] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *ArXiv Preprint ArXiv:1707.02919*.

[2] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 52(1), 273–292.

[3] Qawqzeh Y, Alharbi MT, Jaradat A, Abdul Sattar KN. A review of swarm intelligence algorithms deployment for scheduling and optimization in cloud computing environments. PeerJ Comput Sci. 2021 Aug 25;7:e696. doi: 10.7717/peerj-cs.696. PMID: 34541313; PMCID: PMC8409329.

[4] Cunningham-Nelson, S., Baktashmotlagh, M., & Boles, W. (2017). From review to rating: Exploring dependency measures for text classification. *ArXiv Preprint ArXiv:1709.00813*.

[5] Mohammed M, Khan MB, Bashier Mohammed BE. Machine learning: algorithms and applications. CRC Press; 2016Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36–54.

[6] Li, Baoli & Yu, Shiwen & Lu, Qin. (2003). An Improved k-Nearest Neighbor Algorithm for Text Categorization.

[7] Abdulameer, A. S., Tiun, S., Sani, N. S., Ayob, M., & Taha, A. Y. (2020). Enhanced clustering models with wiki-based k-nearest neighbors-based representation for web search result clustering. Journal of King Saud University-Computer and Information Sciences.

[8] Sabbah, T., Ayyash, M., & Ashraf, M. (2018). Hybrid support vector machine based feature selection method for text classification. *Int. Arab J. Inf. Technol.*, *15*(3A), 599–609.

[9] Sarker IH. Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Comput Sci. 2021

[10] Srivastava, Durgesh & Bhambhu, L. (2010). Data classification using support vector machine. Journal of Theoretical and Applied Information Technology. 12. 1–7. Ślusarczyk B. Industry 4.0: Are we ready? Polish J Manag Stud. 17, 2018.

[11] Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications*, *43*, 82–92.

[12] Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 7370–7377.

[13] Onan, A., Atik, E., & Yalçın, A. (2020). Machine learning approach for automatic categorization of service support requests on university information management system. *International Conference on Intelligent and Fuzzy Systems*, 1133–1139.

[14] Yu, P., Cui, V. Y., & Guan, J. (2021). Text Classification by using Natural Language Processing. *Journal of Physics: Conference Series*, *1802*(4), 42010.