Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 04, April -2018

A REVIEW OF FEATURE EXTRACTION METHODS FOR TEXT CLASSIFICATION

Resham N. Waykole[1], Anuradha D. Thakare[2]

[1]Department of Computer Engineering, Pimpri Chinchwad College of Engineering
[2]Department of Computer Engineering, Pimpri Chinchwad College of Engineering

Abstract —*Natural Language Processing (NLP) and Machine Learning concepts are acclaimed in today's digitalization of data. Over the time, value of the data keeps changing and it is important to tackle that value for performing in depth research in various domains. Over the past decade, natural language processing has gained much importance because it reveals a lot of unseen information in the texts. It is difficult to discover the information of interest from a huge volume of the text data. Thus, information extraction based on computational text processing is necessary. For many of information management goals, the task of recognising phrases and words in free text which falls under particular classes of interest is an important first step. It is crucial to manage huge amount of text being generated dramatically. The text can be for example clinical and biomedical text. Features can be extracted for classification of the documents. Feature extraction is extracting an important subset of features from a data for improving the classification task. Correctly identifying the related features in a text is important. Therefore, applying and expanding NLP techniques can help to better understand and study the data. This paper aims at analysing the clinical literature for cancer. The feature extraction methods such as bag of words, tf-idf, word2vec are compared for clinical text analysis. The extracted features are evaluated against Logistic Regression and Random Forest Classifier.*

Keywords-*Natural Language Processing, Feature Extraction, Classification, Bag of Words, TF-IDF, Word2Vec, Logistic Regression, Random Forest Classifier.*

I. INTRODUCTION

Text data is most simplest form of data which is unstructured in nature. It is generated in huge amount in most scenarios. Humans can clearly perceive and process unstructured text data but it is difficult for machines to understand the same. This voluminous text data is a important source of knowledge and information. Therefore, to use this information extracted from text data effectively in variety of applications, methods and algorithms are needed. NLP has gained a great deal of attention in past few years because of the huge amount of text data gets generated in many forms such as social networks, patient records, news outlets, healthcare insurance data, etc. in a report generated by EMC. It is predicted that, by 2020, the volume of data will grow upto 40 zettabytes[4]. It is difficult for humans to go through all such text data and find the information of interest and to organize large amount of data. To enable the effective transformation and representation of such data, the process includes calculating the word frequencies from the document and in the entire collection of documents. Therefore, it is important to extract the needful information from the unstructured text data[6].

II. RELATED WORK

Extracting information from text helps in analyzing the text data for various applications, reports, clinical records, automated terminology management, research subject identification, data mining and studying effect of research on them, etc. Feature Extraction is vital technique in dimensionality reduction to extract the important features. Samina Khalid et. al [1] have reviewed some common feature selection and feature extraction methods. It is analyzed for determining the effectiveness of these techniques for achieving high performance of learning algorithms. Because this ultimately improves prediction accuracy of the classifier. They have also analyzed some widely used dimensionality reduction techniques for the strengths and weaknesses of the techniques. [2] Several basic text mining tasks and techniques such as text data pre-processing, clustering and classification are described. Also the text mining in healthcare and biomedical domains are briefly explained.

[3] proposed a term frequency (TF) with stemmer-based feature extraction algorithm and the performance of the algorithm is tested using various classifiers. The results shows that the proposed method outperforms other methods. [5][6][7][8] various feature selection methods such as document frequency, information gain etc and feature extraction techniques such as principal component analysis (PCA), latent semantic indexing (LSI), etc are discussed and classifiers used for classification of documents are discussed. [16] [19]several automatically extracted features are compared. The features are extracted for sentiment analysis of twitter.

The commonly used feature extraction method TF-IDF is used. The TF-IDF technique is improved for feature extraction for better accuracy[15][21]. Faheema AG et. al [14] have introduced an efficient technique which increases the accuracy by using bag of visual word representation as a feature selection method. The word2vec is another feature extraction

technique which is widely used. The [9] [10] have discussed the word2vec method. however, [9] have proposed a hybrid method to extract features from the data. The approach is proposed by using both LDA and Word2Vec. The method derives the relationships between topics and documents. It also combines the contextual relationships among the words. The results shows that features generated by this hybrid technique are useful for improving performance of a classification.

III. Data

The datasets is taken from kaggle. The dataset is related to cancer which contains the genes, genetic mutations caused by cancer and clinical text. The datasets are provided via two different files - training and test. One called as training/test_variants which has the information about the genetic mutations. The other training/test_text provides the clinical text which are clinical research papers related to cancer which are used by human experts used to classify the genetic mutations. Both files can be linked via the ID field. The training dataset contains 3000 instances.

IV. METHODS

Text feature extraction is the process of taking out a list of words from the text data and then transforming them into a feature set which is usable by a classifier. This work emphasizes on the review of available feature extraction methods. The following techniques can be used for extracting features from text data.

4.1. Bag of words

The bag of words is the most common and the simplest among all the other feature extraction methods; it forms a word presence feature set from all the words of an instance. It is known as a "bag" of words, since the method doesn't care about how many times a word occurs or the order of the words, all what matters is whether the word is present in a list of words. The features can be used in modelling with machine learning algorithms. This method is very flexible and simple. It is usually used for extracting features from text data in various ways. A bag of words is the presentation of text data. It specifies the frequency of words in the document. It includes: 1. A lexicon of known words 2. A frequency of the existence of those known words. The complexity of bag of words model is both in determining how to score the presence of familiar words and how to design the vocabulary of familiar words.

4.2. TF-IDF

A problem with bag of words approach is that the words with higher frequency becomes dominant in the data. These words may not provide much information for the model. And due to this problem domain specific words which does not have larger score may be discarded or ignored.
To resolve this problem, the frequency of the words is rescaled by considering how frequently the words occur in all the documents. Due to this, the scores for frequent words are also frequent among all the documents are reduced.
This way of scoring is known as Term Frequency – Inverse Document Frequency.
• Term Frequency (TF) is the frequency of the word in the current document.
• Inverse Document Frequency (IDF) is the score of the words among all the documents.
These scores can highlight the words that are unique that is the words that represent needful information in a specified document. Therefore the IDF of an infrequent term is high, and the IDF of a frequent term is low.

4.3. Word2Vec

Word2Vec is used to construct word embeddings. The models created by using word2vec are shallow meaning two-layer neural networks. Once trained, they reproduce semantic contexts of words. The model takes a huge corpus of text as an input. It then creates a vector space which is usually of hundreds of dimensions. Each distinctive word in the corpus is alloted with corresponding vector in the space. The words with common contexts are placed in near proximity in vector space. Word2vec can use one of the two architectures: continuous skip gram or continuous bag of words (CBOW). In the continuous skip gram, the current word is considered to predict the neighbouring window of context words. In this architecture the nearby context words are considered more heavily than words with distant context. In the continuous bag of words architecture, the sequence of context words does not impact the prediction as it is based on bag of words model.

V. EXPERIMENTATION RESULTS

In this work, for experimentation, the dataset is taken and text features are extracted using Bag of words, TF-IDF and Word2vec techniques. The extracted text features are then evaluated for text classification. The effectiveness of transformed free text is evaluated using logistic regression and random forest classifier with 3-fold stratified cross-validation.

5.1. Logistic Regression

Logistic regression is used classification problems. It is used to predict the group in which the current object under consideration belongs to. Classification is portioning the data into groups based on particular features. Most commonly used example of LR, Suppose a tumor needs to be classified as benign or malignant based on various features like location of the tumor, size of the tumore, etc. Logistic regression is named for the function called the logistic function. The logistic function is also called as the sigmoid function.

In LR, Y is the dependent variable which has G = 2 (usually) distinct values. It is reverted on a set of *p* independent variables *X1, X2, ... , Xp*. Suppose Y is a condition after surgery, absence or presence of a disease, or marital status. Because the names of these divisions are arbitrary, they are referred by successive numbers. That is, Y will have values 1, 2, ... G. Let

$$\mathrm{X} = (X_1, X_2, \cdots, X_p)$$

$$B_g = \begin{pmatrix} \beta_{g1} \\ \vdots \\ \beta_{gp} \end{pmatrix}$$

The LR model is given by,

$$\ln\left(\frac{p_g}{p_1}\right) = \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p$$

$$= \ln\left(\frac{P_g}{P_1}\right) + \mathrm{XB}_g$$

where, *Pg* is the probability of an individual having values *X1, X2, ..., Xp* in outcome *g*. That is,

$$Pg = Pr(Y=g|X)$$

The *P1,P2,...,PG* are the prior probabilities of outcome. If the prior probabilities are equal, then the term *ln(Pg / P1)* becomes zero. If the prior probabilities are not equal, the values of the intercepts in the LR equation are changed. In the logits of *p*, these equations are linear. But considering the probabilities, the equations are nonlinear. The nonlinear equations are

$$\mathrm{p_g} = \mathrm{Prob}(Y = g \,|\, \mathrm{X}) = \frac{e^{\mathrm{XB}_g}}{1 + e^{\mathrm{XB}_2} + e^{\mathrm{XB}_3} + \cdots + e^{\mathrm{XB}_G}}$$

These models are called as logistic regression.

5.2. Random Forest Classifier

The random decision forests or random forests are an object learning technique for used for classification. It creates a multitude of decision trees during training of the data. It then gives the output class that is the classification of individual trees. The decision tree algorithm sometimes overfits the training dataset. In such cases RF is used as the correction to it. It is a supervised classification algorithm. The RF classifier constructs a set of decision trees from randomly selected subset of training set. Then the votes from the different decision trees are aggregated to decide final class of the test object. This algorithm creates the forest with a number of trees. Generally, the more trees in the forest the more robust the forest looks like. This gives the high accuracy results.

The results of the experimentation are shown below:

Sr. No.	Feature Extraction Method	Classification Algorithm	Accuracy	Log Loss
1	Bag of Words	Logistic Regression	48%	1.65
	Bag of Words	Random Forest Classifier	50%	1.44
2	TF-IDF	Logistic Regression	46%	1.50
	TF-IDF	Random Forest Classifier	51%	1.35
3	Word2Vec	Logistic Regression	53%	1.32
	Word2Vec	Random Forest Classifier	57%	1.21

Table 1. Accuracy and Log Loss comparison for different methods

VI. CONCLUSION

Analyzing the text data is critical. This work is focused on applying common techniques such as bag of words, TF-IDF and word2vec to preprocess and vectorize free text and are evaluated its effectiveness by running them through vanilla Logistic Regression and very basic Random Forest classifier. The results showed that word2vec is the better method for feature extraction along with random forest classifier for text classification. For the better results other emerging feature extraction techniques such as doc2vec, GloVe can be used. There are also feature extraction tools such as cTAKES are specifically available for clinical text analysis.

REFERENCES

[1] Samina Khalid, Tehmina Khalil, Shamila Nasreen, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning", Science and Information Conference 2014, August 27-29, 2014
[2] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", arXiv:1707.02919 [cs.CL], July 2017
[3] S.Vidhya, D.Asir Antony Gnana Singh, E.Jebamalar Leavline, "Feature Extraction for Document Classification", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Special Issue 6, May 2015
[4] John Gantz and David Reinsel. "THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", Technical Report 1. IDC, 5 Speen Street, Framingham, MA 01701 USA, 2012
[5] Foram P. Shah, Vibha Patel, "A Review on Feature Selection and Feature Extraction for Text Classification", IEEE WiSPNET conference, 2016
[6] F. Beil, M. Ester, X. Xu, "Frequent term-based text clustering", Proc. of Int'l Conf. on knowledge Discovery and Data Mining KDD '02, pp. 436–442, 2002
[7] Dixa Saxena, S. K. Saritha, PhD, K. N. S. S. V. Prasad, "Survey Paper on Feature Extraction Methods in Text Categorization", International Journal of Computer Applications, May 2017
[8] J. J. Verbeek, "Supervised feature extraction for text categorization" Tenth Belgian- Dutch Conference on Machine Learning (Benelearn '00), Dec 2000
[9] Zhibo Wang, Long Ma, and Yanqing Zhang, "A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec", IEEE First International Conference on Data Science in Cyberspace, 2016
[10] Zhang W., Xu W., Chen G., Guo J. "A Feature Extraction Method Based on Word Embedding for Word Similarity Computing", 2014
[11] Zong, C., Nie, J.-Y., Zhao, D., Feng, Y. (Eds.), "Natural Language Processing and Chinese Computing" Communications in Computer and Information Science, vol 496. Springer, 2014
[12] Long Ma, Yanqing Zhang, "Using Word2Vec to process big text data", IEEE International Conference on Big Data (Big Data), 2015
[13] Bradford Heap, Michael Bain, Wayne Wobcke, Alfred Krzywicki, Susanne Schmeidl, "Word Vector Enrichment of Low Frequency Words in the Bag-of-Words Model for Short Text Multi-class Classification Problems", arXiv:1709.05778 [cs.CL], 2017
[14] Faheema AG and Subrata Rakshit, "Feature Selection using Bag-Of-Visual-Words Representation", IEEE 2nd International Advance Computing Conference (IACC), 2010
[15] Leena H. Patil, Mohammed Atique, "A Novel Approach for Feature Selection Method TF- IDF in Document Clustering", IEEE 3rd International Advance Computing Conference (IACC), 2013
[16] Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, Fazal Masud Kundi, "A Review of Feature Extraction in Sentiment Analysis", Journal of Basic and Applied Scientific Research, 2014
[17] N. Elavarasan, Dr. K.Mani, "A Survey on Feature Extraction Techniques" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2015
[18] Vineet John, "A Survey of Neural Network Techniques for Feature Extraction from Text", arXiv:1704.08531 [cs.CL], 2017
[19] Nicolas Tsapatsoulis, Constantinos Djouvas, "Feature Extraction for Tweet Classification: Do the Humans Perform Better?", 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), 2017
[20] Masoumeh Zareapoor, Seeja K. R, "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection", I.J. Information Engineering and Electronic Business, 2, 60-65, 2015
[21] LI-PING JING, HOU-KUAN HUANG, HONG-BO SHI, "IMPROVED FEATURE SELECTION APPROACH TFIDF IN TEXT MINING", Proceedings of the First International Conference on Machine Learning and Cybernetics,Beijing, 4-5 November 2002
[22] Yun Fu, Shuicheng Yan, and Thomas S. Huang, "Classification and Feature Extraction by Simplexization", IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 3, NO. 1, MARCH 2008

