Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230669

Comparison and analysis of the machine learning in the movie subtitle document classification

Buxiao Chu[1,†], Dannong Xu[2,†] and Dongze Wu [3,4,†]

[1]Faculty of Computer Intelligence, Peking University, Beijing, Beijing, 100091, China
[2]Faculty of Science, University of Melbourne, Parkville, Melbourne, 3053, Australia
[3]Faculty of Engineering, University of Sydney, Camperdown, Sydney, 2006, Australia

[4]dongzew@student.unimelb.edu.au
[†]These authors contribute equally to this work.

Abstract. With the rapid iteration of film and technology, the quantity of movie related subtitle document has been rapidly expanded. However, the issue of classifying these documents is becoming non-negligible. This paper aims to compare different methods of classification of every machine learning algorithm on movie subtitle document and then analyse the result of each method. For the classification of movie subtitle document, this paper mostly does the following features of the work: firstly, retrieve of movie subtitle document and pre-processing step of text document. Secondly, deletion of the irrelevant text and extraction of the related texts, the methods are unsupervised machine learning approach and supervised machine learning approach respectively. Furthermore, analyse each classifier in various model which applied with different approach. Lastly, compare the result of different approach and each model then display the results.

Keywords: Movie subtitle classification, supervised learning, unsupervised learning.

1. Introduction

Together with the enormous rise of virtue online society, the scale of text documents has been inflating in an exaggerated speed. That contributes to a high requirement of a method for Internet users to search for the information they need even in an uncommon and specific field, such as the movies.

Nowadays many movie spectators tend to search for their interested movies in some online movie collection websites. The key words they use could mostly be the movie genres like romance, action, science fiction, comedy and so on. Moreover, each movie could be tagged with several different labels, and it might contain the elements of different genres to different degrees. For example, the movie 'Pacific Rim' belongs to action, sci-fi along with a little romance.

The text clustering algorithms have been widely applied in classifying the text documents such as browsing the posts in a social media site which are relevant to some specific event. As for the movies, the subtitle documents could be the materials for the algorithms to extract the features and tag the movies with different genre labels. Due to the characteristics of subtitle document classification, the classification methods are quite essential. The choice of method would determine whether the users could find their favored movies to some extent.

DOI: 10.54254/2755-2721/5/20230669

The text classification methods could be categorized into two algorithms: unsupervised machine learning and supervised machine learning. On the one hand, supervised learning uses common methods such as Support Vector Machine (SVM) and Decision Tree. On the other hand, for unsupervised methods, K Nearest Neighbors (KNN), Agglomerative clustering and K-means clustering are widely studied. To evaluate the advantages and disadvantages of those algorithms, their effects are observed under a multilabel classification mission to simulate their application on movie websites.

To extract the features from the subtitle documents, there are two common algorithms to achieve that goal [1]. For instance, Term Frequency and Inverse Document Frequency (TF-IDF) method and Bag of Word (BOW). Several dependent variables such as precision, recall and F score could always be deployed to evaluate various algorithms' performances.

Many prior studies address synopsis texts, trailer contents such as video and so on. There still are more strategies which can be applied to movie reviewing and comparing the domains of them, for instance, semantic orientation, which expresses whether an opinion is positive, negative, or neutral. In addition, other classifiers and other algorithms which could have profound influence on identifying movie genres are still unrevealing.

The following parts of the paper will be organized like this: In section 2, the unsupervised algorithms from previous work would be introduced and analyzed, including its advantages and disadvantages and current state of research. Section 3 introduced the Parameter Optimized Hybrid Classifier and the classification process. And in section 4, the classification result of each algorithm is discussed, which combined with the classifier on subtitle documentation data, demonstrating better classification results compared to traditional methods.

2. Description of Unsupervised Machine Learning Approach

2.1. Pre-processing

Before applying machine learning algorithms to classify movie subtitle documents, their data are pre-processed at first. Salton establishes a model of vector space, which is a normal model of text document representation [2]. Their data pre-processing can be concluded in 5 steps:

A. Fetch 500 entries of data from YIFY movie subtitle website, those files are originally .srt format when downloaded, then is modified into .txt format [3].

B. To get pure movie subtitle datasets, they decide to discard non-alpha characters and markup language tags, namely standard generalized markup language (SGML) and HyperText Markup Language (HTML). Furthermore, they process every subtitle document so that all characters are transformed into lower cases.

C. A non-negligible part to reduce their work's dimensionality is stemming, which reduced various forms of a word into a normal form, including plural form, adverb, the present continuous, past tense and so on [4]. The figure 1 illustrates an example of a text document after stemming process.

D. Stop-words such as prepositions and pronouns are removed since these words are less relevant to content, in consequence, dimensionality of the feature space would decrease dramatically.

E. Researchers combine TF-IDF and BOW [1], two algorithms of term weighting together. The former one is widely applied term weighting method; they increase term's weights according to its occurrence in a single document while reduced its weight when appeared in most subtitle documents. Moreover, the latter algorithm measures a term's weight by organizing document as an unordered collection of words, in which under the circumstances of not taken grammar and sequence of word into consideration.

$$w_i = tf_i \cdot log \frac{N}{N_i} \tag{1}$$

DOI: 10.54254/2755-2721/5/20230669

Figure 1. Figure of the Residual structure.

(Photo credit: Original)

2.2. Methodology

Firstly, the implemented model was grouped similar movie subtitle data in clusters, after pre-processed these data, term-document matrix was generated based on term's weights, for the purpose of ordering subtitle documents. After that, by applying TF-IDF and BOW algorithms, the input vectors were produced. In addition, the performance was returned after used K-means, bisecting K-means and Agglomerative Clustering technique. Lastly, to achieve a better result, they finally measured the performance through analysing cluster, along with the calculation of the cluster and centroid similarity. More accurately, Hartigan discussed that K-Means Clustering refers to the mean of documents was supposed to be the centroid of that cluster. During their experiment, they found out that online K-Means was way more powerful than the batch one. At beginning, initial centroids could randomly choose subtitle documents from corpus, then the iteration will begin to assign subtitle documents to its adjacent centroid, and it stops when there are no more subtitle documents which need to be relocated.

Following method is the bisecting K-Means Clustering, which employed repeatedly the basic K-Means algorithms to obtain a cluster of hierarchy. Through conducting two research under the clusters with either minimum or maximum overall similarity, they discovered two experiments had close performance, while they chose the larger one to display the result.

Last algorithm is the Agglomerative clustering algorithm, it basically merges most relevant separate clusters from each document, then ends iteration when met the stop criteria. Figure 2 described the definition of inter-cluster similarity which classify from single ties, complete links as well as decent link.

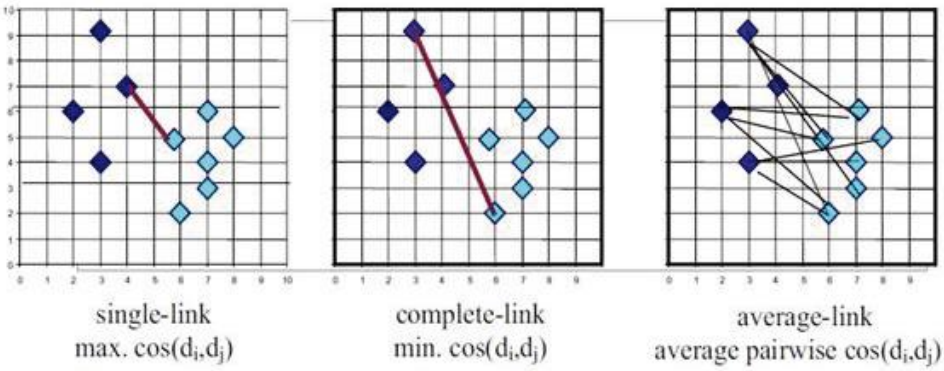

Figure 2. Inter cluster similarity [5].

DOI: 10.54254/2755-2721/5/20230669

2.3. Model Evaluation

Internal quantity measurement and external quality measurement are two respectively two methods evaluating cluster quality in unsupervised machine learning [6]. On the one hand, the measurement of external quality requires purity as a quality standard based on labelled movie title documents, meanwhile, F-measure and entropy are two regularly quantifications of external quality in data mining [7]. On the other hand, the measurement of internal quality only makes use of internal resources and calculate overall similarity for internal quality.

3. Description of Parameter Optimized Hybrid Classifier

3.1. Pre-processing

Before classifying movie subtitles, the initial step would be using a feature vector to replace the dataset. The pre-processing is consisting of 5 steps- text filtering, tokenization, remove stop words, stemming, and N-Gram [8].

1. Movie subtitles are consisting of various texts such as numbers, non-English texts, and tags. The method for normalizing movie documents is to use regular expression generated by python [8].
2. Break down movie subtitles by using tokenization. In this way, classifying subtitles into their appropriate genre is much easier [9].
3. Identify stop words based on pre-defined stop word lists. Then remove these stop words because they are barely meaningful to movie contexts [10].
4. Remove affixes from movie subtitle texts. The dimensionality of the classification system will be reduced a lot by applying this step [11]. This method can save lots of time and increase the accuracy of the classification process.
5. The method of dividing the dataset into N parts is known as the N-Gram. The movie subtitle texts will be separated into N character sequences [12].

3.2. Methodology

After the pre-processing, keywords will be extracted from movie subtitle texts. Then, by applying these two algorithms – TF-IDF and BOW, the dataset will be presented as feature vectors. However, TF-IDF and BOW are performing differently in classifying process of movie subtitle texts. TF-IDF can measure the occurring frequency of one term in one document. While BOW will use the occurring frequency of one term instead of considering the occurring sequence of each term [8].

Several classifiers are introduced here for generalizing various movie subtitle texts. The first classifier is Support Vector Machine (SVM). This model can classify both linear and nonlinear regression. The second classifier is K Nearest Neighbour (KNN). This model can calculate the proportion of the k nearest neighbour in a feature space that significantly belong to a given category [13]. The third classifier is Decision Tree (DT). This model can learn basic decision-making principles to distinguish between nodes and features in order to get the target value [14]. The final classifier is Parameter Optimized Hybrid Classification Approach (POHC).

4. Comparison

The figure 3describes the performance of two learning method of unsupervised learning. We analysed advantages and disadvantages of each algorithm under supervised learning and unsupervised learning respectively.

4.1. Unsupervised learning

In unsupervised learning, considering overall similarity, purity, entropy and F-measure, single-link algorithms present a poor outcome for TF-IDF and BOW methods. A main limitation of single-link algorithms is any two subtitle documents which not belong to same class could be nearest neighbours because they may have same weighting terms. Moreover, complete-link assumes every subtitle text are similar in the cluster, however, the dimensional diversity does not take this assumption into account, for

DOI: 10.54254/2755-2721/5/20230669

instance, the algorithm has not evaluated synonyms, hyponyms and so on. Additionally, average-link algorithm shows a better result among all agglomerative clustering algorithms. To conclude, when assessing entropy, purity and f-measure under TF-IDF method, bisecting k-means and k-means outperform than the single-link, average-link and complete-link clustering algorithms. In the meantime,

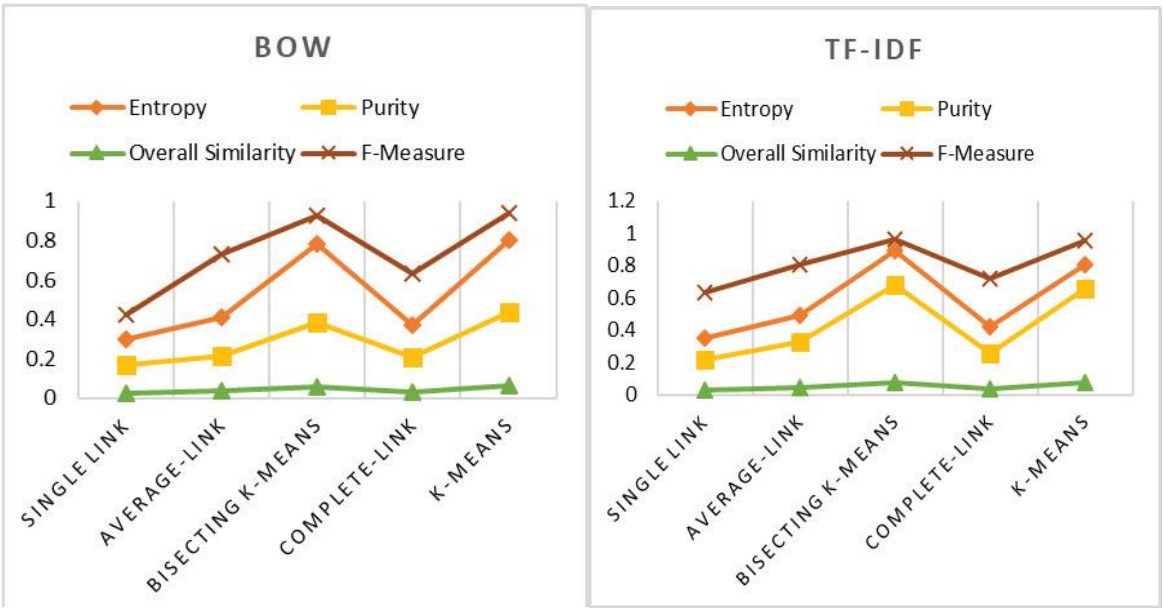

Figure 3. Performance for BOW and TF-IDF method under unsupervised learning [5].

bisecting k-means and k-means algorithms have beat other clustering algorithms when evaluating entropy and f-measure for BOW methods. However, all algorithms have similar overall similarity for two representation methods.

4.2. Supervised learning

In supervised learning, the POHC gets highest points for precision, recall, and F1-score which means it performs best compared to other three classifiers in both BOW and TF-IDF algorithms. In TF-IDF algorithm, POHC performs better than it performs in BOW. The DT classifier does not perform well in either BOW or TF-IDF algorithms compared to other classifiers. DT performs better in TF-IDF than it does in BOW. For SVM and KNN classifiers, they are also performing better than they do in BOW. The phenomenon may occur because structures will be affected significantly even though only a few data changed in BOW (Table 1).

Table 1. Performance for classifiers in BOW and TF-IDF [8].

Algorithm	Feature Extraction	Precision	Recall	F1-Score
BOW	KNN	0.86	0.83	0.83
	SVM	0.84	0.80	0.83
	DT	0.80	0.79	0.81
	POHC	0.92	0.90	0.91
TF-IDF	KNN	0.90	0.87	0.89
	SVM	0.88	0.83	0.87
	DT	0.85	0.83	0.84
	POHC	0.97	0.94	0.96

DOI: 10.54254/2755-2721/5/20230669

5. Conclusion

With the rapid population of computer devices and the Internet, the requirement of online keyword search has been experiencing a continuing increase, which leads to the development of classification technique based on enormous texts. In this study, the authors synthesize several existing methods and compare them under different applications. TF-IDF and BOW are the two methods to generate the features of the input texts. For the unsupervised methods, the author chooses the prevailing one K-Means, its expanding version Bisecting K-Means, and its simple version agglomerative clustering algorithms. By calculating the indicators: entropy, purity and F-score, although the average-link performs the best among clustering algorithms, it totally falls behind K-means and Bisecting K-Means. At the same time, when considering K-means and its expanding version, they show almost equally effect under those assessing indicators. These results are generally equal under TF-IDF and BOW methods. For supervised classifiers, the author concludes the most practical methods, SVM, KNN, DT and POHC. With the features of Bi-GRAM and Tri-GRAM, Precision, recall and F1 are tested to compare their accuracy and stability. As was expected, POHC which combines SVM with DT performs the best and a high robust. As a popular method, KNN is competitive compared with single SVM and DT. And DT also provides acceptable basic result.

References

[1] Pimpalkar, A. P., and Raj, R. J. R 2020 Influence of pre-processing strategies on the performance of ML classifiers exploiting TF-IDF and BOW features. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 9, 49.

[2] A, Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, 7th Edition, 67-179.

[3] YIFY Subtitles for English Movies, Available at: https://yts-subs.com/, Accessed: 10 December 2020.

[4] Liu, L., Kang, J., Yu, J., and Wang, Z 2005 A comparative study on unsupervised feature selection methods for text clustering. In 2005 International Conference on Natural Language Processing and Knowledge Engineering, 597-601.

[5] Hasan, M. M., Dip, S. T., Kamruzzaman, T. M., Akter, S., and Salehin, I 2021 December. Movie Subtitle Document Classification Using Unsupervised Machine Learning Approach. In 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), 219-224.

[6] Rajawat A.S., et.al 2022 Efficient Deep Learning for Reforming Authentic Content Searching on Big Data. In: Bianchini M., Piuri V., Das S., Shaw R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, 218.

[7] F. Liu, G. Zhang and J. Lu, 2020 Heterogeneous Domain Adaptation: An Unsupervised Approach, IEEE Transactions on Neural Networks and Learning Systems, 31, 5588-5602.

[8] Hasan, M. M., Dip, S. T., Rahman, T., Akter, M. S., and Salehin, I 2021 Multilabel Movie Genre Classification from Movie Subtitle: Parameter Optimized Hybrid Classifier. In 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 1-6.

[9] S. C. Dharmadhikari, M. Ingle and P. Kulkarni 2011 Empirical Studies on Machine Learning Based Text Classification Algorithms, Advanced Computing: An International Journal (ACIJ), 2, No.6.

[10] Y. F. Huang and S. H. Wang 2012 Movie Genre Classification Using SVM with Audio and Video Features, Lecture Notes in Computer Science, 1–10.

[11] Sentiment Symposium Tutorial, Available: http://sentiment.christopherpotts.net/codedata/happyfuntokenizing.py, Accessed: 03 January 2020.

[12] Bitbucket, Available at: https://bitbucket.org/jaganadhg/twittertokenize/src/, Accessed: 12 March 2020.

[13] F. Sebastiani 2002 Machine learning in automated text categorization, ACM Computing Surveys,

DOI: 10.54254/2755-2721/5/20230669

34, 1–47.
[14] C. Moral, A. D. Antonio, R. Imbert and J. Ramirez 2014 A survey of stemming algorithms in information retrieval, Politecnica de Madrid, 19.