Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT (JISTM)
www.jistm.com

GLOBAL ACADEMIC EXCELLENCE

A REVIEW OF FEATURE EXTRACTION METHODS ON MACHINE LEARNING

Mustazzihim Suhaidi[1*], Rabiah Abdul Kadir[2], Sabrina Tiun[3]

[1] Institute of IR4.0, Universiti Kebangsaan Malaysia, Malaysia
Email: p100520@siswa.ukm.edu.my
[2] Institute of IR4.0, Universiti Kebangsaan Malaysia, Malaysia
Email: rabiahivi@ukm.edu.my
[3] Faculty of Information Science And Technology , Universiti Kebangsaan Malaysia, Malaysia
Email: sabrinatiun@ukm.edu.my
[*] Corresponding Author

Article Info:

Article history:
Received date: 10.06.2021
Revised date: 15.07.2021
Accepted date: 20.08.2021
Published date: 01.09.2021

To cite this document:

Suhaidi, M., Kadir, R. A., & Tiun, S. (2021). A Review Of Feature Extraction Methods On Machine Learning. *Journal of Information System and Technology Management, 6* (22), 51-59.

DOI: 10.35631/JISTM.622005.

Abstract:

Extracting features from input data is vital for successful classification and machine learning tasks. Classification is the process of declaring an object into one of the predefined categories. Many different feature selection and feature extraction methods exist, and they are being widely used. Feature extraction, obviously, is a transformation of large input data into a low dimensional feature vector, which is an input to classification or a machine learning algorithm. The task of feature extraction has major challenges, which will be discussed in this paper. The challenge is to learn and extract knowledge from text datasets to make correct decisions. The objective of this paper is to give an overview of methods used in feature extraction for various applications, with a dataset containing a collection of texts taken from social media.

Keywords:

Feature Extraction, Classification, Machine Learning

Introduction

Text feature extraction, which extracts text information, is used with the aim of representing a text message; it is the basis of a large number of text processing methods. The basic unit of the feature is called text features. Selecting a set of features from some effective ways to reduce

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

the dimension of feature space is called feature extraction. During feature extraction, uncorrelated or superfluous features will be deleted. As a method of data pre-processing of learning algorithms, feature extraction can better improve the accuracy of learning algorithms and shorten the time. Selection from a document part can reflect the information on the word contents, and this calculation of weight is called text feature extraction. Common methods of text feature extraction include filtration, fusion, mapping, and clustering (Liang *et al.*, 2017).

Text feature extraction holds a crucial role in text classification because it directly influences the classification accuracy. Feature extraction is based on vector space model, where a text is viewed as a dot in an N-dimensional space. Each dimension of the dot represents one feature of the text in digital form (Dzisevic and Sesok, 2019).

Fig.1 shows the flow of text classification. It includes four steps: document pre-processing, feature selection, feature extraction, and text classification. For document pre-processing, techniques like stop word removal, stemming and pruning of word can be used. Feature selection can be defined as a process of selecting a subset from the original feature set on the basis of importance of features. There are three categories of feature selection methods: wrappers, filters, and embedded methods.

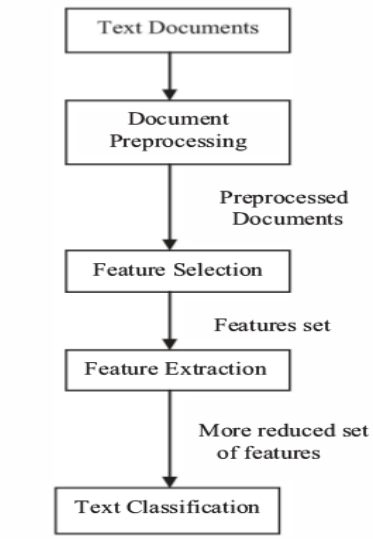

Fig. 1 Text Classification Flow

Source: (Dzisevic and Sesok, 2019)

Feature extraction can be defined as a process of extracting a set of new features from the features set that is generated at the feature selection stage. Feature extraction methods include principal component analysis (PCA), latent semantic indexing (LSI), and clustering methods. Among these methods, PCA is one of the most powerful ones (Shah and Patel, 2016).

This paper outlines the common methods used in text feature extraction first, and then expands frequently used deep learning methods in text feature extraction and its applications and forecasts the application of deep learning in feature extraction. The main review of this paper can be presented as follows:

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

- By reading a large amount of articles, the text feature extraction method and machine learning method is summarized
- A large number of articles have been collected to summarize most of the applications of the present text feature extraction method
- Most of the applications of machine learning in text feature extraction have been summarized
- The application of machine learning method in text feature extraction is prospected and summarized.

Literature Review

This section briefly reviews the literature on feature extraction methods.

Feature Extraction

Principle Component Analysis (PCA): Principal Component Analysis is a dimension reduction technique. PCA extracts the information from various datasets. The goal of PCA is to produce a lower-dimensional feature set from the original dataset. In PCA it is very important to determine the number of principal components. The p number of principal components to be chosen among all of the principal components should be the principal components to represent the data at their very best. There are certain criteria in determining the optimal number of principal components such as broken-stick model, cross-validation, Velicier's partial correlation procedure, Kaiser's criterion, Barlett's test for equality of eigen-values, Cattell's screen-test, and cumulative percentage of variance (Shah and Patel, 2016).

One of the simplest types of feature extraction models is called Bag of Words. The name Bag of Words refers to the fact that this model does not take the order of the words into account. Instead one can imagine that every word is put into a bag, where the ordering of the words gets lost. Although there exist a few different variations of this model, the most common one is to simply count the number of occurrences of each word within a document and keep the result in a vector (Eklund, 2018).

In feature extraction, the original feature space is converted to a more compact, new space. All the original features are transformed into the new reduced space without deleting them, but by replacing the original features with a smaller representative set. When the number of features in input data is too large to be processed, then the input data will be transformed into a reduced representative set of features (Zareapoor and K. R, 2015).

PCA is a well-known technique that can reduce the dimensionality of data by transforming the original attribute space into smaller space. In other words, the purpose of principle component analysis is to derive new variables that are combinations of the original variables and are uncorrelated. This is achieved by transforming the original variables Y = [y1, y2,..., yp] (where p is number of original variables) to a new set of variables, T = [t1, t2,..., tq] (where q is number of new variables), which are combinations of the original variables. Transformed attributes are framed by first, computing the mean (μ) of the dataset, and then calculating a covariance matrix of the original attributes (Zareapoor and K. R, 2015).

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

There are two main types of composite feature extraction methods in text categorization: n-gram and termset (Wan *et al.*, 2019).

1) n-gram
The n-gram extraction process entails using a window with length n to slide through an entire corpus. Then, all the sets of consecutive words or characters in each window are extracted. The purpose of n-gram is to get the composite features that appear continuously to alleviate the ambiguity of individual words. Commonly used n-grams are bigram and trigram. However, the influence of text structure, such as punctuation and stop words, is not considered.

2) termset
A termset is completely different to an n-gram, where composite features are only extracted based on their cooccurrence, irrespective of the order and position of the member terms [18], [19]. In other words, termsets could be defined as arbitrary paired combinations in vocabulary. One problem stems from this combination however, that it is a combination explosion even for 2-termsets. It means that there will be n^2 kinds of combinations for a vocabulary size of n.

Feature Selection
Feature selection process: It consists of four basic steps (shown in Fig. 2), namely, subset generation, subset evaluation, stopping criterion, and result validation (Mutlag *et al.*, 2020).

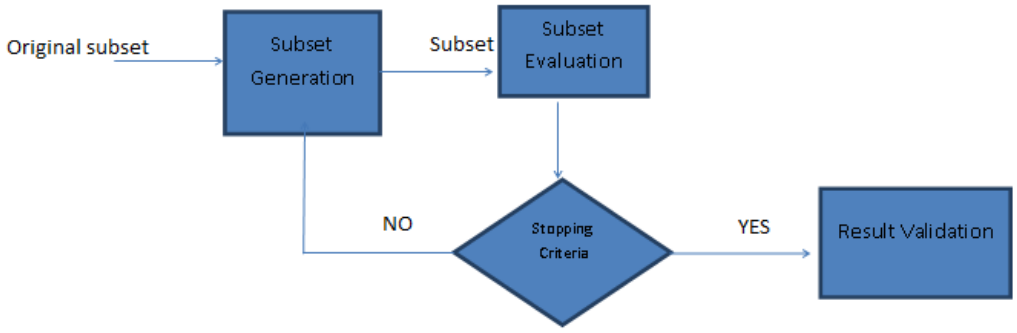

Fig. 2 Four Steps Of Feature Selection

Source: (Mutlag *et al.*, 2020)

When the input data to an algorithm is too large to be processed and suspected to be redundant (e.g., the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). This process is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data (Mutlag *et al.*, 2020).

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

Feature selection is commonly used to reduce the feature dimension and improve the performance of a text classifier. During the process of feature selection, the score of each feature is usually calculated by a general criterion, and then the top N features are selected in the feature subset (N is an experimentally determined number). Chi-square is a widely known statistical method that has played an important role in assessing individual distinguishing power , whose formula is as follows (Wan *et al.*, 2019):

$$\chi^2(t_i) = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}.$$

Where a and c are defined as the number of documents that contain t_i in the positive and negative classes, respectively, and b and d indicate the number of documents that do not contain t_i in the positive and negative classes, respectively. The total number of documents in the training set is N = a + b + c + d.

$$\tilde{\chi^2}(t_{ij}) = \frac{N(\tilde{a}\tilde{d}-\tilde{b}\tilde{c})^2}{(\tilde{a}+\tilde{c})(\tilde{b}+\tilde{d})(\tilde{a}+\tilde{b})(\tilde{c}+\tilde{d})},$$

Where t_{ij} denotes the 2-termset which is made up of terms t_i and t_j , ã and c̃ respectively indicate the number of documents which contain both or either of t_i and t_j in the positive and negative classes, and b̃ and d˜ are the number of documents that not contain any of t_i and t_j in the positive and negative classes, respectively. This means that a subset of the members can also convey information.

In text categorization, we aredealing with a huge feature space. This is why we need a feature selection mechanism. The most popular feature selection methods are document frequency thresholding (DF), the X^2 statistics (CHI), term strength (TS), information gain (IG) , and mutual information (MI). The X^2 statistic measures the lack of independence between the text feature term t and the text category *C* and can be compared to the X^2 distribution with one degree of freedom to judge the extremeness (Masih and Grant, 2017).

Suitable Feature Extraction Algorithm

Feature extraction is highly subjective in nature; it depends on what type of problem you are solving. There is no generic feature extraction algorithm which works in all cases. Like classifiers, it is not possible to say which is the best algorithm for feature selection or extraction. It depends on your application. The questions you can answer and choose the best algorithm are:

- What kind of problem are you solving? (classification, regression, clustering, etc.)
- Do you have a huge set data?
- Does your data have very high dimensionality?
- Is the datalabeled?
- What do you want to perform, feature extraction or feature selection?
- Which method do you want to use, a supervised or an unsupervised method?

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

Selection of text feature items is a basic and important matter for text mining and information retrieval. Feature extraction means that according to the certain feature extraction metrics, the extract is relevant to the original feature subsets from initial feature sets of test sets, so as to reduce the dimensionality of feature vector spaces. During feature extraction, the uncorrelated or superfluous features will be deleted. As a method of data pre-processing of the learning algorithm, feature extraction can better improve the accuracy of learning algorithm and shorten the time. Compared with other machine learning methods, deep learning is able to detect complicated interactions from features, learn lower level features from nearly unprocessed original data, mine characteristics that are not easily detectable, hand class members with high cardinal numbers, and process untapped data.

Discussion

Effects of Pre-processing

N-grams

It seems as though the use of N-grams makes the classifier more restrictive. This probably introduces more false negative predictions, as is evident by examining the deteriorating recall scores. However, since the classifier becomes more restrictive, the chance of false positives decreases, which is shown in the increasing micro-precision scores. One possible reason for this happening could be that certain keywords that would be recognized as useful features on their own, are not recognized as such when combined with one or more other keywords.

An example of this could be two phrases from the same class being: "Microsoft's revenue decreased" and "Apple's revenue remained constant". Here the word "revenue" would be recognized as a common distinguishable keyword for both examples when using 1-grams, but when using 2-grams the keywords would be "Microsoft's revenue" and "revenue decreased" for the first example. These keywords would not match any of the second example's keywords: "Apple's revenue", "revenue remained" and "remained constant".

Stop-Word Elimination

By omitting the use of stop-word elimination, the performance of all the classifier and extraction combinations generally decreases.

Stemming

When the use of stemming is not applied to the data, there are many varying performance differences versus when it is applied. For the TF-IDF method used with the ANN classifier we see a slight increase in performance for the micro-metrics and a slight decrease for the macro-metrics. For the Count Vector method, we see a sharp decline in micro precision. It seems as though the TF-IDF method and Count Vector method are more sensitive to different forms of the same words than the GloVe extraction method, which is logical since these methods do not preserve the semantic meanings of the terms examined.

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

Extraction Methods

TF-IDF

One method that has proven itself to be both simple and effective for feature extraction is Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is an information retrieval technique that can be used to determine the relevance of terms in documents in relation to a query.

Bag of Words/Count Vector

The Count Vector is similar to the TF-IDF vector, but it does not take the length of the current document, nor other documents within a class, into consideration. It simply counts the number of times that each word occurs within a document and stores the results in a vector. A maximum feature length of 100 is also chosen for this extraction method.

Glove

The GloVe feature extractor is implemented by employing a pre-trained GloVe model. The model has been trained on data from Wikipedia and Gigaword 5. This data contains 6 billion tokens, and a vocabulary of 400,000 words. The length of the vectorized word representations is chosen to be 100 in order for this approach to be fairly compared with the TF-IDF and Count Vector method.

Bag of Words vs. Word Embeddings

The Bag of Words model looks to be more sensitive to the different pre-processing methods that are used, while the Word Embedding model yields similar results regardless. Which approach is better is debatable, and depends largely on which metric performance is most important for the problem at hand.

Concerns

More accurate multi-label classifications would hopefully lead to a more efficient use of both energy and time. When applied to a recommender system, it would hopefully mean that people could get relevant items or articles recommended to them both faster and more accurately. When it concerns articles regarding pensions, which was the original starting point of this work, it would mean that people would hopefully be more informed regarding their financial decisions and both their future and current well-being. Moreover, it would ideally reduce the need for direct communication between the Swedish Pensions Agency and people with questions regarding their pension. This could lead to reduced spending of tax-money, but could also mean that some people would lose their jobs.

Table 1: Summarization of Feature Extraction Method

Feature Extraction Methode	
Author (Year)	***Feature Extraction Method***
Kuang, S. and Davison, B. D. (2017)	Word Embeddings With Chi-Square(Kuang and Davison, 2017)
Masih, M. and Grant, A. (2017)	Chi square feature extraction based SVMS

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

Carducci, G. *et al (2018)*	Word Embeddings based SVMS(Carducci *et al.*, 2018)
Yang, H. *et al* (2019)	Feature Scoring and Extraction(Yang *et al.*, 2019)
Kim, S., Kim, J. and Chun, H. W (2018)	Wave2vec (Kim, Kim and Chun, 2018)
Liu, Q. *et al.* (2016)	Word Embeddings with weighted contexts based on part-of-speech (POS) relevance weights(Liu *et al.*, 2016)
Kholghi, M. *et al.* (2016)	Word Embeddings Features(Kholghi *et al.*, 2016)
Brennan, P. M. *et al.* (2017)	GloVe: Global Vectors for Word Representation(Brennan *et al.*, 2017).

Conclusion

One remarkable conclusion can be obtained from a review of some papers not all feature extraction methods are beneficial to classification performance. Nor is there an extraction method that performs the best across both classifiers. However the extraction method can have a significant impact on the results of multi-label classification. The best choice of extraction method depends on the what the multi-label classifications are to be used for. If the priority lies with not producing false negatives, then the results of this work indicate that the GloVe extraction method is the best choice. If, however, not producing false positives is the highest priority, then the Bag of Words extraction method is the superior choice.

The structure of the data that is to be used is also an important factor to take into consideration when choosing an extraction method. The dataset used for this work consisted of heavily imbalanced real-world data. For this dataset, the best overall result was achieved by using the TF-IDF method used in conjunction with stop-word elimination, using stemming, not using N-grams, and an SVM classifier.

Removal of stop-words generally had a positive effect on the results. Only when using the GloVe extraction method did it prove to be detrimental for some metrics. Not removing stop-words had a great negative effect on the SVM classifier when the TF-IDF extraction method was used.

Acknowledgement

The authors would like to express gratitude to the National University of Malaysia (UKM) for providing the opportunity and funding under the Research Project Code TAP-K020558.

References

Brennan, P. M. *et al.* (2017) 'Pre-operative obesity does not predict poorer symptom control and quality of life after lumbar disc surgery', *British Journal of Neurosurgery*, 31(6), pp. 682–687. doi: 10.1080/02688697.2017.1354122.

Carducci, G. *et al.* (2018) 'TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning', *Information (Switzerland)*, 9(5), pp. 1–20. doi: 10.3390/info9050127.

Dzisevic, R. and Sesok, D. (2019) 'Text Classification using Different Feature Extraction Approaches', *2019 Open Conference of Electrical, Electronic and Information Sciences, eStream 2019 - Proceedings*, pp. 1–4. doi: 10.1109/eStream.2019.8732167.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 6 Issue 22 (September 2021) PP. 51-59
DOI: 10.35631/JISTM.622005

Eklund, M. (2018) 'Comparing Feature Extraction Methods and Effects of Pre-Processing Methods for Multi-Label Classification of Textual Data', *Degree Project Computer Science and Engineering*.

Kholghi, M. *et al.* (2016) 'The Benefits of Word Embeddings Features for Active Learning in Clinical Information Extraction'. Available at: http://arxiv.org/abs/1607.02810.

Kim, S., Kim, J. and Chun, H. W. (2018) 'Wave2Vec: Vectorizing electroencephalography bio-signal for prediction of brain disease', *International Journal of Environmental Research and Public Health*, 15(8). doi: 10.3390/ijerph15081750.

Kuang, S. and Davison, B. D. (2017) 'Learning word embeddings with chi-square weights for healthcare tweet classification', *Applied Sciences (Switzerland)*, 7(8). doi: 10.3390/app7080846.

Liang, H. *et al.* (2017) 'Text feature extraction based on deep learning: a review', *Eurasip Journal on Wireless Communications and Networking*, 2017(1), pp. 1–12. doi: 10.1186/s13638-017-0993-1.

Liu, Q. *et al.* (2016) 'Part-of-Speech Relevance Weights for Learning Word Embeddings'. Available at: http://arxiv.org/abs/1603.07695.

Masih, M. and Grant, A. (2017) 'Chi square feature extraction based SVMS arabic language text categorization system', *Talent Development and Excellence*, 9(2), pp. 18–26. doi: 10.3844/jcssp.2007.430.435.

Mutlag, W. K. *et al.* (2020) 'Feature Extraction Methods: A Review', *Journal of Physics: Conference Series*, 1591(1), pp. 22558–22577. doi: 10.1088/1742-6596/1591/1/012028.

Shah, F. P. and Patel, V. (2016) 'A review on feature selection and feature extraction for text classification', *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*, pp. 2264–2268. doi: 10.1109/WiSPNET.2016.7566545.

Wan, C. *et al.* (2019) 'Composite Feature Extraction and Selection for Text Classification', *IEEE Access*, 7(c), pp. 35208–35219. doi: 10.1109/ACCESS.2019.2904602.

Yang, H. *et al.* (2019) 'Dynamic Slide Window-Based Feature Scoring and Extraction for On-Line Rumor Detection with CNN', *IEEE International Conference on Communications*, 2019-May, pp. 1–6. doi: 10.1109/ICC.2019.8761288.

Zareapoor, M. and K. R, S. (2015) 'Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection', *International Journal of Information Engineering and Electronic Business*, 7(2), pp. 60–65. doi: 10.5815/ijieeb.2015.02.08.