JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 18, 2020

Taxonomy on Text classification Methods and Viewpoint

Sowmya M S[1], Dinesh R[2]

Research Scholar, Jain University[1], Visiting Professor, School of Engineering and Technology, Jain University[2]
mssowmya.sbmjce@gmail.com[1], dr.dineshr@gmail.com[2]

Received: 28 May 2020 Revised and Accepted: 06 July 2020

ABSTRACT: Now a days, there is tremendous progress in the quantity of composite documents and transcripts that need an in-depth knowledge of machine learning concepts to precisely classify texts in real time. Though there are many different methods that has given good results in literature, but they are dependent on interpreting difficult models and non-linear relationships within data. Even with all these, selecting relevant structures and practices for text classification is tricky task for many of us. Here we provide, a brief discussion on text classification procedures, highlighting the limits of each of its method in real-world applications.

KEYWORDS: Feature extractions, Dimensionality reduction, Classification techniques.

I. INTRODUCTION

Text classification hitches is extensively considered and instructed for several applications [1–9] from many ago. Specially with new innovations in text mining and Natural Language Processing (NLP), many scholars are currently concerned in developing applications that influence text classification methods. Most text classification and document categorization systems can be decomposed into the subsequent four stages: Extraction of feature, Reduction of dimension, selection of classifier, and estimations.

Considering documents formed of sentences, words and texts. Each segment is named by means of a class rate from a set of k dissimilar separate rate indices [1,8]. A structured set used for training purpose is called as Feature Extraction. The dimensionality reduction is used to reduce the features by bringing feature space in other-dimensional space. Text classification levels can be in document level, paragraph level, sentence level or sub-sentence-level. Choosing the best algorithm for document classification be a need always in real-time.

A. Feature Extraction

Most of the documents and texts are unstructured in terms sets of data. These are converted to structured feature using mathematical modeling. The input data is cleaned to remove unwanted characters and words then feature extraction methods are applied to it. Few methods of extraction of feature includes, Word2Vec [11], Global Vectors for Word Representation (GloVe) [12], Term Frequency (TF) [10] and combination of TF_IDF(Term Frequency-Inverse Document Frequency).

B. Dimensionality Reduction

Many times, data sets of document or text comprise of numerous inimitable words, pre-processing step for data may lag in time and complexity because of this. Though the inexpensive algorithms work in this context, but do not perform as expected. Dimensionality reduction (DR) is used to decrease the memory and time intricacy for many uses and pre-processing becomes more efficient. Some of the DR techniques are Non-negative Matrix Factorization (NMF), Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA).

C. Classification approaches

Choosing best classifier is main step of the text classification. It is important to know the text classification algorithms, some are like Rocchio classification, bagging and boosting are the ensemble-based learning techniques, logistic regression (LR), The Naïve Bayes Classifier, Non-parametric techniques are Support Vector Machine (SVM) [21,22], k-nearest neighbor (KNN) [13], Tree-based classifiers like decision tree.

D. Assessment

Assessment is a stage to check how the model performs. Many methods are available for this to assess supervised techniques. These won't work for unbalanced data sets [14]. Some of the Assessment methods are Fb Score, area under the ROC (Receiver Operating Characteristics) with the Area Under the curve (AUC), receiver operating characteristics, Matthews Correlation Coefficient.

II. PREPROCESSING

Cleaning the text datasets is all about pre-processing and it is very important step in text categorization applications. Feature extraction methods for texts can be Embedding word techniques and Word weighting technique. Some of the other tasks in preprocessing are,

- Removing stopwords, misspelling, slang etc. are the objective of text cleaning.
- Tokenization is one of the pre-processing methods, it segments the input text into words, phrases, symbols called tokens [15,16]. Its goal is to find the words in a sentence.
- Stop words are the words that do not contribute to classification of text and document.
- It is good idea to reduce every letter to lower case and avoid capitalized words in identifying good feature or term.
- As per the literature, it is suggested to convert the slang and abbreviation into formal language to avoid incompleteness and ambiguity in words.
- Unnecessary characters like punctuation and special symbols need to be removed. It is called Noise Removal.
- Spelling correction methods like context-sensitive and Hashing-based can be used for spelling correction and typo mistakes of users.
- Stemming is a way to know the root words by removing ing, es, ed, er etc. When extracting feature the words with these may not be contributing always.
- Lemmatization is an improved form of stemming as needed for many applications. The result of lemmatization is more meaningful compared to stemming which may end up giving words that are not in dictionary at all. It is a process of NLP which tries to replace a word suffix with other possible suffixes of a word entirely to get a base form of word called as lemma [17-19].

A. Word Representation

It is also important to keep track of syntax and semantics between words. Though there are different techniques available in this category, all has its own pros and limitations. Some of the techniques that works for syntax in statements are N-Gram, basically Bag-of-words (BOW) like 1-gram, 2-gram and N gram). N-gram means the set of n-words that comes in the same order in the input text/document.

B. Technique of Weighting Words

Basically, it is about finding the frequency of the word/term t_i. One such method is TF (Term frequency). Other methods that uses outcomes of TF will utilize the frequency of the word as a Yes/No scaled weighting. Here, each document is converted to a vector with occurrence count of the words in respective document. But the technique has restrictions from the dominating languages to such representations. Unlike to TF, the technique that assigns a highest weight to high or low frequency term in the respective document. The blend of Term Frequency-Inverse document frequency (TF-IDF) is another needed approach as per the application need. Depending on the accuracy expected, only TF or IDF or combination of TF-IDF techniques can be applied.

C. Embedding terms/words

Another technique used is word embedding, it maps each word to a vector of real numbers of dimension N. Some of the approaches that comes here are Word2Vec, GloVe, and FastText. Embedding methods are used to translate unigrams as a meaningful ML algorithm.

III. DIMENSIONALITY REDUCTION

Based on vector models, the Text sequences will have more features. It makes these methods to consume more memory and it is very expensive. To address this issue, many experts in literature make use of dimensionality reduction method to decrease the dimensions of feature space. Some of the main algorithms to reduce the dimension reduction algorithms are Non-Negative Matrix Factorization (NMF), Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA).

Principal component analysis (PCA) is techniques for multivariate analysis and dimension reduction. A subspace is identified in this method to find where data lies approximately lies [20]. PCA is used in noise reduction algorithms and to avoid over-fitting problem [21]. Independent component analysis is another technique that is a method of statistical modeling, in this, the data is represented as a linear transformation after detecting it [22].

To reduce the dimension and to classify the data, Linear Discriminant Analysis (LDA) technique is very helpful [12,23]. It evaluates the performance on randomly generated test data when there is inequal within-class occurrences. LDA approaches comes under class-dependent and class-independent transformation. Non-negative matrix approximation is used for very high dimensional text and sequences [24, 25]. It is one of best dimension reduction method in terms of the results it gives.

IV. TEXT CLASSIFICATION METHODS

From literature, text classification algorithms are Rocchio algorithm, bagging and boosting are learning algorithms in ensemble category. Scientific community still make use of traditional methods like k-nearest neighbor, logistic regression and Naive Bayes. Kernel SVM or Support vector machines (SVMs), is mostly used as technique for classification. For classification of document the algorithms like Tree-based gives accurate and fast results. Some of tree-based methods are random forest and decision tree.

The Rocchio algorithm is developed based on the Vector Space Model. The algorithm works on the assumption that user has the conception of being a document is relevant or not. Instead of Boolean features, this algorithm makes use weights generated by TF-IDF weights for W_i, a useful word. It builds a vector as an example for Ci, the class, over the training document, an average vector is C_i, then it allocates t_i, the test document with the highest resemblance class between the prototype vectors and the test document [1, 26]. Its algorithms limitations are that the handler can only fetch a few relevant documents and it shows results using semantics.

Bagging and boosting are two ensemble techniques. These algorithms generate class predictions by combining the predictions of many other sub-classifiers. Bagging uses statistical analyses to improve the approximation of one by combining the approximations of many. Boosting approach generates consecutive base classifiers that are told to place greater importance on the mis-classified examples from the training data. Like bagging, the results of boosting are combined to produce a meta-prediction. The limitations of these techniques are computational complexity and loss of interpretability [27], it can’t find the feature importance.

Logistic regression is used for predicting binary classes. It uses statistical method. It calculates the probability of an event occurrence. It gives constant output. Detecting cancer patient is one of the examples to this. It is assessed using Maximum Likelihood Estimation (MLE) method. It determines the most likely parameters to provide the data that is observed. It assigns variance and mean as most likely parameters that determines the values of specific parametric for a model given. In normal distribution, these are used for forecasting the needed data. The limitation is that it works well for predicting categorical outcomes.

Naïve Bayes Classification is based on Bayes rule, it can provide accurate results without much training data. Its representation should be very simple. Suppose our input is review of a movie in text form. the check is to test whether the review class is positive or negative [5]. So, having a bag of negative words and a bag of positive words, the frequency of the word appears in a document can be counted to place the document in the category of the class either positive or negative. Topic classification uses this to classify academic documents into different topics like computer science, biology, mathematics.

K Nearest Neighbor (KNN) algorithm, before predicting the closest matches using label, it finds the K nearest matches in training data to classify [9]. To find the closest match the help of Euclidean distance is good. KNN works by first finding the k nearest neighbor for the given test document x, and based on the class of k neighbors, it scores the candidates in the category. Neighbor’s document di and x’s can be the neighbor documents category score. In case of Multiple KNN documents, with respect to the test document x, similarity score of class k will be the summation of these scores. The highest scored candidate to the class from the test document x is assigned by the algorithm after the score values are sorted [1, 28]. KNN performance is based on meaningful distance function that is found. KNN has limitation for large search problems in terms of data storage. These all makes KNN a algorithm that is data dependent.

One of the supervised models of machine learning is Support vector machine (SVM). It uses classification algorithms to solve the two-group classification problems [3]. It is an algorithm to determine the decision boundary that is simplest between vectors. It belongs to a vector and category given. These are often applied to vectors that encode any relevant data. It recommends to leverage the facility of SVM text classification for the texts that need to be transformed into vectors.

In Decision tree, the thought is to make a tree supported the attribute for categorized data points, but the most challenge is chosen which feature might be in parents' level and which one should be in child level. to unravel this problem, for feature selection in tree, De Mántaras [29] presented statistical modeling [7]. The tree based on decision may be a in no time algorithm for both learning and prediction and extremely sensitive to within data minor perturbations [30], and may be easily overfit [31]. Validation methods and pruning are used to negate these effects [30]. Out-of-sample prediction is the limitation of this model.[32].

V. LIMITATIONS OF TEXT CLASSIFICATION ALGORITHMS

The Rocchio algorithm [2,26] has limitation to user, i.e, using this model, they can only retrieve relevant couple of documents. Loss of interpretability and computational complexity are drawbacks of Boosting and bagging [27] methods. For the categorical outcome's prediction, the LR method is good. But this type of prediction needs, the data be independent.

Naive Bayes algorithm [28] has limitation in the form of the information distribution by making an assumption robustly. It also suffers from data scarcity but it handles multi-class cases naturally. Also, KNN has limitation for finding closest neighbors by data storage in case of search problems that are massive. KNN performance depends on distance function seeking, so it comes in the category of the data-dependent algorithm. SVM method is not much transparent in results with high dimensions. It also suffers from variable financial ratios rate. It gets easily overfit and has limitations of out-of-sample prediction.

Considering the limitations of the discussed best classification techniques, we will be demonstrating the improved results after applying few of these techniques into our application and deliberating about the method which gave us good result in out next paper.

VI. CONCLUSION

One among the foremost crucial problems in machine learning is the classification task. Text classification becomes an important issue in supervised machine learning algorithms, when there is proliferate text and document data sets. To have a far better document categorization system puts these algorithms in discriminating state. But these prevailing algorithms work more proficiently if we've a good knowledge about extraction of features and way to gauge them properly. Here, we are presenting a quick outline on techniques of classifying texts, pre-processing process and a brief about evaluation. Restrictions of existing classification techniques are highlighted. The most challenging task is in opting an effective classification method and is in knowing the common and uniqueness of techniques in several pipeline steps.

VII. REFERENCES

[1] Kamran Kowsari, Kiana Jafari Meimandi, "Text Classification Algorithms: A Survey", Information 2019, 10, 150; doi:10.3390/info10040150 , PP 1-68.

[2] Jiang, M.; Liang, Y.; Feng, X.; Fan, X.; Pei, Z.; Xue, Y.; Guan, R. Text classification based on deep belief network and softmax regression. Neural Comput. Appl. 2018, 29, 61–70.

[3] Kowsari, K.; Brown, D.E.; Heidarysafa, M.; Jafari Meimandi, K.; Gerber, M.S.; Barnes, L.E. HDLTex: Hierarchical Deep Learning for Text Classification. Machine Learning and Applications (ICMLA). In Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017.

[4] McCallum, A.; Nigam, K. A comparison of event models for naive bayes text classification. In Proceedings of the AAAI-98Workshop on Learning for Text Categorization, Madison, WI, USA, 26–27 July 1998; Volume 752, pp. 41–48.

[5] Kowsari, K.; Heidarysafa, M.; Brown, D.E.; Jafari Meimandi, K.; Barnes, L.E. RMDL: Random Multimodel Deep Learning for Classification. In Proceedings of the 2018 International Conference on Information System and Data Mining, Lakeland, FL, USA, 9–11 April 2018; doi:10.1145/3206098.3206111.

[6] Heidarysafa, M.; Kowsari, K.; Brown, D.E.; Jafari Meimandi, K.; Barnes, L.E. An Improvement of Data Classification Using Random Multimodel Deep Learning (RMDL). IJMLC 2018, 8, 298–310.

[7] Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent Convolutional Neural Networks for Text Classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 333, pp. 2267–2273.

[8] Aggarwal, C.C.; Zhai, C. A survey of text classification algorithms. In Mining Text Data; Springer: Berlin/Heidelberg, Germany, 2012; pp. 163–222.

[9] Aggarwal, C.C.; Zhai, C.X. Mining Text Data; Springer: Berlin/Heidelberg, Germany, 2012.

[10] Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 1988, Volume 24, 513–523. [CrossRef]

[11] Goldberg, Y.; Levy, O.Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv 2014, arXiv:1402.3722.

[12] Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Volume 14, pp. 1532–1543.

[13] Li, L.;Weinberg, C.R.; Darden, T.A.; Pedersen, L.G. Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics **2001**, 17, 1131–1142. [CrossRef]

[14] Huang, J.; Ling, C.X. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. **2005**, 17, 299–310. [CrossRef]

[15] Gupta, G.; Malhotra, S. Text Document Tokenization for Word Frequency Count using Rapid Miner (Taking Resume as an Example). Int. J. Comput. Appl. **2015**, 975, 8887.

[16] Verma, T.; Renu, R.; Gaur, D. Tokenization and filtering process in RapidMiner. Int. J. Appl. Inf. Syst. **2014**, 7, 16–18. [CrossRef]

[17] Sampson, G. The'Language Instinct'Debate: Revised Edition; A&C Black: London, UK, 2005.

[18] Plisson, J.; Lavrac, N.; Mladeni´c, D. A rule based approach to word lemmatization. In Proceedings of the 7th International MultiConference Information Society IS 2004, Ljubljana, Slovenia, 13–14 October 2004.

[19] Korenius, T.; Laurikkala, J.; Järvelin, K.; Juhola, M. Stemming and lemmatization in the clustering of finnish text documents. In Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management,Washington, DC, USA, 8–13 November 2004; pp. 625–633.

[20] Abdi, H.;Williams, L.J. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. **2010**, 2, 433–459. [CrossRef]

[21] Ng, A. Principal components analysis. Generative Algorithms, Regularization and Model Selection. CS **2015**, 229, 71.

[22] Hyvärinen, A.; Hoyer, P.O.; Inki, M. Topographic independent component analysis. Neural Comput. **2001**, 13, 1527–1558. [CrossRef] [PubMed]

[23] Sugiyama, M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. J. Mach. Learn. Res. **2007**, 8, 1027–1061.

[24] Pauca, V.P.; Shahnaz, F.; Berry, M.W.; Plemmons, R.J. Text mining using non-negative matrix factorizations. In Proceedings of the 2004 SIAM International Conference on Data Mining, Lake Buena Vista, FL, USA, 22–24 April 2004; pp. 452–456.

[25] Tsuge, S.; Shishibori, M.; Kuroiwa, S.; Kita, K. Dimensionality reduction using non-negative matrix factorization for information retrieval. In Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics, Tucson, AZ, USA, 7–10 October 2001; Volume 2, pp. 960–965.

[26] Korde, V.; Mahender, C.N. Text classification and classifiers: A survey. Int. J. Artif. Intell. Appl. **2012**, 3, 85.

[27] Geurts, P. Some enhancements of decision tree bagging. In European Conference on Principles of Data Mining and Knowledge Discovery; Springer: Berlin/Heidelberg, Germany, 2000; pp. 136–147.

[28] Jiang, S.; Pang, G.;Wu, M.; Kuang, L. An improved K-nearest-neighbor algorithm for text categorization. Expert Syst. Appl. **2012**, 39, 1503–1509. [CrossRef]

[29] De Mántaras, R.L. A distance-based attribute selection measure for decision tree induction. Mach. Learn. 1991, 6, 81–92. [CrossRef]

[30] Giovanelli, C.; Liu, X.; Sierla, S.; Vyatkin, V.; Ichise, R. Towards an aggregator that exploits big data to bid on frequency containment reserve market. In Proceedings of the 43rd Annual Conference of the IEEE Industrial Electronics Society (IECON 2017), Beijing, China, 29 October–1 November 2017; pp. 7514–7519.

[31] Quinlan, J.R. Simplifying decision trees. Int. J. Man-Mach. Stud. 1987, 27, 221–234. [CrossRef]

[32] Jasim, D.S. Data Mining Approach and Its Application to Dresses Sales Recommendation. Available online: https://www.researchgate.net/profile/Dalia_Jasim/publication/293464737_main_steps_for_doing_data_mining_project_using_weka/links/56b8782008ae44bb330d2583/main-steps-for-doing-datamining-project-using-weka.pdf (accessed on 23 April 2019).