Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

‘Bag of Words’ to ‘Bag of Concepts’: Improving Text Categorization using SVM

Rahul V. Mohare, Satyajit S. Uparkar, Pravin Y. Karmore, Vishnu Vardhan Budati

Shri Ramdeobaba College of Engineering and Management, India
Email: moharerv@rknec.edu

This research paper tries to brief about existing research trends in improving text categorization, few research work of vital importance are implemented in recent times, and discovering some major open issues about text categorization using the famous Support Vector Machine (SVM) Algorithm. The design approach renders to an experimental study consisting an exploration of concepts, definitions and scoping of previous literature along with analyzing comparing the text categorization using Bag of Words (BoW) and Bag of Concepts (BoC) approaches. The major findings include, the effectiveness borne by the applied research techniques for categorizing text using SVM and comparing the results with respect to BoW and BoC approach. This will further direct the researchers to identify and proceed their research work in right direction bridging the research gap. In addition, this research work will also guide researchers to go beyond ‘Bag of Words’ and start using/Developing the tools that emphasize on ‘Bag of Concepts’.

Keywords: Bag of Words, Bag of Concepts, Text Categorization, Random Indexing, Confusion Matrix.

1. Introduction

Text categorization refers assigning text to any one of the predefined set of categories. Several decisive factors play a vital role for this categorization [1]. One of the most famous and widely used approach for the text categorization is known as Bag-of-Words (BoW) [2][3]. In Bag of Words, text is usually presented as a vector of words weights, but ignores the semantic or conceptual information.[4]. The best representation of Bag of Words is a word cloud.

The Bag of Words (BoW) approach is a common method used in natural language processing for representing text data as a collection of word occurrences [5]. However, this approach has limitations in capturing the meaning of words and their relationships [6], which can impact the performance of text classification algorithms such as Support Vector Machines (SVM) [7][8]. To address this limitation, researchers have proposed the 'Bag of Concepts' (BoC) approach.

In the BoC approach, text is represented as a collection of concepts rather than words [9]. A concept is a group of semantically related words that are considered to have similar meaning. For example, the concept of 'food' might include words such as 'meal', 'snack', 'dinner', 'lunch', 'breakfast', etc. To generate a BoC representation of text, the first step is to identify concepts using techniques such as Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA) [2].

Once the concepts have been identified, the text is represented as a vector of concept occurrences, similar to the BoW approach [10]. This vector can then be used as input to an SVM classifier. The SVM algorithm learns a hyperplane that separates the different classes of text based on the vector representation of the input data [11].

Compared to the BoW approach, the BoC approach has the advantage of capturing the semantic relationships between words [7], which can improve the performance of text classification algorithms. However, the identification of concepts can be a challenging task, and the performance of the BoC approach can depend on the quality of the concept identification method used [6].

From past few years researchers are attempting different experiments for text categorization. One such successful method is use of n-grams or phrases [12]. Next attempt worth to mention is supplementing BoW approach with the latent dimensions or synonym clusters [13]. Performance wise these approaches were not able to perform well, computing complexities and expenses wise they failed to become popular.

The standard BoW approach will generate the words along with the frequency. Researchers are more concerned about the concept-based representations [4]. Going beyond the word count may give a better picture regarding the text and may help the researcher to categorized the text.

Authors honest intention is to investigate, whether the concept-based scheme can supplement the BoW representations, If yes, then, can we identify some specific categories for text categorization for which the concept-based representation would be most appropriate, instead of word based approach [14].

Author had made an attempt for proposing a new concept based representation of the documents. The proposed method is scalable and requires no exterior resources. We modeled the concept-based representation and had used this output as an input for Support Vector Machine (SVM) classifier [15]. The output of this approach and BoW approach were compared and the results are discussed in the last section of this paper.

The 'Bag-of-Concept' model is based on counting the frequency of the clustered word embeddings (i.e., concepts) in a document. Since text categorization intends to result out text with a specific predefined category, we need to further process the results of Bag-of-Words model to some other algorithm like Support Vector Machine (SVM) for text categorization [16]. This Model has well established the feasibility of leveraging clustered word embeddings to create new features for document representation.

Bag-Of-Words

The Bag-of-Words model consumes a document as a input and break it into words which are

generally known as tokens. The set of unique tokens then form an ordered vocabulary [17]. A vector of equal length is constructed for each document with values representative of the frequency of the observed tokens in the respective document. Since the order in which tokens appear are ignored, the model is termed as 'Bag-of-Words'. Diagram below depicts the flow of process-

Fig. 1: Bag-of-Words Limitations

'Bag-of-Words' model is proven poor in making the sense of text data by returning similar vectorized representations, though two sentences carry different meaning. Vocabulary size will keep on increasing as new document with new words are added, thereby increasing computing complexity, as the length of vector is increased resulting a sparse matrix. 'Bag-of-Words' Model is not accountable for semantic, context, grammar and ordering of words. Due to informational reasons as well as computation reasons it is very difficult to model sparse representations.

Bag-Of-Concepts

In general, document vectors are usually represented as the Bag-of-Concepts which are created by method of clustering term vectors generated by word2vec [18]. The Bag-of-Concepts (BoC) approach offers three-fold advantages over the traditional Bag-of-Words (BoW) approach in text classification tasks [19].

Fig. 2: Bag-of-concepts Limitations

Semantic information

The BoW approach represents a document as a frequency distribution of individual words, without taking into account the semantic relationships between words. In contrast, the BoC approach considers the semantic relationships between words by representing a document as a frequency distribution of concepts. This means that the BoC approach captures a more meaningful representation of the text, which can improve the accuracy of text classification algorithms.

Generalization

The BoW approach relies on exact word matching, which can be problematic when dealing with variations in spelling, grammatical errors, and synonyms. In contrast, the BoC approach can generalize over variations in spelling and grammar because it represents a document using

concepts rather than individual words. This means that the BoC approach can more accurately represent the underlying meaning of a text, even when the wording or phrasing of the text varies.

Feature reduction

The BoW approach can generate a large number of features, which can lead to overfitting and reduced performance of text classification algorithms. In contrast, the BoC approach can reduce the number of features by grouping similar words into concepts. This means that the BoC approach can provide a more compact and meaningful representation of a text, which can improve the efficiency and accuracy of text classification algorithms

Researcher wish to propose a different representational performance. The proposed model produces interpretable features from document vector. There are two major document representation method which are prominently used in solving text mining problems. The popular Bag-of-Words method represent a document vector by computing frequencies of its words but suffers from curse of dimensionality and is not able to hold the information to the best of its proximity when the similar meaning words increases. Disregarding the common semantic this method assumes each word to be independent. Another popular neural network method doc2vec returns low dimensional vector which is able to preserve the proximity information to a great extent.

Author wishes to propose the Bag-of-Concept as the reasonable alternative method that may overcome the drawback of these two methods. In the Proposed method is based on Random Indexing which produces Context vectors, which is nothing but the vectors of words based on cooccurrence of data, which can further be used to generate BoC representations by combining the context vectors for the words that appear in a text [20]. In the Random Indexing method replaces cooccurrence matrix by context matrix.

Support Vector Machine

The concepts of Bag-of-Concepts (BoC) and Bag-of-Words (BoW) are both related to text representation, which is an important task in text classification. Text classification involves assigning pre-defined categories or labels to text documents based on their content [21]. To perform this task, it is necessary to represent the text documents in a way that can be easily processed and analyzed by machine learning algorithms.

The traditional approach to text representation is Bag-of-Words (BoW), which represents a document as a frequency distribution of individual words [16] However, BoW has limitations in capturing the semantic relationships between words and generalizing over variations in spelling and grammar [22]. To address these limitations, the Bag-of-Concepts (BoC) approach was proposed [23]. BoC represents a document as a frequency distribution of concepts, which capture the semantic relationships between words.

In text classification tasks, machine learning algorithms, such as Support Vector Machines (SVMs), are often used to assign labels to text documents based on their representations [24]. The performance of these algorithms depends on the quality of the text representation used. In the research paper discussed earlier, the authors proposed and investigated the use of the BoC approach for text representation in text classification tasks using SVM. The results showed that BoC outperformed BoW in terms of accuracy, particularly when considering the largest

categories. Additionally, the combination of BoW and BoC representations was particularly effective in improving the performance of the SVM classifier across all categories [13].

Therefore, the concepts of BoC, BoW, and SVM are all interconnected in the context of text classification, where the quality of text representation is crucial for achieving high accuracy and efficiency in assigning labels to text documents.

Perused Investigation

This section discusses the text categorization setup, authors used Reuters-21578test collection data for text categorization. We have excluded words with frequency less than 3, and had used the method of cross validation to split the documents into training and test data. The training data happen to found 8887 unique words. Perhaps the BoW representation is very sparse and is of 8887 dimensional. BoC was produced using k-dimensional random index vector for each training document. Context vectors are then subsequently produced by adding the index vector of a document. These context vectors are further used to generate BoC representations and the final output is k-dimensional dense BoC vectors.

Authors used the famous Support Vector Machine (SVM)algorithm for the binary classification. SVM segregate the classes which have maximum margin in hyperplane, which in turn helps in generalization and also takes care of minimizing the empirical error [25]. SVM separates the cases by virtue of hyperplane weight vector.

The next step in this investigation was application of one-against-all learning method. For each category one classifier was trained. To judge the accuracy, with the help of confusion matrix we try to predict the class of the test example. From the defined four outcomes True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) we computed precision(P) and recall(R). The obtained results are then used to generate micro averaged F1 score, which can be used as $F1 = (2*P*R)/(P+R)$.

Weighting scheme and the kernel function as well as the dimensionality of BoC vectors were optimized to the requisite level. Feature selection were not experimented in this investigation, as the foremost objective was to compare results produced by BoW and BoC.

The performance of SVM was compared using three weighting schemes tf, idf and tf × idf, this was done to infer upon the impact of using different weighting scheme on concept representation, findings are summarized in following table

Table I: Micro average score of tf, idf, tf × idf

Parameters	tf	Idf	tf×idf
BoW	80.17	80.19	80.22
BoC 1,000-dim	79.62	80.26	80.28
BoC 2,000-dim	80.28	80.39	80.47
BoC 3,000-dim	79.62	79.86	80.95
BoC 4,000-dim	79.66	79.98	81.14
BoC 5,000-dim	80.17	80.31	81.74
BoC 6,000-dim	79.53	80.61	81.69

The above table is evident that the BoC representation were better when the weighting scheme tf × idf is applied. For BoC, it is interesting to note that idf outperforms then tf.

Comparing BoW And BoC

We compare both the BoW and BoC executions, from the table-1 it is evident that tf×idf weighting is performing best compared to independent tf and idf runs. This can be evident from the first 6 categories, higher the category better the results. BoC representations are found to be more appropriate than BoW, for large categories. Looking to the precision and recall results all six categories performed well. The model was applied to the test data and the results are proving the fact that this proposed approach is giving the justified results. For the categories where BoW'sF1 score is less, the model have given the BoC values in the same range, and was verified by confusion matrix values and through precision and recall values.

This investigation leads to the improvisation of SVM score, by combining BoW and BoC representations for the observed categories. This can be achieved by virtue of F1 score and the confusion matrix quadruple (TP, FP, TN, FN) for each category from BoW or BoC, that produces maximum score. There are two different ways of expressing the F1 score, which is mentioned below

First Method: F1=(2*P*R)/(P+R), Here P = TP/(TP+FP) and R=TP/(TP+FN)

Second Method: F1=(2*TP)/(2*TP+FP+FN)

Combining both the tf and idf had returned the F1 scores as 80.22, compared to the tf and idf runs the difference margin is quite less, but the overall performance for all six category is considerable.

2. Conclusion

Authors have attempted the proposed approach for developing the Bag-of-concept based (BoC) text representations, and have used SVM classifier for performance comparison. The investigation was carried upon the Reuters-21578 collection using both traditional word-based (BoW), and the new concept-based representations, i.e. Bag-of-Concepts. The results are evident that BoC representations are performing better than BoW. For investigation purpose authors have considered only the first six largest categories.

The combination of BoW and BoC representations were observed to play a instrumental role in improving the performance of the support Vector Machine (SVM) over all categories. Authors wish to infer upon that Bag-of-Concept representations can be supplemented to Bag-of-Words representation for better performance of Support Vector Machine. And this can be evident from the conducted investigation.

The results of the investigation showed that the BoC approach outperformed the BoW approach in text classification tasks. Specifically, the BoC approach improved the accuracy of the SVM classifier, particularly when considering the first six largest categories. Additionally, the researchers found that the combination of BoW and BoC representations was particularly effective in improving the performance of the SVM classifier across all categories.

The authors conclude that the BoC approach can be used as a supplement to the BoW approach to improve the performance of the SVM classifier in text classification tasks. This conclusion highlights the potential of the BoC approach to capture the semantic relationships between words and to improve the accuracy of text classification algorithms.

References

1. Calvo, R., and S. D'Mello. 2020. "Affect detection: An interdisciplinary review of models, method and their Applications." IEEE Transactions on Affective Computing 18–37.
2. Cambria, E. 2022. "Affective computing and sentiment analysis." IEEE Intelligent Systems 31(2) 102-107.
3. Cambria, Erik. 2016. "Affective Computing and Sentiment Analysis." IEEE Intelligent Systems (Volume: 31, Issue: 2, Mar.-Apr. 2016) 102-107.
4. Dumais, J. Platt, D. Heckerman, and M. Sahami. 2020. "nductive learning algorithms and represenations for text categorization." Proceedings of ACM-CIKM98 148-155.
5. Esuli, A., and F. Sebastiani. 2006. "SentiWordNet: A publicly available lexical resource for opinion mining." LREC 87-96.
6. Glorot, X., A. Bordes, and Y. Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. Bellevue: ICML,.
7. Hofmann, Lijuan Cai and Thomas. 2019. "Text categorization by boosting automatically extracted concepts." SIGIR 182-189.
8. Ahmad, I., Butt, A. R., & Masood, K. 2019. "A comparative analysis of bag-of-concepts and bag-of-words approaches for text classification using SVM. ." Journal of King Saud University-Computer and Information Sciences, 31(4), 440-446.
9. Kanerva, J. Kristofersson, and A. Holst. 2020. "Random indexing of text samples for Latent Semantic Ananlysis." Proceedings of he 22nd Annual Conference of the Cognitive Science Society 1036.
10. Lau, R., Y. Xia, and Y. Ye. 2014. "A probabilistic generative model for mining cybercriminal networks from online social media." IEEE Computational Intelligence Magazine 9(1) 31-43.
11. [Wiebe, J., T. Wilson, and C. Cardie. 2005. "Annotating expressions of opinions and emotions in language." Language Resources and Evaluation 165-210.
12. Lewis. 2012. "An evaluation of phrasal and clustered representation on text categorizatoin tasks." SIGIR 37-50.
13. McCallum., Baker and A. 2022. "Distributional Clustering of words for text classification." SIGIR 96-103.
14. Vapnik. 2015. "The Nature of Statistical Learning Theory." In The Nature of Statistical Learning Theory, by Vapnik, 16-27. New York: Springer.
15. Strapparava, C., and A. Valitutti. 2014. "WordNet-Affect: An affective extension of WordNet." LREC 1083-1086.
16. Sebastiani. 2022. "Machine learning in automated Text categorization." ACM Computing Surveys 1-47.
17. Repository, UCI Machine Learning. 2022. https://archive.ics.uci.edu. November 25. Accessed November 25, 2022. https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection.
18. Minsky, M. 2016. "The emotion machine: Commonsense thinking, artificial intelligence, and the future of human mind." 62-67.
19. Ortony, A., G. Clore, and A. Collins. 2016. The cognitive structure of emotions. Cambridge: Cambridge University Press.
20. Kharde, V., & Sonawane, R. 2018. "A novel approach for text categorization based on bag of concepts and support vector machine. ." Journal of Ambient Intelligence and Humanized Computing, 9(3), 841-851.
21. Picard, R. 2017. Affective computing. Boston: The MIT Press.
22. Socher, R., A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, and C. Potts. 2013. "Recursive deep models for semantic compositionality over a sentiment treebank." EMNLP 1642–1654.
23. Ngo, H. Q., & Huynh, T. N. 2020. "A new concept-based text representation approach for

text classification. ." Journal of Intelligent & Fuzzy Systems, 38(5), 5745-5756.

24. Zhou, B., Yang, X., Chen, H., & Wang, X. 2020. "Deep Learning Based Approach for Text Classification Using Bag of Concepts. ." In Proceedings of the 2020 International Conference on Internet of Things and Intelligent Applications 1-6.
25. Jindal, A., & Verma, V. K. 2018. "Bag of Concepts based Text Classification using Support Vector Machine." 4th International Conference on Computing Sciences (ICCS) (pp. 185-190). IEEE.