

# A Novel Document Weighted Approach for Text Classification

S. Sai Satyanarayana Reddy<sup>1\*</sup>, N. Hanuman Reddy<sup>1</sup>, T. Raghunadha Reddy<sup>2</sup>

<sup>1</sup>Department of CSE, Vardhaman College of Engineering, Hyderabad, Telangana, India.

<sup>2</sup>Department of IT, Vardhaman College of Engineering, Hyderabad, Telangana, India.

\* Corresponding author. Tel.: 09502653333; email: saishn90@gmail.com

Manuscript submitted February 10, 2020; accepted April 30, 2020.

doi: 10.17706/jcp.15.3.105-113

---

**Abstract:** The textual data in the internet is increasing exponentially through blogs, twitter and various social media sites. The users are not specifying the type of text that they are uploading into the internet. In this regard most of the researchers are looking for automated tools for classifying the data or assigning class label to the unknown documents. Text classification is one such area used for classifying the texts. Several solutions were provided for text classification by the researchers. The text classification approaches generally contains collection of training data, preprocessing of the text, features extraction, feature reduction, document representation and finally applying classification algorithms to build the model for class label prediction of a new textual document. In the phases of text classification, the document representation is one important step to increase the efficiency of the accuracy of text classification. In this work, a new document representation approach is proposed. The experimentation conducted on 20-Newsgroup and Reuters-21578 datasets and different types of classification algorithms. Our approach attained best accuracy results for text classification and observed that the results are more promising than most of the popular approaches for text classification.

**Key words:** Accuracy, bag of words model, document representation, document weight measure, term weight measure, text classification.

---

## 1. Introduction

In the last 20 years, Internet has evolved from a network of connected computers to share data among researchers. As a result of this growth and the birth of social networks, blogs and many other websites where users are given the opportunity of easily creating or uploading content. One of the characteristics of the Internet nowadays is that a user can post anonymously in forums, articles, social networks, chat systems etc. Identifying the type of uploaded data becomes an interesting area for the researchers. Text classification is one research area influenced by the researchers to classify the text into known category.

A traditional text classification also known as text categorization paradigm includes preprocessing step, extraction of features, selection of features, document representation and finally categorization phase. The preprocessing step normally includes lowercase conversion, tokenization, removing of stop words, and stemming. After applying the preprocessing techniques, different researchers extracted different types of features from the dataset which were suitable for differentiating the type of category. In the process of features extraction, some researchers experimented with various feature selection algorithms to decrease the features count as well as to identify the most informative features to discriminate the type of category.

The document representation is one important technique to organize the identified features. Vector space model is one famous technique used by the various researchers in text classification for document vector representation. Finally these vectors of documents were directed to the classification algorithms. In the classification step, classification models are generated. Labeled documents are exploited to train the classification algorithms and the learned model is exploited to classify the unlabeled documents [1], [2].

Text categorization has been exploited in various applications such as topic detection, spam e-mail filtering and web page categorization. To categorize the text, the document is represented as multi-dimensional feature vectors. The weight of each feature is calculated using different weight measures proposed by the researchers in various research domains. The role of document representation is crucial in the performance of the text classification. In this paper, a new document representation approach called as Category specific Document Weighted (CDW) is proposed for representing the documents. Most of the approaches for text classification extracted several features for classifying the text. Excessive number of features degrades classification accuracy and increase computational time. The proposed approach used less number of features to represent the document. The experiment performed on two popular datasets such as 20-Newsgroup and Reuters-21578.

This work is planned in 6 sections. The review of previous works in text classification domain is mentioned in Section 2. The features of 20-Newsgroup and Reuters-21578 dataset and performance measures were presented in Section 3. The traditional model for document representation is described in Section 4. Section 5 analyzes the proposed approach along with term and document weight measures. The empirical evaluations of the CDW approach were discussed in Section 6. Finally, the conclusions of this work were mentioned in Section 7.

## 2. Literature Survey

Normally, text classification includes feature extraction step, feature representation in a document and a classifier which performs the categorization process based on labeled data. Nobata *et al.* [1] proposed the state-of-art method for detecting abusive user content online. They experimented with regression classification algorithm with labeled user comments from Yahoo news. They used a combination of character n-grams, word, syntactic, linguistic and distributional semantics features and obtained the highest accuracy. They observed that the performance of character n-grams is almost on par with all the combined features. In another work [2], they experimented with same dataset of user comments from Yahoo news. They experimented with character n-grams where n ranges from 1 to 5. Support Vector Machine (SVM) and Naive Bayes (NB) were used and observed that the F-score is improved by using these features.

There are numerous different learning algorithms that were used for text classification. One popular algorithm for text classification is Random Forest. In [3], they presented a study of categorizing texts into three degrees of fanaticism such as non-fanatic, code-attitude and code-red regarding terrorism using Random Forest as learning algorithm. In their study, bag-of-words were extracted and used as features for representing the document vectors. The performance of the Random Forest was compared with SVM, Naive Bayes and C4.5 and observed that Random Forest was the most reliable classifier and obtained the highest accuracy of 69%. Thorsten Joachims was one of the early adopters of SVM within text classification to categorize texts by topic who compared SVMs with four popular learning algorithms used for text classification. They observed that the SVMs outperformed the other learning algorithms and the study resulted in both theoretical and empirical evidence that SVMs are well suited for text classification [4].

In text classification domain, several term weight measures were proposed by the researchers for giving weight to the features. Lan et al proposed [5] an effective and simple supervised term weighting measures

known as tf.rf. This measure used the information of how the term is distributed in various categories of the training data. They found that the tf.rf measure performance is good when compared with other measures like tf.idf, tf.IG, tf-chi2 and tf.logOR. Wang & Zhang proposed [6] a measure by adding inverse category information in the measure of tf.rf. They observed that the performance of their measure is close to tf.rf measure performance but not in all cases.

Most of the techniques used for text classification are based on words. However, in 2015, LeCun *et al.* presented [7] a character-level CNN which showed to be an effective method for text classification. The authors stated that, CNNs were not used the information of the semantic or syntactic structure of a language. One other approach taken for text classification based on characters is the work done by Douglas Bangall [8], who showed that a simple character level RNN was a useful model for text classification. ANN models such as CNN and LSTM has recently shown to be promising within the area of text classification [7], [8].

### 3. Dataset Characteristics and Performance Measures

In this work, two datasets have been exploited to measure the proposed approach performance. The first dataset is the 20-Newsgroups collection which is a balanced (number of documents per category are equal) dataset. The 20-Newsgroups contain 18828 documents from 20 news categories. The dataset is divided into 66% of documents (12426) for training and 34% of documents (6402) for testing. Another dataset Reuters-21578 consists of 21578 news articles extracted in 1987 from Reuters newswire. The Reuters-21578 collection is not a balanced (the number of documents in each class are different) data set. The categories in Reuter's dataset were classified manually into 135 subcategories. In this work, ten most frequent categories (Acquisition, Crude, Corn, Earn, Interest, Grain, Money-fx, Trade, Ship and Wheat) were used for experimentation.

The researchers used various performance evaluation measures to estimate the efficiency of the text classification in their approaches. In this paper, the accuracy measure is used to test the efficiency of our proposed CDW approach. Accuracy is defined as the proportion of documents classified correctly from a set of test documents.

### 4. Traditional Bag of Words Model

Most of the researchers used BOW model for representing the documents as vectors in text classification. The Fig. 1 shows the BOW model.

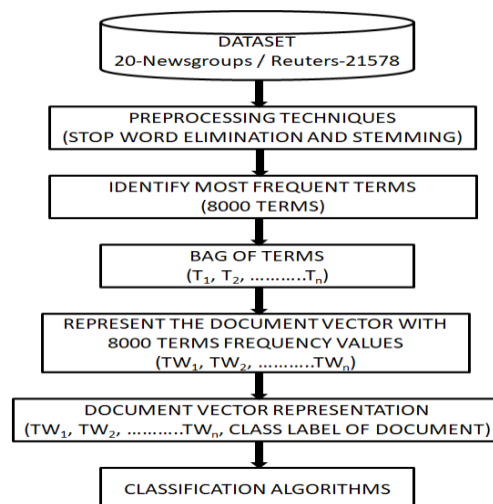


Fig. 1. The steps in BOW model.

In BOW model, first clean the training data to extract the efficient features for text classification. In this work, two cleaning techniques such as elimination of stop words and stemming were applied on the corpus to remove unwanted data for analysis. After performing cleaning, extract the terms which were more frequent in the dataset. We identified most frequent 8000 terms for experimentation. Consider these 8000 terms as bag of terms. In BOW model, every document is represented with the terms that were specified in the bag. The values of the vector are represented with the frequency of the terms in a document.

In Fig. 1,  $(T_1, T_2, \dots, T_n)$  represents the set of terms,  $(TW_1, TW_2, \dots, TW_n)$  represents the weights of the terms. In this model, term frequency is used to specify the weight of a term.  $(TW_1, TW_2, \dots, TW_n, \text{CLASS LABEL OF DOCUMENT})$  represents the document vector with values of  $n$  number of terms and class label of a document. This representation of document vectors was given to classification algorithms such Naïve Bayes Multinomial (NBM), Support Vector Machines (SVM) and Random Forest (RF). These algorithms generate the classification model. The class label of anonymous document is predicted with the classification model. The classification results of two datasets were presented in Table 1.

Table 1. The Accuracies of Text Classification for BOW Model with 8000 Most Frequent Terms

MOST FREQUENT TERMS	20-NEWSGROUP			REUTERS-21578		
	NBM	SVM	RF	NBM	SVM	RF
1000	58.59	62.37	65.78	61.74	77.21	68.63
2000	60.13	64.28	66.46	62.31	78.45	69.27
3000	61.75	65.26	66.81	62.92	80.01	69.83
4000	62.62	65.71	68.37	63.85	82.13	71.25
5000	65.41	67.07	70.14	65.76	83.92	72.46
6000	66.02	67.72	70.53	67.42	85.83	72.84
7000	66.89	69.48	71.87	69.63	86.02	74.33
8000	68.54	70.89	73.71	70.97	71.11	75.48

In Table 1, the accuracies of text classification is increased when the number of terms were increased to represent the document vector. It was observed that the Random Forest classifier obtained good accuracy of 73.71% and 75.48% for text classification in 20-NewsGroup and Reuters-21578 datasets respectively. The RF classifier performance is good when compared with other classifiers.

## 5. Category Specific Document Weighted (CDW) Approach

Most of the researchers in text classification focused on the extraction of features, representation of features as document vectors and the classification algorithms used for generating classification model. The existing solutions for text classification represent the document vector with more number of features thereby high dimensionality problem is occurred and the features independently participated in the generation of classification model thereby no usage of the relationship between features. In this paper, a new approach named as CDW approach is proposed to overcome the drawbacks of previous techniques in text classification. The procedure of CDW approach is shown in Fig. 2.

In this approach, to prepare the data for efficient features extraction, two preprocessing methods such as stop word removal and stemming were used. Once the data is cleaned then extract the high frequency terms from the entire training dataset. After extraction of most frequent terms, each document is represented with these terms as document vector. In this experiment, top frequent 8000 terms were considered for document vector representation. In this model, we used term weight measures to represent the terms weight in the vectors of document. Each term importance is computed by computing the term's weight against each category group of documents. These term's weights of individual category were used to calculate the document weight against to every category group. The document vectors were represented

with weights of the documents. Finally, classification model is created by using these vectors. This model used to know the category of an anonymous text.

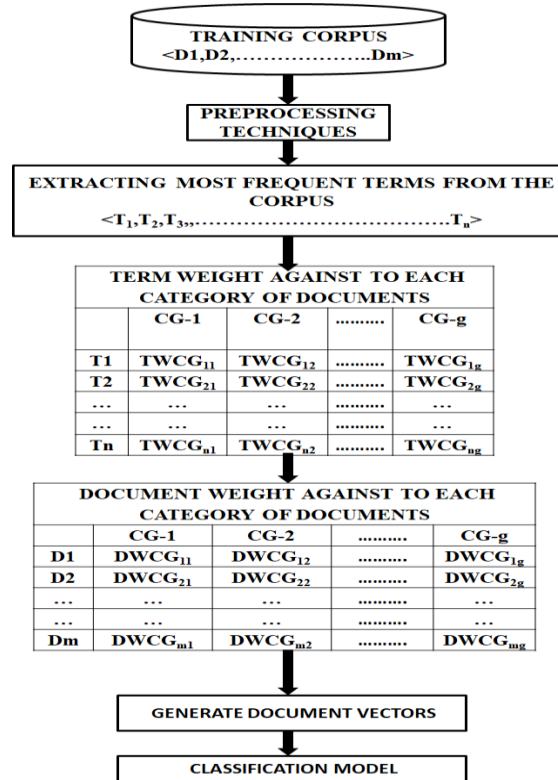


Fig. 2. The steps in CDW approach.

In this CDW approach,  $\{D_1, D_2, \dots, D_m\}$  represents the set of documents in the dataset,  $\{T_1, T_2, \dots, T_n\}$  represents the set of most frequent terms,  $\{CG-1, CG-2, \dots, CG-g\}$  represents a set of categories in the dataset.  $TWCG_{ng}$  represents term  $T_n$  weight in the category group  $CG-g$ ,  $DWCG_{mg}$  represents the document weight of  $D_m$  in the category group  $CG-g$ . The next two sections discuss the weight measures used in this approach.

### 5.1. Supervised Unique Term Weight Measure (SUTW)

In this work, we used a SUTW measure proposed by Reddy et al., [9] for computing the weight of the term. This measure considers the importance of the distributional information of terms. In general, most of the weight measures functioning based on the distribution of the term within a category of documents, within a document and across categories of documents. This measure considers all three information to calculate the weight of the term. Equation (1) shows the SUTW measure.

$$W(T_i, CG-g) = \sum_{k=1, D_k \in CG-g}^m \left( \frac{TF(T_i, D_k)}{TF(T_i, CG-g)} \left[ \frac{\log(DT_k)}{0.2 * UT_k + 0.8 * AVGUT_k} \right] \right) \times \frac{P_{ig}}{(Q_{ig} + R_{ig})} \times \frac{R_{ig}}{(R_{ig} + S_{ig})} \quad (1)$$

where,  $W(T_i, CG-g)$  is the term  $T_i$  weight in the  $g^{th}$  category,  $TF(T_i, D_k)$  is frequency of  $T_i$  in document  $D_k$ ,  $TF(T_i, CG-g)$  is the frequency of  $T_i$  in  $g^{th}$  category of documents,  $DT_k$  is the  $k^{th}$  document count,  $UT_k$  and  $AVGUT_k$  is the frequency of unique terms and the average number of unique terms in  $k$  document respectively. Average number of unique terms is the ratio among number of unique terms and total number of terms.

Table 2. Confusion Matrix

	<i>CG-g</i>	<i>Except CG-g</i>
<i>T<sub>i</sub></i>	<i>P<sub>ig</sub></i>	<i>R<sub>ig</sub></i>
<i>No T<sub>i</sub></i>	<i>Q<sub>ig</sub></i>	<i>S<sub>ig</sub></i>

Table 2 represents the distribution of term  $T_i$  in categories of documents. Where,  $P_{ig}$  and  $Q_{ig}$  is the count of  $g^{th}$  category documents contain the  $T_i$  term and doesn't contain the  $T_i$  term respectively,  $R_{ig}$  is the count of other than  $g^{th}$  category documents contain the  $T_i$  term,  $S_{ig}$  is the count of other than  $g^{th}$  category documents does not contain the  $T_i$  term.

### 5.2. Document Weight Measure (DWM)

In this approach, the DWM computes the weight of document against a category of documents. This measure uses the individual term weights computed in the term weight measure and TFIDF weight of individual term is computed in each document. The measure is specified in equation (2).

$$W(D_k, CG-g) = \sum_{T_i \in D_k, D_k \in CG-g} W(T_i, CG-g) * TFIDF(T_i, D_k) \tag{2}$$

In this measure,  $W(D_k, CG-g)$  is the weight of  $D_k$  document in  $g^{th}$  category of documents,  $W(T_i, CG-g)$  is the term  $T_i$  weight in the  $g^{th}$  category,  $TFIDF(T_i, D_k)$  is  $TFIDF$  of term  $T_i$  in  $D_k$  document.

To avoid high dimensionality problem, the proposed approach used less number of features for document representation. The number of features considered for representing the document depends on the number of categories in the dataset. The CDW approach used 20 features for 20-NewsGroup dataset and 10 features for Reuters-21578 dataset to represent the document vectors.

## 6. Empirical Evaluations

In the experimentation, the dataset is divided into training and test data. The classification algorithms adjust the internal parameters based on the training data and generate a model. The test data is used for evaluating the finished model and predicting the unknown information. This experiment is carried out on two datasets such as 20 NewsGroup and Reuters-21578 using accuracy as a performance measure.

The proposed approach achieved good accuracies based on the effectiveness of the term and document weight measures. The accuracies of two datasets and different classifiers were presented in Table 3. In Table 3, it was identified that when the count of terms increased from 1000 to 8000 with an interval of 1000 terms for computing the document weight, the improvement in accuracy was increased. It was also witnessed that, the accuracies are increased in all the classifiers when the count of terms were increased for calculating the document weight.

Table 3. The Accuracies of Text Classification for CDW Approach

MOST FREQUENT TERMS	20-NEWSGROUP			REUTERS-21578		
	NBM	SVM	RF	NBM	SVM	RF
1000	75.09	81.79	82.55	75.09	77.21	80.21
2000	76.29	82.44	84.90	77.29	78.45	83.06
3000	79.48	85.74	86.31	78.48	80.01	84.12
4000	83.31	86.43	87.56	80.31	82.13	86.67
5000	87.39	88.39	90.81	83.17	83.92	87.71
6000	88.16	90.81	91.65	84.87	85.83	88.47
7000	89.67	91.39	93.32	86.39	86.02	90.23
8000	90.71	92.17	94.71	87.16	88.11	92.89

The RF classifier attained a best accuracy of 94.71% for 20-newsgroup dataset and 92.89% accuracy for Reuters-21578 dataset when RF classifier is used. Among all classifiers, the RF classifier attained best accuracies in both BOW model and CDW approaches. It was also identified that the proposed CDW results were good compared to the results of BOW model as well as most of the approaches of text classification.

The existing approaches extracted more features like content based, character based, word based, syntactic features for document representation and these features are individually participating in the process of classification. In the proposed approach, all the features were participated collaboratively in the classification process that's why our proposed approach achieved good accuracies than existing approaches for text classification.

## 7. Conclusions

In this work, a new CDW approach was proposed for classifying the text. This approach used the document weights instead of feature weights in the document representation. The document weight is calculated against the category of documents. Each document weight is determined with DWM by combining the feature weights within a document and within category of documents. Two datasets namely 20-Newsgroup and Reuters-21578 used in this experiment and three classifiers namely NBM, RF and SVM were used for classification. The BOW model obtained the accuracies of 73.71% and 75.48% for text classification in 20-NewsGroup and Reuters-21578 datasets respectively. The CDW approach achieved best accuracy of 94.71% for text classification in 20-newsgroups dataset and 92.89 % for text classification in Reuters-21578 dataset. The RF classifier obtained best results for text classification in both BOW and CDW models. The CDW approach performed well when compared with BOW model.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

Dr. Sai Satyanarayana Reddy conducted the research. He analyzed the various methods implemented in text classification and developed a new method to improve the accuracy of text classification. Dr. T. Raghunadha Reddy identified the various datasets used in the experiments of text classification. He analyzed the datasets which are used in this work. Mr. N. Hanuman Reddy having good writing skills. He wrote the paper with good flow of understanding.

## References

- [1] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web* (pp. 145-153).
- [2] Mehdad, Y., & Tetreault, J. (2016). Do characters abuse more than words? *Proceedings of the SIGdial 2016 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 299-303).
- [3] Amasyalı, M. F., & Diri, B. (2006). Automatic Turkish text categorization in terms of author, genre and gender. *Proceedings of the International Conference on Application of Natural Language to Information Systems* (pp. 221-226).
- [4] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning, ECML '98* (pp. 137-142).
- [5] Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(4), 721-735.

- [6] Wang, D., & Zhang, H. (2013). *Inverse Category Frequency Based Supervised Term Weighting Scheme for Text Categorization*.
- [7] Zhang, X., Zhao, J., & Yann, L. C. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, pp. 649-657, Curran Associates, Inc..
- [8] Bagnall, D. (2015). Author identification using multi-headed recurrent neural networks. *CoRR*, vol. abs/1506.04891.
- [9] Raghunadha, R. T., Vishnu, V. B., & Vijayapal, R. P. (2016). Profile specific document weighted approach using a new term weighting measure for author profiling. *International Journal of Intelligent Engineering and Systems*, 9(4), 136-146.

Copyright © 2020 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).



**Seelam Sai Satyanarayana Reddy** is the principal of Vardhaman College of Engineering, Hyderabad and he is a professor in computer science and engineering. Alumni of BITS, Pilani (Rajasthan). He is an active researcher and published good research papers with 26 years of teaching experience. He is the guiding research scholars, a member of IETE, SIEEE, CSI, FIE, FIETE, IET, ASEE, GEDC, ISTE and IAENG. He is certified professional in microsoft, IBM and CISCO. His area of interest are Data mining, cloud computing, machine learning and artificial intelligence. He is very hard working in nature, good integrator and driven by nature. He has published more than 50 papers in international conferences and journals.



**N. Hanuman Reddy** is an associate professor and have 10 years of experience in teaching. He received the M. Tech in software engineering at JNTU, Hyderabad, and received the UG degree in B. Tech in computer science and engineering in 2003 from EVP College of Engineering, University of Madras. He published more than 10 Scopus indexed papers. His area of interest are cloud computing and machine learning. He is familiar with software engineering, computer networks, internet of things, programming in CIT workshop.



**T. Raghunadha Reddy** received the B.Tech, M.Tech PhD in computer science and engineering from Jawaharlal Nehru Technological University (JNTU), Hyderabad, Telangana. He has 15 years of teaching experience. Currently, he is working as an associate professor in the Department of Information Technology, Vardhaman College of Engineering, Shamshabad, Hyderabad, He published 45 research papers in reputed international journals and international conferences such as text classification, natural language processing, information retrieval and cloud computing. He has memberships in CSI, IET, IFERP and IAENG.