Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5516262, 8 pages
https://doi.org/10.1155/2021/5516262

Research Article

Hybrid Feature Selection for Amharic News Document Classification

Demeke Endalie and Getamesay Haile

Factuality of Computing and Informatics, Jimma Institute of Technology, Jimma, Ethiopia

Correspondence should be addressed to Demeke Endalie; demeke.endalie@ju.edu.et

Received 29 January 2021; Revised 23 February 2021; Accepted 27 February 2021; Published 12 March 2021

Academic Editor: Ali Ahmadian

Today, the amount of Amharic digital documents has grown rapidly. Because of this, automatic text classification is extremely important. Proper selection of features has a crucial role in the accuracy of classification and computational time. When the initial feature set is considerably larger, it is important to pick the right features. In this paper, we present a hybrid feature selection method, called IGCHIDF, which consists of information gain (IG), chi-square (CHI), and document frequency (DF) features' selection methods. We evaluate the proposed feature selection method on two datasets: dataset 1 containing 9 news categories and dataset 2 containing 13 news categories. Our experimental results showed that the proposed method performs better than other methods on both datasets 1and 2. The IGCHIDF method's classification accuracy is up to 3.96% higher than the IG method, up to 11.16% higher than CHI, and 7.3% higher than DF on dataset 2, respectively.

1. Introduction

Amharic is one of the Ethiopian languages, grouped under Semitic branch of Afro-Asiatic language. Amharic serves as the official working language of Ethiopia Federal Democratic Republic, and it is the second most spoken Semitic language in the world after Arabic with a total speakers of around 25 million, as per the census of 2007 [1].

Nowadays, the volume of Amharic digital document has increased rapidly on Internet. Due to this, Amharic document classification is strongly needed. The method by which tags or categories are allocated to text according to their content is called text classification [2]. It is one of Natural Language Processing (NLP) task with wide applications for subject labeling, spam detection, and intent detection [3]. Owing to the existence of unstructured data in the form of text, emails, social media, conversations, web pages, and survey answers are everywhere. The extraction of insights from these media is challenging and time-consuming. Automatic text classification plays a great role in the dynamic management of textual information gained from these sources.

In five simple steps, automatic text classification can be performed. These are the preprocessing, term-weighing, document representation, dimension reduction, and classification. The accuracy of classification is influenced by the methods used in certain text classification stages. Curse of dimensionality is one of the major problems which affects the classification accuracy and computational complexity [4]. Dataset with large number of attributes or features results in difficulties in analyzing or visualizing data to identify patterns and train the machine learning model [5]. Text documents can contain hundreds or thousands of unique terms. If we use all terms in the classification, we may get a poor result because some terms are not helpful for the classification and some of the terms mislead the classifier.

Dimensionality reduction is used to reduce time complexity and space complexity, saving the cost of observing all features and making the system robust by using the most relevant features of the dataset [6]. Feature selection is one of the dimensionality reduction methods that help to choose the most relevant terms before applying the learning algorithm [7]. Feature selection methods can be divided mainly into three categories: filter, wrapper, and embedded [5]. Filter methods do not depend on the classification algorithm rather it focused on evaluating the importance of features in the classification process [8]. The wrapper method of feature

selection considers features' dependency and provides the interaction between feature subset selection and learning model used [9], whereas the embedded method is similar to the wrapper method since they also used to optimize the objective function or performance of the learning algorithm or model [10]. Their basic difference is that an intrinsic model building metric is used during the learning phase. The filter method is used in this paper, since the wrapper and embedded methods are not feasible for large feature size, biased to learning algorithm and computationally expensive [8, 11].

Many feature selection methods were discussed for improving English and Arabic text classification [8, 12, 13]. However, there are not many feature selection works for Amharic document classification. The major aim of the existing Amharic text classification focused on performance of text classification algorithm [14–16], but not on the feature selection method. Feature selection methods such as information gain (IG), chi-square (CHI), and document frequency (DF) can be used to overcome the curse of dimensionality by eliminating irrelevant features and selecting the most valuable features from the corpus. Better selection of terms or features leads to better classification results [3].

Therefore, the aim of this paper is to present a hybrid feature selection strategy for Amharic news document classification for improving the performance of the classifier's accuracy. The proposed feature selection method consists of IG, CHI, and DF as feature selection method, union to combine highly-ranked features, and intersection to join least-ranked features selected by IG, CHI, and DF methods.

This article is structured as follows. Related works are listed in Section 2. Our method of hybrid feature selection is defined in Section 3. The experimental findings are defined in Section 4, and we conclude the paper in Section 5.

2. Related Works

Feature selection (FS) is the process of choosing a small subset of relevant features from the original features by eliminating irrelevant, redundant, or noisy features [6]. FS is very important in pattern recognition and classification because the existing learning algorithms are not designed to deal with higher dimensional feature space. Feature selection usually leads to better learning accuracy, lower computational cost, and better model interpretability [11].

The feature selection method used in Amharic document affects the classification accuracy and computational complexity of the machine learning model. A number of research studies have attempted to use different feature selection techniques to overcome the problem of curse of dimensionality. The work of [17] is a pioneer in the automatic categorization of Amharic documents using statistical techniques. The purpose of the study is to design a prototype that automatically classifies news items from the Ethiopian News Agency (ENA) into their predefined class based on their content. The author used the term "frequency thresholding" to reduce feature space. In this type of dimension reduction, the resulting feature space may have features that are noncontent bearing and may lose content bearing features that have a lower term frequency.

The classification of Amharic news documents was carried out using Artificial Neural Network (ANN) [18]. The author tries to see the potential application of Learning Vector Quantization (LVQ) over Amharic document classification. They used the technique of single dimensionality reduction, i.e., DF feature selection method and manual term selection, which is a key word for a particular class and does not fit the given DF threshold value used in their paper. For example, in their experiment, the word "ኢትዮጵያ" (Ethiopia) appears in 46 document of bank and insurance category and satisfies DF threshold. However, they manually exclude it and take the other term "ኢንሹራንስ" (insurance) having a DF value 8 as a key word. Manual identification and rearrangement of terms might not make the system automatic. In the previous studies of Amharic text classification, only single dimension reduction technique was used to reduce the feature space, which may increase the computational cost, memory storage, and underfitting or overfitting.

A hybrid feature selection for Arabic text clustering has been proposed in paper [13]. Three separate feature selection methods are incorporated into this model, such as CHI, mutual information, and term frequency-inverse document frequency. To combine selected features with three feature selection methods, they used a union merger approach, resulting in an increase in the feature size for larger datasets.

The authors in [8] proposed a hybrid dimension reduction method for English text clustering by the combined uses of feature selection and feature extraction methods together. For feature selection, they used DF and term variance (TV) and for feature extraction, PCA. To pick the most representative words that perform best in DF and TV function scoring metrics, they used global thresholding. However, the selected features of the global threshold are influenced by the most frequently occurring news category in the document collection [19].

The authors in [12] proposed a hybrid feature selection method for Arabic text classification. Their hybrid feature selection approach incorporates DF and IG feature selection methods. In order to minimize the feature size for the classification, DF and IG thresholding was used. DF was used to eliminate rare terms, and IG was used to get the most informative term from the remaining terms. Feature selection techniques select features which are irrelevant for the classification. Therefore, feature extraction techniques are highly required to furtherly refine the feature subset [20].

The authors in [21] propose a feature selection method for Arabic text classification using an improved chi-square to enhance the performance of classification. They compared their enhanced feature selection method with three other features' selection metrics, namely, mutual information, IG, and CHI. The authors experiment their work with a dataset of 5070 Arabic documents classified into six independently classes with SVM classifier. The authors conclude their proposed method improves the performance of the Arabic text classification model. The best f-measures obtained from their model is 90.50%, when the number of features is 900. Since the number of classes has an effect on classification

accuracy, the authors did not experiment their model by varying the number of categories in the dataset.

In [22], the authors proposed a new feature selection method for Arabic text classification by using Firefly Algorithm. The authors use SVM as a text classifier. The authors use three evaluation performance measures, including precision, recall, and F-measure, for the classification accuracy. The authors conclude that their proposed method achieves a precision value equal to 0.994 and the efficiency of their proposed feature selection method in improving Arabic text classification accuracy. However, the author's test their proposed algorithm only in single Arabic dataset.

In [23], the authors propose a hybrid feature selection approach based on LSI for classification of Urdu text. For classification of Urdu text, they used the SVM classifier. They evaluate their proposed method on Urdu dataset of 29,931 news articles with 16 different categories. The authors integrate CHI, IG, and gain ratio (GR) feature selection methods with the Latent Semantic Indexing (LSI) method. They conclude their proposed approach show a better classification with promising accuracy and better efficiency. They did not evaluate their proposed method with different classifiers.

Reviewing the previous research studies of Amharic document classification, we noted that they used DF thresholding as a feature shrinking method. However, DF cannot take terms which have different characteristics at the same time. For other languages, we noted the problem with feature merging strategy, testing with datasets having different number of categories and experimenting with different classifier. We are researching a feature selection scheme consisting of the FS-FS-FS in this paper. This technique of hybrid feature selection decreases the number of features with a minimal effect on the accuracy of classification.

3. The Proposed Method

The proposed hybrid feature selection method for Amharic document classification starts with a collection of news documents. The classifier consists of five basic components. These components are preprocessing, document representation, feature selection, term weighing, and, lastly, classifier module. The preprocessing module has tokenizer, normalization, stop-word removal, and stemmer subcomponents. Detail description of each step is given below.

3.1. Preprocessing

3.1.1. Normalization. There are different characters with the same sound in the Amharic writing system which are called homophones. For instance, ሰ and ሠ, ሀ, ሐ, ኸ, and ኀ, and ጸ and ፀ are Amharic alphabet consonants having the same meaning and sound. Inconsistency in the writing of words by the characters mentioned above can be resolved by replacing the characters of the same sound in one canonical form.

3.1.2. Tokenization. It is breaking a text chunk in smaller parts, whether it is breaking paragraph into sentences, sentence into words, or word in characters [24]. The length of tokens varies from a single term to consecutive *n* terms. In this study, we use single term for the representation of documents.

3.1.3. Stop Words' Removal. In Amharic, the common words, e.g., እንደ፣ሁሉ፣እስከ፣ነው፣ላይ፣ና, and others that score less weightage in the document classification are called stop words. To decrease the dimension of the feature space and to avoid misleading or decreasing the efficiency of the classification, stop words are eliminated. Amharic does not have a well-prepared list of stop words. However, we eradicate stop words in accordance with Eyob [25].

3.1.4. Stemming. In this paper, we used HornMorpho for stemming purposes, which is a Python program developed by Michael Gasser. HornMorpho produces morphemes of a given Amharic word (meaningful portions).

3.2. Document Representation. Document representation is concerned with how textual documents for various tasks such as text classification, information retrieval, content discovery, and text mining should be portrayed [26]. We used the vector space model (VSM) in this paper, which is the simplest way of representing documents. Syntactic structure and semantic dependence between words are ignored by VSM representation.

3.3. Feature Selection. In this section, we define document frequency, information gain, chi-square, and proposed hybrid feature selection for Amharic text classification.

3.3.1. Document Frequency. Document frequency (DF) counts the number of documents in a term occurs. The fundamental concept behind DF is terms that are not relevant for the classification contained in a smaller number of documents. For the next step of text classification, words scoring a DF value greater than the threshold are used. DF is determined as [9]

$$\mathrm{DF}(ti) = \sum_{i=1}^{m} (Ai). \tag{1}$$

3.3.2. Information Gain. Information gain (IG) calculates the amount of information available for categories through recognizing the existence of absence of a given word in a document. In other words, IG measures the worthiness of features for classification, and the IG of a term t can be evaluated as follows [27]:

$$IG(t) = -\sum_{i=1}^{m} P(Ci)\log P(Ci) + P(t)\sum_{i=1}^{m} P(Ci \mid t)\log P(Ci \mid t) + P(\neg t)\sum_{i=1}^{m} P(Ci \mid \neg t)\log P(Ci \mid \neg t), \quad (2)$$

where m is the number of categories, $P(Ci)$ is the probability of the i^{th} category, $P(t)$ and $P(\neg t)$ are the probabilities of the presence and absence of term t, and $P(Ci \mid t)$ and $P(Ci \mid t)$ are the probabilities of Ci with the presence or absence of term t, respectively.

3.3.3. Chi-Square. Chi-square (CHI) test is a mathematical method used to calculate two events' level of independence. The level of independence, in this case, is between categories and terms. We use chi-square to verify whether or not a particular term and particular class are independent. The higher chi-square score value means that more likely the word is to be associated with the class. CHI (C, t) is determined mathematically as [28]

$$\mathrm{CHI}(C,t)\frac{N\ (AM - MP)(AM - MP)}{PM(N - P)(N - M)}, \quad (3)$$

where N is the total number of documents in the set, M is the number of instances containing t, P is the number of positive instances, and A is the number of positive instances containing t.

3.3.4. Proposed Hybrid Feature Selection Method (IGCHIDF) and Feature Merging. The subsets of terms generated by each method are combined to form one set that is used by the next phase of classification. As shown in Figure 1, we used IG, CHI, and DF in a hybrid manner to pick the most important terms. The feature selection methods and feature merging strategy used in IGCHIDF is shown in Figure 1.

The IGCHIDF method is given as per the following context:

(1) Implement the method of FS and retain all terms with a score greater than the value of the threshold
(2) Get set1, set2, and set3 by repeating I, for IG, CHI, and DF, respectively
(3) Sort set1, set2, and set3 by the scoring value in ascending order
(4) Set1, set2, and set3 are divided by a ratio of 75 to 25 into two sections
(5) Combine section one by intersection and section two from 4 by union
(6) Unify the intersection and union outcome from step 5

In Figure 2, the flowchart demonstrates how IGCHIDF is implemented and how its components are connected.

3.4. Term Weighting. In distinguishing one document and the other, all the words in the document are not equally significant. The significance of terms in the classification is measured using term frequency by inverse document frequency (TF*IDF) [29]. This method of term weighting can be formulated as follows:

$$\mathrm{TF} * \mathrm{IDF}(t,d) = \mathrm{TF}_{t,d} * \log\frac{N}{\mathrm{DF}_t}. \quad (4)$$

4. Experiment

All the experiments are conducted in a window 10 environment on a machine having core i7 processor and 8 GB RAM. Multilayer perceptron (MLP) classifier, support vector machine (SVM), K-nearest neighbor (KNN), and decision tree classifiers were used in our experiment. Since MLP classifier is an efficient classifier than others classifiers [30], we examined DF, IG, CHI, and IGCHIDF feature selection techniques with TF*IDF term weighting. To train the MLP classifier models, we consider the following parameters: number of hidden layers = 1; neuron number per layer is equal to ((input feature*2)/3+output classes); other default parameters. For our experiment, the dataset is collected from Ethiopian News Agency (ENA), Walta Information Center, Amhara Mass Media, and Fana Broadcasting Corporation from 2018 to 2019 (GC). A total of 2666 Amharic news documents from 13 major news categories were collected. To evaluate the classifier model, we used two datasets with different numbers of categories: dataset1 with 9 categories and dataset2 with 13 categories as presented in Tables 1 and 2.

4.1. Performance Measure. Different metrics may assess the efficacy of document classification. We used a well-known accuracy metric in our experiment, which is widely used in text mining [31].

Accuracy = TP + TN/TP + TN + FP + FN*100% (4)

where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

5. Results

We conducted experiments for Amharic news document classification with MPL, KNN, SVM, and decision tree classifiers. According to our experimental results, MPL classifier performs the other three classifiers as shown in Table 3.

We evaluate the performance of the proposed hybrid feature selection method using the MLP classifier because MPL classifier outperforms the other three classifiers used in this paper. The experiments are conducted over dataset1 and dataset2. The result is presented in Tables 4 and 5 below, respectively.

The results from Tables 4 and 5 showed that the best accuracy was obtained by the IGGHIDF feature selection method. From our experiment, we conduct the classification accuracy achieved is 90.52%, 84.93%, 88.49%, and 93.7% by IG, CHI, DF, and IGCHIDF, respectively, on dataset1. On dataset2, we obtained 85.17%, 77.52%, 81.83%, and 89.13%

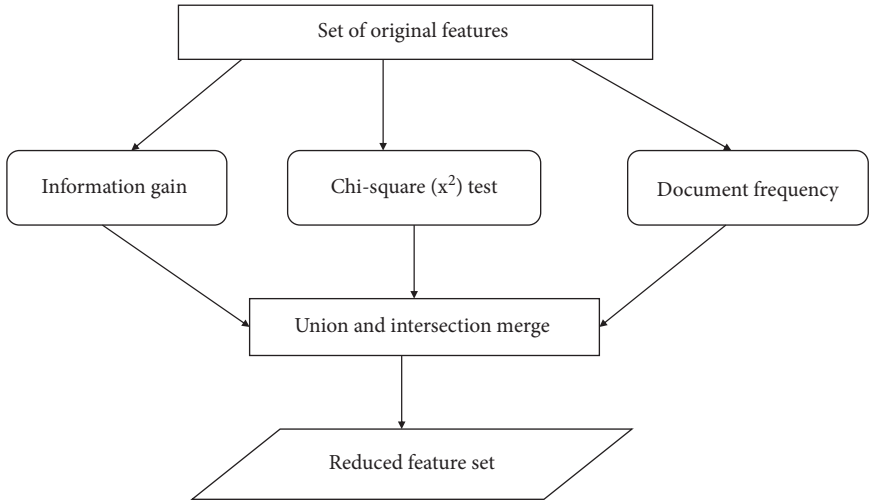

Figure 1: Pictorial description of hybrid feature selection (IGCHIDF).

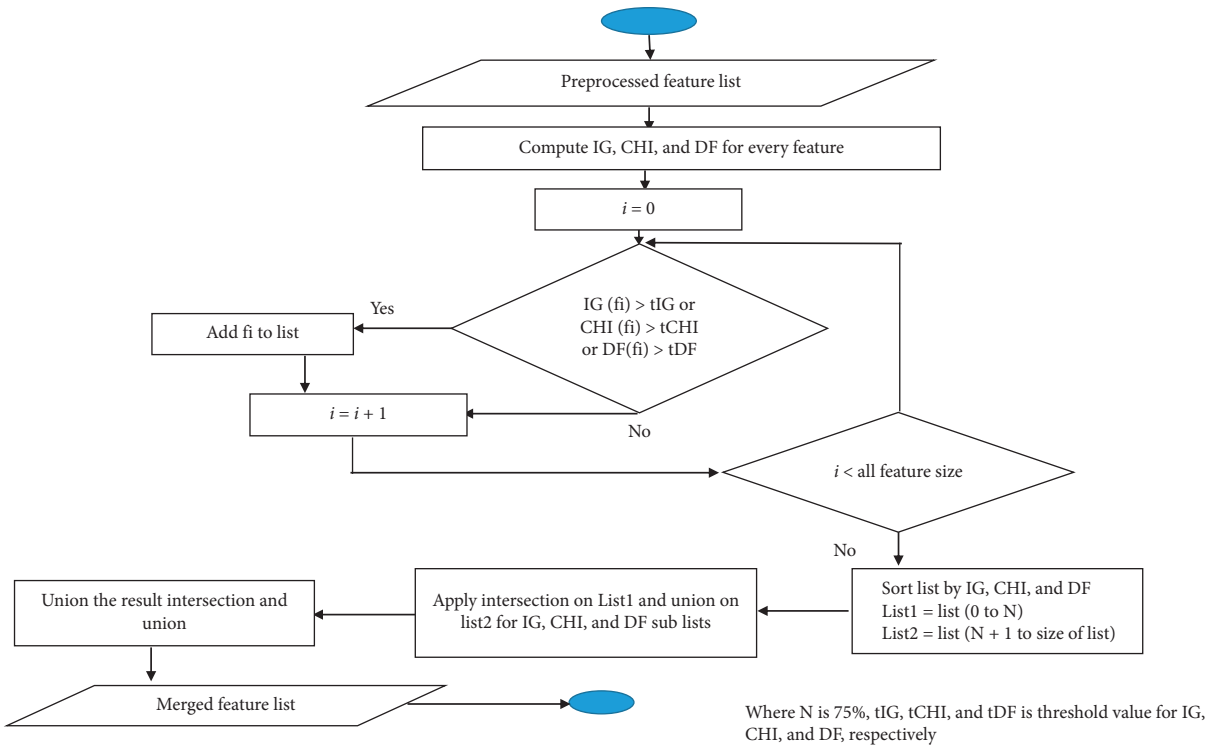

Figure 2: The flowchart for the proposed feature selection (IGCHIDF).

Table 1: The distribution of dataset1.

No.	Category	#training docs	#testing docs
1	Economy	173	44
2	Education	175	34
3	Sport	183	37
4	Tourism	162	43
5	Accident	161	46
6	Environmental	165	45
7	Diplomacy	162	43
8	Law and justice	117	28
9	Agriculture	160	45
	Summary	**1,458**	**365**

Table 2: The distribution of dataset2.

No.	Category	#training docs	#testing docs
1	Economy	169	45
2	Education	170	46
3	Sport	171	37
4	Tourism	173	36
5	Accident	166	39
6	Environmental	180	38
7	Diplomacy	167	39
8	Law and justice	111	30
9	Agriculture	150	49
10	Army	162	48
11	Technology	170	47
12	Politics	171	44
13	Health	172	36
	Summary	**2132**	**534**

Table 3: Comparisons of MLP, KNN, SVM, and decision tree classifiers.

No.	Machine learning model	Classification accuracy (%)	Dataset
1	MLP classifier	89.13	Dataset2
2	KNN classifier	67.91	
3	SVM classifier	80.3	
4	Decision tree	67	

Table 4: Comparisons of IG, CHI, DF, and IGCHIDF on dataset1 using MLP classifier.

No.	Learning model	FS method	Accuracy	Training time
1	MLP classifier	IG	90.52	29.5
2	MLP classifier	CHI	84.93	60
3	MLP classifier	DF	88.49	12.83
4	MLP classifier	IGCHIDF	93.70	14.72

Table 5: Comparisons of IG, CHI, DF, and IGCHIDF on dataset2 using MLP classifier.

No.	Learning model	FS method	Accuracy (%)	Training time (seconds)
1	MLP classifier	IG	85.17	40
2	MLP classifier	CHI	77.52	100
3	MLP classifier	DF	81.83	14.70
4	MLP classifier	IGCHIDF	89.13	16.45

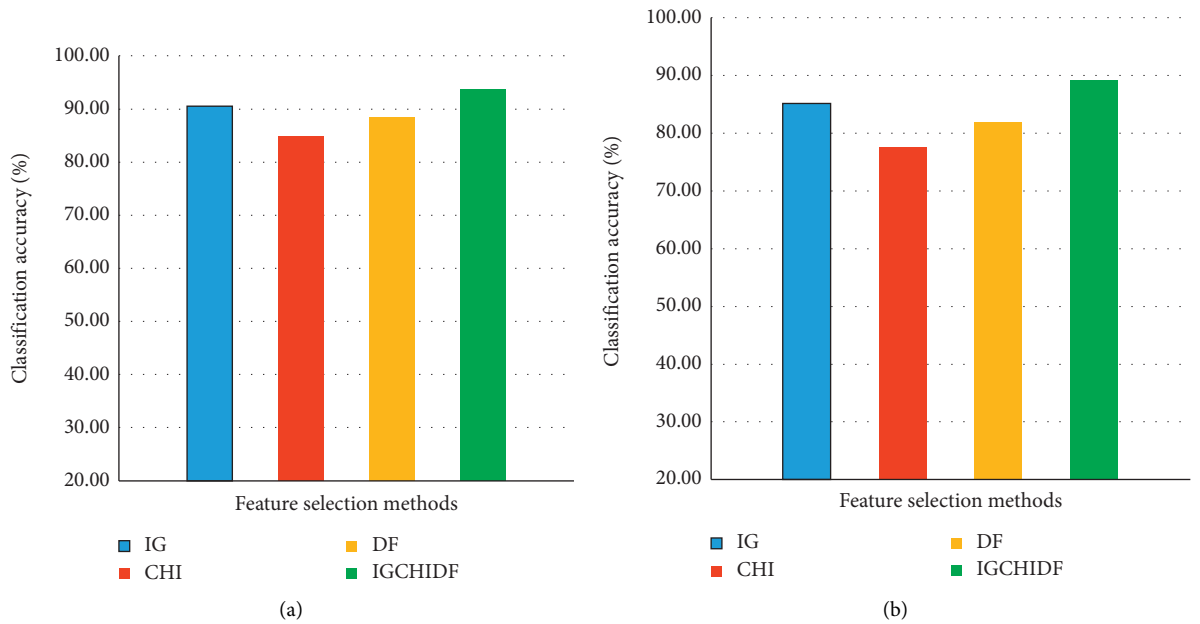

Figure 3: The comparison IGCHIDF with IG, CHI, and DF with dataset 1 and 2.

by IG, CHI, DF, and IGCHIDF, respectively. The reason behind the use of dataset 1 and dataset 2 is to see a loss of classification accuracy as the number of categories increases from nine to thirteen.

In Figure 3, the proposed method, IGCHIDF, is performing significantly better than the other methods in terms of classification accuracy. This is due to the fact that IGCHIDF uses a union merging strategy on top ranking features (terms) and intersections on the lowest ranked terms (to reduce the dimension of feature space). In our experiment, the accuracy of IGCHIDF is 3.96% higher than IG (in the case of dataset 2), 11.16% higher than CHI in the case of dataset (2), and 7.3% higher than DF.

The results show that the combination of different feature selection methods can improve the performance for Amharic text classification as they recompense the weaknesses of the individual method.

6. Conclusion

The purpose of this study was to show how the proposed feature selection method improves the classification accuracy. To validate the performance of the proposed feature selection method, several experimentations and comparisons with the state-of-the art methods are performed on two different datasets having different numbers of categories. The results show that the proposed IGCHIDF method gives promising results when combined with the MLP classifier. Therefore, the proposed feature selection method deserves to be used in different applications where Amharic document classification is required such as automatic document organization, topic extraction, and information retrieval. However, some improvements could be brought to IGCHIDF to reduce the loss in classification accuracy as the number of features and categories increases. Additional categories and datasets will be explored in our future work. The integration of features remains an interest in the future studies.

Data Availability

The dataset and the source code of this research work are publically available at GitHub (https://github.com/demekeendalie/feature-selection).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work was performed by two academic staff at Jimma Institute of Technology, Jimma University, Ethiopia. The authors would like to thank the institute for supporting through different resources. The authors would like to thank Jimma University for support during the research work.

References

[1] S. Zakaria, *Population and Housing Census of Ethoipia Central Statistical Authority*, Addis Ababa, Addis Ababa, Ethiopia, 2019.

[2] V. Korde, "Text classification and classifiers:A survey," *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 2, pp. 85–99, 2012.

[3] A. K. Uysal and S. Gunal, "Text classification using genetic algorithm oriented latent semantic features," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5938–5947, 2014.

[4] T. H. Nguyen, D. L. Tuan, H. N. Nguyen, and T. N. Vu, "A hybrid feature selection method for Vietnamese text classification," in *Proceedings of the Seventh International Conference on Knowledge and Systems Engineering*, Ho Chi Minh city, Vietnam, October 2015.

[5] K. U. Alper and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Systems*, vol. 36, pp. 226–235, 2012.

[6] N. J. Nilsson, *Introduction to Machine Learning*, Stanford University, Kerala, India, 1998.

[7] M. Rostami, S. Forouzandeh, K. Berahmand, and M. Soltani, "Integration of multi-objective PSO based feature selection and node centrality for medical datasets," *Genomics*, vol. 112, no. 6, pp. 4370–4384, 2021.

[8] K. K. Bharti and P. K. Singh, "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3105–3114, April 2015.

[9] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[10] M. Lu, "Embedded feature selection accounting for unknown data heterogeneity," *Expert Systems With Applications*, vol. 119, 2018.

[11] D. A. Said, *Dimensionality Reduction Techniques for Enhancing Automatic Text Categorization*, Faculty of Engineering at Cairo University Master of science, Cairo, Egypt, 2007.

[12] A. A. Mena Badieh Habib, "A hybrid feature selection approach for Arabic documents classification," *Egyptian Computer Science Journal*, vol. 28, no. 3, pp. 1–7, 2006.

[13] H. Alghamdi, "The hybrid feature selection k-means method for Arabic webpage classification," *Jurnal Teknologi*, vol. 70, no. 5, pp. 73–79, 2014.

[14] A. w. Yohannes, *Automatic Amharic Text Categorization Using Support Vector Machine Approach*, Addis Ababa University, Addis Ababa, Ethiopia, 2007.

[15] T. Surafel, *Automatic Categorization of Amharic News Text: A Machine Learning Approach*, Addis Ababa University, Addis Ababa, Ethiopia, 2003.

[16] A. Hilu, *Amharic Document Categorization Using Itemsets Method, Addis Ababa*, Addis Ababa University, Addis Ababa, Ethiopia, 2013.

[17] Z. Sntayehu, *Automatic Classification of Amharic News Items: The Case of Ethiopian News Agency*, Addis Ababa University, Addis Ababa, Ethiopia, 2001.

[18] W. Kelemework, "Automatic Amharic text news classification: aneural networks approach," *Ethiopian Journal of Science and Technology*, vol. 6, no. 2, pp. 127–137, 2013.

[19] F. Balabanian, E. Sant'Ana da Silva, and H. Pedrini, "Image thresholding improved by global optimization methods," *Applied Artificial Intelligence*, vol. 31, no. 3, 2017.

[20] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019.

[21] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text

classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, 2020.

[22] S. Larabi Marie-Sainte and N. Alalyani, "Firefly algorithm based feature selection for Arabic text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 3, pp. 320–328, 2020.

[23] I. Rasheed, H. Banka, and H. M. Khan, "A hybrid feature selection approach based on LSI for classification of Urdu text," *Studies in Computational Intelligence*, vol. 907, pp. 3–18, 2020.

[24] E. Pimentel and E. Boulianne, "Special issue:blockchain Technology: promises and perils," *Jornal of Corporate Accounting and Finance*, vol. 31, no. 2, pp. 1–73, 2020.

[25] D. Y. Eyob and E. D. Dejene, "Topic-based Amharic text summarization with probabilistic latent semantic analysis," in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, Tokyo, Japan, September 2012.

[26] J. Sang, S. Pang, Y. Zha et al., "Design and analysis of a general vector space model for data classification in Internet of Things," *Journal of Wireless Networking and Communications*, vol. 263, 2019.

[27] Tehseen Zia, M. P. Akhter, and Q. Abbas, "Comparative study of feature selection approaches for Urdu text categorization," *Computer Science Malaysian Journal of Computer Science*, vol. 28, no. 2, pp. 93–109, 2015.

[28] M. Zhu, W. Xiong, and Yi-F. B. Wu, "Learning to rank with only positive examples," in *Proceedings of the 2014 13th International Conference on Machine Learning and Application ICMLA '14*, Detroit, MI USA, December 2014.

[29] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 30, no. 9, 2019.

[30] A. Patra and D. Singh, "Neural Network approach for text classification using relevance factor as term weighing method," *International Journal of Computer Applications*, vol. 68, no. 17, pp. 37–41, 2013.

[31] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, and W. Meira, "Word co-occurrence features for text classification," *Information Systems*, vol. 36, no. 5, pp. 843–858, 2011.