A New Supervised Term Weight Measure Based Machine Learning Approach for Text Classification

T. Raghunadha Reddy, P. Vijaya Pal Reddy, and P. Chandra Sekhar Reddy

Abstract Text classification is a technique of predicting assigning a class label of an anonymous document or classifying the documents into known classes. The content of a text is a primary source for classifying the data in text classification. The researchers used the content of a text in different ways like most frequent terms, TFIDF scores of terms, N-grams of word and character for text classification. In this work, a term weight measure-based machine learning approach is proposed for text classification. In this approach, we propose a new term weight measure to represent the importance of a term in a document. The terms which are more frequent in the dataset are extracted to represent the documents as vectors. The term value in vector representation is determined by using term weight measure. Four term weight measures are used in this experiment to compute the weight of a term. Machine learning algorithms are trained by using these vectors to generate the model for classification. The performance of a proposed system is predicted by using this classification model. Accuracy measure is used as a performance evaluation measure. Six machine learning algorithms and two datasets namely, IMDB and Enron Spam datasets are used in this work. The proposed term weight measure-based approach efficiency is good when compared with results of popular approaches to text classification.

1 Introduction

The textual information in the Internet is increasing tremendously through different social media platforms. The type of text is very important for better organization of text and effective accessing of information in the Internet. Text classification (TC) is one such technique to classify the documents into different categories as

T. Raghunadha Reddy · P. Vijaya Pal Reddy (✉)
Department of CSE, Matrusri Engineering College, Hyderabad, Telangana, India

P. Chandra Sekhar Reddy
Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India

V. S. Reddy et al. (eds.), *Intelligent Systems and Sustainable Computing*,
Smart Innovation, Systems and Technologies 289,
https://doi.org/10.1007/978-981-19-0011-2_50

well as detecting the class label of an anonymous document [1]. The researchers proposed several approaches based on stylistic features, content-based features, word n-grams, character n-grams, term weight measures (TWMs), feature selection techniques, similarity based techniques, machine learning (ML) techniques and deep learning (DL) techniques for text classification [2]. In general, the text classification approaches are divided into several steps such as data collection, data preprocessing, feature extraction, dimensionality reduction, document vector representation, machine learning techniques and exploration of results.

Several researchers focused on the development of TWMs to compute the term weight in a document [3]. The TWMs are majorly categorized into two types such as supervised TWMs (STWM) and unsupervised TWMs (UTWM) by considering the utilization of class membership information. The STWMs used class membership information while computing the term weight. The UTWM not used class membership information while determining the term weight [4]. In this work, we concentrated on the proposal of a novel supervised TWM to determine weight of terms based on the term distributions in various classes. The proposed measure used the information of the way the term is distributed in positive class of documents, negative class of documents, within a document and total dataset to calculate the term weight in a document.

In this work, a TWM based machine learning method is proposed for text classification. In this approach, TWMs are used to determine the term value in vector representation of documents. Various ML techniques such as Decision Tree (DT), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbor (KNN) [5], Support Vector Machine (SVM) and linear SVM are used to generate the model to assess the efficiency of the proposed approach. Two datasets namely Enron spam dataset and IMDB movie reviews datasets are used in this experiment. The proposed TWM performance is compared with various popular TWMs such as TFIDF, TFRF and TFIDFICF.

This chapter is structured in 5 sections. Dataset characteristics are presented in Sect. 2. The proposed approach is explained in Sect. 3. The existing and proposed term weight measures are described in Sect. 4. The experimental results are discussed in Sect. 5. The conclusions and future scope are explained in Sect. 6.

2 Dataset Characteristics

The characteristics of datasets are mentioned in Table 1.

The dataset play a major role in the text classification process. The datasets which are collected from known sources with correct labels are obtaining good accuracies in the classification process. In this work, two datasets such as IMDB and Enron spam datasets are used.

Table 1 The dataset characteristics

S. No.	Dataset name	Number of documents/samples
1	IMDB [6]	50,000 (positive and negative reviews)
2	Enron spam dataset [7]	Ham (3672 files), spam(1500 files)

3 A Term Weight Measure-Based ML Approach for Text Classification

In this work, a TWM-based machine learning approach is proposed for text classification. In this approach, term weight measures and machine learning algorithms are used to improve the accuracy of classification. The proposed model is represented in Fig. 1.

In this approach, firstly, apply preprocessing techniques like tokenization, lowercase conversion, removal of punctuations, stop words removal and stemming. Stop words are words which are commonly used by the users like articles, prepositions, etc. but they are not useful in the process of classification [8]. Stemming is a technique of converting a word into its root form [9]. After cleaning the data from the dataset, the dataset contains all informative words which are useful for classification. Next step is extracting the important terms for experiment. The most frequent terms are selected as important terms. The documents are represented with these selected terms as document vectors. The value of a term in vector representation is determined by using TWM. Once the vectors representation is ready the ML algorithms are used to train on the document vectors and generate a classification model and it is used to determine the performance of the proposed approach as well as detect the class label of a test document.

4 Term Weight Measures (TWMs)

The TWMs determine the value of a term based on the importance of a term in document. In this work, we proposed a new TWM and compared with three popular term weight measures.

4.1 Term Frequency and Inverse Document Frequency (TFIDF)

TFIDF assign weight to the terms based on term frequency and documents that are containing the term at least one time [10]. Equation (1) is used to compute the TFIDF

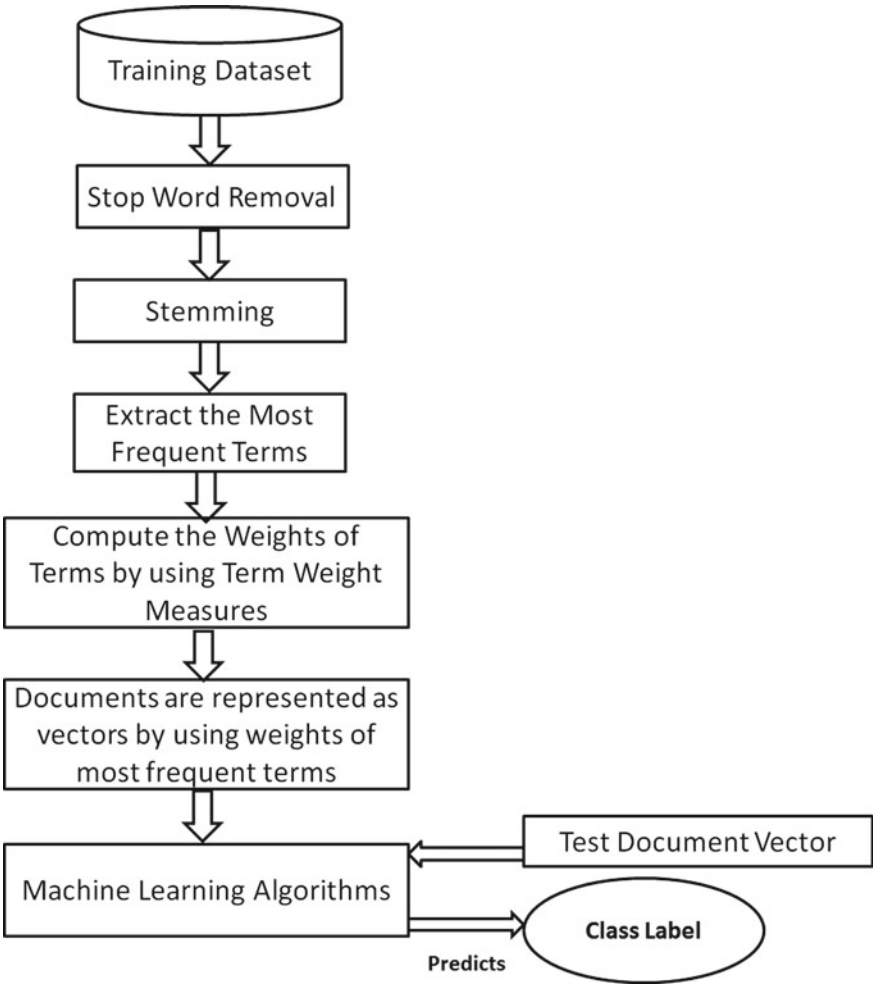

Fig. 1 The proposed approach

measure.

$$\mathrm{TFIDF}(\mathrm{T}_i, \mathrm{D}_k) = \mathrm{TF}_{ik} \times \log\left(\frac{N}{\mathrm{DF}_i}\right) \tag{1}$$

In this measure TF_{ik} is the occurrence count of T_i term in D_k document, N is to total count of documents in whole dataset, DF_i is the documents count in dataset which contain term T_i at least one time.

4.2 Term Frequency and Relevance Frequency (TFRF) Measure

TFRF measure considers the information of term presence of documents in negative and positive class [11]. Equation (2) is used to compute the TFRF measure.

$$\mathrm{TFRF}(\mathrm{T}_i, D_k) = \mathrm{TF}_{ik} \times \log\left(2 + \frac{A}{\max(1, C)}\right) \tag{2}$$

In this measure, TF_{ik} is frequency of term T_i in kth document, A is count of positive class documents have a term T_i, C is the documents count of negative class that have a term T_i.

4.3 Term Frequency Inverse Document Frequency and Inverse Class Frequency (TFIDFICF) Measure

TFIDFICF measure considers the information of term frequency of a term, count of documents contains term and count of classes contains term [12]. Equation (3) is used to determine the TFIDFICF measure.

$$\mathrm{TFIDFICF}(\mathrm{T}_i, \mathrm{D}_k) = \mathrm{TF}_{ik} \times \log\left(\frac{N}{\mathrm{DF}_i}\right) \times \log\left(\frac{|C|}{\mathrm{CF}_i}\right) \tag{3}$$

In this measure, TF_{ik} is term T_i occurrence count in D_k document, N is total documents in dataset, $\mathrm{DF_i}$ is documents count that contain term T_i, $|C|$ is classes count in dataset and CF_i is classes count which contain term T_i.

4.4 Proposed Class Specific Supervised Term Weight Measure (CSTWM)

The proposed STWM considers different types of information of terms in the dataset, in other words, the way terms are distributed in the dataset. This measure gives more weight to the terms that are occurred more times in the dataset, the terms that are discussed in less number of documents in the dataset, the terms that are discussed in less number of classes and the terms that are occurred more times in positive class of documents when contrasted with negative class of documents.

$$\mathrm{CSTWM}(\mathrm{T}_i, \mathrm{D}_k) = \mathrm{TF}(\mathrm{T}_i, \mathrm{D}_k) \times \log\left(\frac{|D|}{1 + \mathrm{DF}_i}\right)$$

$$\times \log\left(\frac{|C|}{1+\mathrm{CF}_i}\right) \times \left(\frac{A}{1+C}\right) \tag{4}$$

where $|D|$ is total documents in dataset, $|C|$ is classes count in dataset, CF_i is count of classes in dataset which contain ith term.

5 Experimental Results of Term Weight Measures

The experiment is conducted with the four term weight measures and six machine learning algorithms for text classification. In this work, 8000 terms that are more occurred in the dataset are extracted for vectors representation of documents. The experiment started with 2000 terms and increases the number of terms in next iterations by 2000. It was observed that the text classification accuracies are reduced when experimented with more than 8000 terms. The documents are represented with extracted features as vectors. The machine learning algorithms trained on these vectors and gives the accuracy of proposed system. The text classification accuracies of different classifiers on IMDB dataset is showed in Table 2.

In Table 2, the combination of most frequent 8000 terms and RF algorithms achieved accuracy 0.848 for text classification when the TFRF measure is used to compute the term weight in vector representation. The combination of most frequent 8000 terms and RF algorithms achieved accuracy 0.851 for text classification when the TFIDFICF measure is used to compute the term weight in vector representation. The combination of 8000 terms and RF algorithms achieved accuracy 0.859 for text classification when the proposed CSTWM is used to compute the weight of terms in vector representation. Overall, the proposed CSTWM attained best accuracies for text classification when compared with other TWMs. It was found that the RF classifier shows best efficiencies for TC when contrasted with other machine learning algorithms.

Table 3 displays the accuracies of TC when experimented with most frequent terms and machine learning algorithms on Enron Spam dataset.

In Table 3, the combination of 4000 terms and LR algorithm achieved accuracy 0.910 for text classification when the TFIDF measure is used to compute the term weight in vector representation. The combination of 4000 terms and LR algorithm achieved accuracy 0.913 for text classification when the TFRF measure is used to determine the weight of terms in vector representation. The combination of 6000 terms and LSVM algorithm achieved accuracy 0.915 for text classification when the TFIDFICF measure is used to compute the weight of terms in vector representation. The combination of most frequent 8000 terms and LSVM algorithm achieved accuracy 0.919 for text classification when the proposed CSTWM is used to calculate the weight of terms in vector representation. Overall, the proposed CSTWM obtained good accuracies for text classification when compared with other TWMs. It was

Table 2 The accuracies of text classification on IMDB dataset

Term weight measures/machine learning algorithms/most frequent terms		2000	4000	6000	8000
TFIDF	LR	0.805	0.828	0.832	0.846
	KNN	0.606	0.574	0.589	0.568
	SVM	0.793	0.812	0.815	0.831
	LSVM	0.821	0.824	0.812	0.828
	GNB	0.731	0.691	0.637	0.635
	DT	0.659	0.656	0.667	0.656
	RF	0.821	0.819	0.821	0.839
TFIDFICF	LR	0.805	0.827	0.839	0.849
	KNN	0.606	0.574	0.593	0.574
	SVM	0.793	0.811	0.814	0.835
	LSVM	0.821	0.827	0.819	0.830
	GNB	0.731	0.691	0.639	0.645
	DT	0.656	0.672	0.661	0.660
	RF	0.829	0.826	0.839	0.851
TFRF	LR	0.814	0.823	0.827	0.836
	KNN	0.579	0.582	0.574	0.567
	SVM	0.795	0.81	0.819	0.825
	LSVM	0.806	0.798	0.802	0.803
	GNB	0.731	0.691	0.629	0.618
	DT	0.665	0.668	0.659	0.658
	RF	0.818	0.828	0.841	0.848
CSTWM	LR	0.801	0.828	0.835	0.833
	KNN	0.601	0.569	0.580	0.573
	SVM	0.788	0.816	0.823	0.834
	LSVM	0.803	0.796	0.803	0.807
	GNB	0.731	0.692	0.621	0.618
	DT	0.663	0.665	0.667	0.666
	RF	0.830	0.834	0.845	0.859

identified that the LSVM and LR classifier shows best performance when contrasted with other machine learning algorithms.

Table 3 The accuracies of TC on enron spam dataset

Term weight measures/machine learning algorithms/most frequent terms		2000	4000	6000	8000
TFIDF	LR	0.899	0.910	0.909	0.909
	KNN	0.851	0.850	0.840	0.842
	SVM	0.896	0.903	0.903	0.905
	LSVM	0.869	0.878	0.878	0.877
	GNB	0.731	0.769	0.769	0.801
	DT	0.889	0.881	0.874	0.888
	RF	0.905	0.903	0.901	0.906
TFIDFICF	LR	0.905	0.913	0.911	0.909
	KNN	0.867	0.866	0.859	0.845
	SVM	0.901	0.894	0.895	0.886
	LSVM	0.906	0.910	0.915	0.911
	GNB	0.731	0.772	0.774	0.803
	DT	0.892	0.879	0.874	0.889
	RF	0.904	0.901	0.902	0.906
TFRF	LR	0.902	0.913	0.912	0.909
	KNN	0.866	0.866	0.863	0.853
	SVM	0.897	0.886	0.886	0.889
	LSVM	0.869	0.891	0.893	0.887
	GNB	0.729	0.773	0.774	0.809
	DT	0.889	0.882	0.878	0.885
	RF	0.906	0.901	0.902	0.906
CSTWM	LR	0.906	0.915	0.913	0.913
	KNN	0.879	0.883	0.877	0.863
	SVM	0.901	0.893	0.894	0.891
	LSVM	0.907	0.910	0.905	0.919
	GNB	0.733	0.778	0.780	0.908
	DT	0.890	0.883	0.879	0.885
	RF	0.904	0.903	0.899	0.907

6 Conclusions and Future Scope

In this work, a new a term weight measure-based ML approach is proposed for text classification. Four TWMs are used to determine the value of a term in document vector representation. A new TWM is proposed and observed that the proposed CSTWM attained best accuracies for text classification when compared with existing TWMs. For IMDB dataset, the combination of RF classifier and proposed CSTWM

attained an accuracy of 0.859 for text classification. For Enron Spam dataset, the combination of proposed term weight measure and SVM classifier attained an accuracy of 0.919 for text classification.

In future work, we are planned to implement to a new vector representation to avoid the problems in existing document representations. It was also planned to implement DL techniques to increase the accuracy of TC.

References

1. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: A survey on author profiling techniques. Int. J. Appl. Eng. Res. **11**(5), 3092–3102 (2016)
2. Khatoon, T., Govardhan, A., Sujatha, D.: Improving document relevant accuracy by distinguish Doc2query matching mechanisms on biomedical literature. In: IEEE 10th International Conference on Cloud Computing, Data Science and Engineering, pp. 29–31 (2020)
3. Chandra Sekhar Reddy, P.: Gender classification using central fibonacci weighted neighborhood pattern flooding binary matrix (CFWNP_FBM) shape primitive features. Int. J. Eng. Adv. Technol. **8**(6), 5238–5244 (2019). (IJEAT) ISSN: 2249-8958
4. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: A document weighted approach for gender and age prediction. Int. J. Eng. Trans. B Appl. **30**(5), 647–653 (2017)
5. Khatoon, T.: Query expansion with enhanced-BM25 approach for improving the search query performance on clustered biomedical literature retrieval. J. Digital Inform. Manag. **16**(2) (2018)
6. https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews?
7. https://www.kaggle.com/wanderfj/enron-spam
8. Chandra Sekhar Reddy, P., Vara Prasad Rao, P., Kiran Kumar Reddy, P., Sridhar, M.: Motif shape primitives on fibonacci weighted neighborhood pattern for age classification. In: Wang, J., Reddy, G., Prasad, V., Reddy, V. (eds) Soft Computing and Signal Processing. Advances in Intelligent Systems and Computing, vol 900. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-3600-3_26
9. Avanthi, M., Chandra Sekhar Reddy, P.: Human facial expression recognition using fusion of DRLDP and DCT features. In: Satapathy, S.C., Bhateja, V., Favorskaya, M.N., Adilakshmi T. (eds) Smart Computing Techniques and Applications. Smart Innovation, Systems and Technologies, vol. 224. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1502-3_20
10. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: Author profile prediction using pivoted unique term normalization. Indian J. Sci. Technol. **9**(46) (2016)
11. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: Profile specific document weighted approach using a new term weighting measure for author profiling. Int. J. Intell. Eng. Syst. **9**(4), 136–146 (2016)
12. Ren, F., Sohrab, M.G.: Class-indexing-based term weighting for automatic text classification. Inf. Sci. **236**, 109–125 (2013)