

Evaluating the influence of feature selection-based dimensionality reduction on sentiment analysis

Gowrav Ramesh Babu Kishore, Bukahally Somashekar Harish, Chaluvegowda Kanakalakshmi Roopa

Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, India

Article Info

Article history:

Received Apr 2, 2024

Revised Mar 21, 2025

Accepted Jun 8, 2025

Keywords:

Sentiment analysis

Pre-processing

Dimensionality reduction

Feature selection

Classification

ABSTRACT

As social media has become an integral part of digital medium, the usage of the same has increased multi-fold in recent years. With increase in usage, the sentiment analysis of such data has emerged as one of the most sought research domains. At the same time, social media texts are known to pose variety of challenges during the analysis, thus making pre-processing one of the important steps. The aim of this work is to perform sentiment analysis on social media text, while handling the noise effectively in the data. This study is performed on a multi-class twitter sentiment dataset. Firstly, we apply several text cleaning techniques in order to eliminate noise and redundancy in the data. In addition, we examine the influence of regularized locality preserving indexing (RLPI) technique combined with the well-known word weighting methods. The findings obtained from experiment indicate that, RLPI outperforms other algorithms in feature selection and when paired with long short-term memory (LSTM), the combination outperforms other classification models that are discussed.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Chaluvegowda Kanakalakshmi Roopa

Department of Information Science and Engineering, JSS Science and Technology University

Mysuru, 570006, Karnataka, India

Email: ckr@jssstuniv.in

1. INTRODUCTION

Now-a-days social media has garnered a lot of attention. It is a multimedia platform where people can share or consume information in any format that they want, be it image, video, audio, or text. Thanks to its instantaneous global accessibility, it has become a vital part of digital media. As people started using social media in large numbers, the need to analyze the same became necessary. The analysis started taking place on all possible aspects. If one section of research community focused on the optimal use of computing resources, the other section focused on the effective information retrieval techniques for the same.

One of the trending areas in information retrieval is sentiment analysis, where the given data is analyzed in order to obtain the intended opinion or emotion. There are many ways to express sentiments. The most popular methods to categorize them is either based on polarity or based on emotion. When it comes to polarity, the sentiments might be one among positive, negative, or neutral. Such labelling is best suited when the aim of the analysis is to get the inference only at higher level. On the other hand, for emotion, there is wide range of terms to express, such as happy, sad, sarcastic, ironic, and metaphorical; and such sentiment labelling works best when the analysis calls for the inference of particular opinion.

In recent years, sentiment analysis on the social media text has gained a lot of momentum. Whether it is analyzing amazon reviews for market research, or analyzing tweets to gauge audience sentiment, the research is being conducted on all conceivable fronts. Although social media is widely recognized as a valuable data source, the text data collected from these platforms can have a number of issues. Issues like

emojis, hashtags, emojis, mimicking spoken word prolongations, misspellings, and special characters occasionally cause noise in the data, thus making it difficult to process it directly. Processing and analyzing social media texts can also be difficult because of their non-uniform nature, as they don't always adhere to linguistic norms. Usually, such issues are not encountered in other standard sources, such as newspapers or e-books, as they adhere to language standards. Therefore, pre-processing steps such as text cleaning or dimensionality reduction becomes necessary, in order to handle superfluous or high-dimensional social media text data [1]. Additionally, it is critical that the model be able to understand the context and sentiment from short-texts, as social media platforms also impose limits on the number of words.

Pre-processing of a text involves several important steps, where with each step the least important part of the data is dropped. Sometimes, more than the analysis, pre-processing itself takes more time [2]. The removal of special symbols and stop-words reduces the dimensionality in the term space [3]. However, certain cleaning procedures do not require the complete removal of the term from the data, as for example, lemmatization and stemming merely require the term to be reduced to its basic forms. It also greatly aids in removing redundancy and noise, so that only the most important components are left for further analysis. Often, even after cleaning the input text, the final corpus size will surpass the processing capability of the system. So, in order to reduce the dimensionality of input furthermore, feature engineering is performed. Feature engineering techniques are used mainly to extract or select most relevant set of features. In case of text, the first and foremost task is features extraction, where the text is represented in machine understandable numerical form. Subsequently, feature selection is employed to isolate the most significant features, whose contribution is more during the classification.

Figure 1 shows the categories of dimensionality reduction techniques. Generally, in feature extraction, the original set of features is transformed to get a lesser number of meaningful and relevant feature set. Some of the well-known algorithms are principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). In feature selection, subsets of features are selected from the original set of features by eliminating the redundant or irrelevant ones. Some well-known methods are recursive feature elimination (RFE), correlation and mutual information-based algorithms.

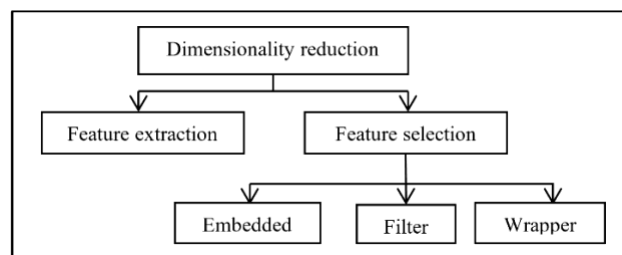


Figure 1. Categories of dimensionality reduction techniques

In this paper, we evaluate the performance of feature selection techniques when paired with regularized locality preserving indexing (RLPI) algorithm. Also, examine the behavior of the selected set of features with various neural network-based classification models. The purpose of this study is to gain a deeper understanding on using various pre-processing techniques in combination with RLPI dimensionality reduction technique that affect the performance sentiment classification. The primary focus of this research is on feature selection approaches and their effect on sentiment text classification performance. The following discussion provides some initial insights on the prominent feature selection techniques and their impact on sentiment classification.

Term frequency-inverse document frequency (TF-IDF) is one of the well-known feature extraction methods, where it is generally used for extracting numerical features out of text data [4]. However, Patil and Atique [5] shows how feature selection can be implemented with TF-IDF, by adding threshold parameters to the terms in order to select the key terms. While Qu *et al.* [6] proposed an improved TF-IDF approach by including document's relation with multi-class information, and based on the weights obtained, the top K vocabulary terms for each document are identified. Li *et al.* [7] applied regularized least squares-multi angle regression and shrinkage (RLS-MARS) model to determine the least significant features. The proposed method assigns less weight to the least significant features. According to Wang and Zhang [8], a feature selection method is presented based on TF-IDF by combining it with Kullback–Leibler (KL) divergence, whereby considering the mutual information as the criterion, the authors proposed an improved classification approach. Song *et al.* [9] introduced an entropy index along with TF-IDF in order to get the

entropy information of a term with-in and among the classes, which will then be used for text feature selection. Nafis and Awang [10] proposes a two stage feature selection approach. Where in first stage, the variance obtained for entire TF-IDF matrix is used as threshold to select the features. Then in second stage, the support vector machine (SVM)-RFE is applied on the new feature set to re-evaluate the features.

The authors in [11] proposed a feature selection method based on the combination of information gain and divergence for text categorization models based on statistics, where it chooses every feature based on a combination of information gain and novelty criteria resulting in reduced redundancy among the selected features. The behavior of the information gain-based feature selection method combined with the genetic algorithm is demonstrated in [12], demonstrating the method that lowers the text vector's dimension. Shang *et al.* [13] proposed a maximizing global information gain approach, which is an enhanced version of information gain algorithm. Along with avoiding the redundancy in the features, global information gain metric is said to be more informative, distinctive and also perform faster when compared to the traditional information gain. Pereira *et al.* [14] discusses the performance of information gain based feature selection, and compares the same against other multi-label feature selection methods. Omuya *et al.* [15] proposes a hybrid dimensionality reduction technique that uses information gain and PCA to extract and choose relevant features. The approach's effectiveness was assessed against the naive Bayes model, where the training time is shortened while enhancing performance.

The chi-square test is one of the widely used statistical functions and the work in [4] demonstrates the use of chi-square test for feature selection, along with K-nearest neighbor (KNN) as the classification algorithm. On the other hand, Zhai *et al.* [16] shows its ability to effectively select the better performing set of features than the information gain algorithm. Jin *et al.* [17] proposes an enhanced version of chi-square statistics approach called as term frequency and distribution based CHI for feature selection in order to address the inability of the original approach to consider and identify the term distribution in each class. Li [18] proposed an enhanced version of chi-square approach based on Chi-square rank correlation factorization where it is claimed that the algorithm does not need any prior knowledge and can offer generalized text categorization. Haryanto *et al.* [19] show the behavior of SVM classifier upon feeding the inputs which are normalized and features are selected using the chi-square approach.

Sel *et al.* [20] presents the feature selection method, which is performed based on the mutual information, thus showing the effectiveness of the approach in improving the classification performance despite of drastic reduction in the number of features. Liu *et al.* [21] proposes a dynamic mutual information algorithm by introducing a general criterion function for feature selection, which is expected to get most information measurements in previous algorithms together and was evaluated against various existing methods. Agnihotri *et al.* [22] demonstrate use of the mutual information to obtain the sample variance in order to measure the variations in term distribution and to select the features. Meanwhile, Ding and Tang [23] presents an enhanced mutual information method by introducing the feature frequency in class and the dispersion of feature in class, leading to an efficient and improved text categorization. While Darshan *et al.* [24] shows the ability of RLPI to effectively extract the discriminative features, which in turn reduces the complexity during the representation thus by reducing the total number of final feature set. Revanasiddappa *et al.* [25] proposed a framework based on meta-cognitive neural network constituting RLPI, where RLPI is used along with term document matrix (TDM) as feature selection approach in order to reduce the dimensionality.

The rest of the paper is organized as follows: in section 3, details regarding the dataset considered for the experiment, text cleaning and feature selection techniques that are employed during the pre-processing stage, details on the classification models used, followed by the working principle of the experiment. Section 4 presents the experiment results along with discussion. Finally, section 5 concludes the work along with future scope.

2. METHOD

Since the study focuses on text-based sentiment analysis, there are steps in the process that must be completed in order to clean the data, reduce its dimensionality, and get it ready for training. This section covers the specifics of the dataset that was used, as well as the approaches employed for each stage.

2.1. Dataset

For this study we use a twitter dataset, which is created by combining 2 datasets which were earlier separate. Originally, the differentiating factor between the two datasets was their labelling. One dataset with 1.6 million samples were labelled based on polarity, while the other dataset with about 98,000 tweet samples were labelled based on feelings such as sarcasm, figurative, irony, and regular. The final dataset consists of 97,000 samples, where they are categorized among 5 sentiment classes namely positive, negative, neutral,

sarcasm, and figurative. While creating the final dataset, samples were randomly selected such that each category contains samples ranging from 15,000 to 20,000 tweets.

2.2. Text cleaning methods

This is a crucial and widely employed stage in text-based research, since it facilitates the extraction of useful information from textual data. In this study we employed many text cleaning procedures, and they are as follows:

- Text casing: the same word can be perceived as a single token by changing its case which will otherwise be considered as a different token, such standardization of case helps prevent redundancy in the original corpus. The majority of the time, the text is changed to lower case and the same is followed during this study as well.
- Removing punctuation: depending on the design and final goal of the model, punctuations that are often used to indicate separate sentences or the end of sentences such as commas, periods, and semicolons are preserved or dropped. Since we are concentrating more on the tokens in this instance, the punctuations are dropped.
- Removing special symbols: since the study is primarily focused on preserving only the important tokens, as previously noted, any characters other than alphanumeric such as ampersand, dollar, pipe, and percentage. that are known to be often used in Twitter posts, are excluded.
- Removing stop words: from a non-linguistic point of view, stop-words don't carry much information [5] hence removing them will not only help in reducing the noise, but it also helps in saving space. Stop-words can be identified and dropped using both manual and automatic approach.
- Stemming or Lemmatization: this is the processes of reducing the words to their root form. It was noticed that lemmatization helps better when compared to stemming in giving the meaningful root form. Example; while stemming reduces 'studies' is reduced 'studi', lemmatization reduces the same to 'study', and hence in the work lemmatization is applied on the text samples.
- Handling emojis: emojis can be handled in a number of ways, either by removing them completely or substituting them with their text equivalent. In this study, emoticons are omitted.
- Handling word contractions: in this action, we convert the combined short forms of words back to their original forms. Example: 'don't' is converted to 'do not'. This can also be achieved in both manual and automated ways.
- Spell checking: checking the spelling of the token is equally important as lemmatization, it helps in avoiding unnecessary additional tokens that may be present due to some wrong spellings.

2.3. Feature selection methods

As conveyed in the beginning, since this work is mainly focused on the feature selection approach for dimensionality reduction, it is very important to know more about the approaches that are there for feature selection. It is mainly classified into 3 types namely, filter method, wrapper method and embedded method. In this study, we restrict the experiment to filter and wrapper methods.

In filter method, the features are selected using statistical tests in order to get the correlation scores. They are known to be inexpensive and fast and some of the techniques used under this method are:

- TF-IDF: a way of calculating a word's weight within a collection of documents, taking into account the fact that some terms are more common than others. The weight is calculated using (1):

$$W_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right) \quad (1)$$

Where $tf_{x,y}$ is frequency of x in y, df_x is Number of documents containing x and N is the total number of documents.

- Chi-square test: this measure [21] is used to identify the degree of independence between the term t_i and class C_k , and it is given in (2)

$$\chi^2 = N * \frac{(a*d-b*c)}{(a+c)(b+d)(a+b)(c+d)} \quad (2)$$

Where a is the number of documents in the positive category that contain this term (t_i); b is the number of documents in the positive category that do not contain this term (t_i); c is the number of documents in the negative category that contain this term (t_i); and d is the number of documents in the negative category that do not contain this term (t_i); and N is the total number of documents.

- Information gain: the information gain [26] provides the dependency between a term and a class and is given as (3). Where a, b, c, d, and N mean the same as in (2).

$$ig = \frac{a}{N} * \log \frac{a*N}{(a+c)*(a+b)} + \frac{b}{N} * \log \frac{b*N}{(b+d)*(a+b)} + \frac{c}{N} * \log \frac{c*N}{(a+c)*(c+d)} + \frac{d}{N} * \log \frac{d*N}{(b+d)*(c+d)} \quad (3)$$

- Mutual information: it is a maximum class-based score for the term t_i which is highly influenced by the marginal probabilities, that assigns higher weight for the rare terms as compared to the commonly occurring term. The metric helps in measuring the information contained by the term t_i to represent the class C and it is given as [22].

$$MI(t_i) = \max_{i \leq j \leq r} \log \frac{p(t_i, C_j)}{p(t_i) * p(C_j)} \quad (4)$$

Where $p(t_i)$ is the probability of the word t_i which is $(a + b)/N$, $p(C_j)$ is the probability of class given as $(a + c)/N$ and $p(t_i, C_j)$ is the probability of the word t_i for being in class C_j which is given by a/N .

- RLPI: is a multistep algorithm applied in order to get the meaningful set of features, which involves adjacency graph construction, Eigen decomposition and regularized least square. RLPI embedding is given as [27].

$$x \rightarrow z = A^T x \quad (5)$$

Where z is a d -dimensional representation of the document x and A is the transformation matrix.

- Word embeddings: word embedding is a representation method, where a particular term is represented in the form of a numerical vector. In this study RLPI is incorporated with some of the well-known word embedding methods for feature selection in an attempt to reduce the dimensionality of the original feature vectors.

In the wrapper method, the model is trained using a subset of features, and feature additions and deletions are determined by the conclusions derived from the results obtained. One such technique considered for the study is, RFE. It is one of the computationally expensive techniques, due to its greedy approach. In this technique, the model is trained iteratively with a subset of features until all the features are exhausted, ultimately identifying the best performing set of features.

2.4. Classification methods

In this study, sentiment classification is performed with some of the widely known neural network-based models. We assess the classification performance of both basic and recurrent neural networks (RNN) based models. Firstly, the classification performance of basic feed forward neural network (FNN) model is assessed. Because of their non-cyclic information flow, FNNs are highly straightforward and easier to verify [28]. Then, the behavior of radial basis function network (RBFN) is evaluated against the selected set of features. It is widely used for common approximation problems, where hidden layer will use the radial basis function. It is much faster when compared to back propagation network, and can even outperform the classification performance if the proper set of features are selected [29].

We then examine the classification performance of models that are designed for sequential or time series data. Firstly, the classification performance of RNN is evaluated. Though it is bit slower than basic FNNs, its ability to retain information about a sequence in hidden layers makes it most suitable for processing sequential data such as text. However, the vanishing gradient issue in their memory state limits their ability to retain only short window of the prior inputs. In order to handle this issue, long short-term memory (LSTM) was introduced. One big advantage of LSTM is its relative insensitivity to gap length, so the classification performance of LSTM is also evaluated against the selected feature set. Finally, we evaluate the performance of gated recurrent units (GRU). It is also an RNN based network and an alternative to LSTM. But GRU's fundamental principle is to update the network's hidden state only on a chosen subset of time steps, by means of gating methods. It is simpler in structure and easier to train than LSTM.

2.5. Experimentation

The experiment set-up starts with twitter sentiment data being considered as an input to the classification system, which will first undergo the pre-processing with the methods that are discussed in the section 2.2. Figure 2 presents the flow diagram, where the input data first undergoes cleaning, followed by the dimensionality reduction. For dimensionality reduction, first in order to obtain locality information, the RLPI is applied on the samples, which is then coupled with the feature selection techniques covered in section 2.3 of this work. The resulting set of relevant features from the respective combination is then used for training the model. For classification, most commonly known neural network based models viz., FNN, RNN, RBFN, LSTM, and GRU are used. Upon obtaining the classification results, the effectiveness of

each of the dimensionality reduction techniques and classification performance of the models are evaluated and analyzed.

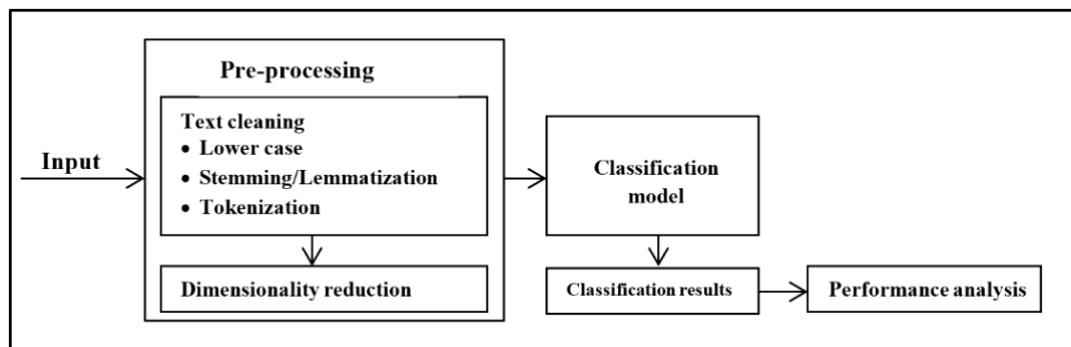


Figure 2. Workflow of text-based sentiment analysis

3. RESULTS AND DISCUSSION

During the dimensionality reduction stage of the experiment, the feature selection was performed for several iterations as seen in Table 1. During this study, the upper and lower limits were defined to obtain the most relevant set of features. With minimum of 300 and maximum of 700 being the empirically defined standard thresholds for the number of features, the experiments were carried out for each combination of feature selection methods. Table 1 shows the outcomes of each trial. It can be observed from the table, that the RLPI has selected an interestingly less number of features in each trial when compared to other methods. Figure 3 is showing the range of features by using maximum and minimum count as the extremes to indicate the count of features selected by each of the approaches mentioned in Table 1.

Table 1. Number of features selected by various selection methods

Feature selection methods	Number of features selected					
	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
TF-IDF	575	538	357	399	412	419
Chi-square	600	562	552	457	547	552
Information gain	549	552	656	453	479	490
Mutual information	427	477	479	360	380	411
Word2Vec	340	342	341	353	361	379
Glove	435	415	426	421	405	445
RFE	494	412	485	433	530	540
RLPI	69	58	49	93	100	210

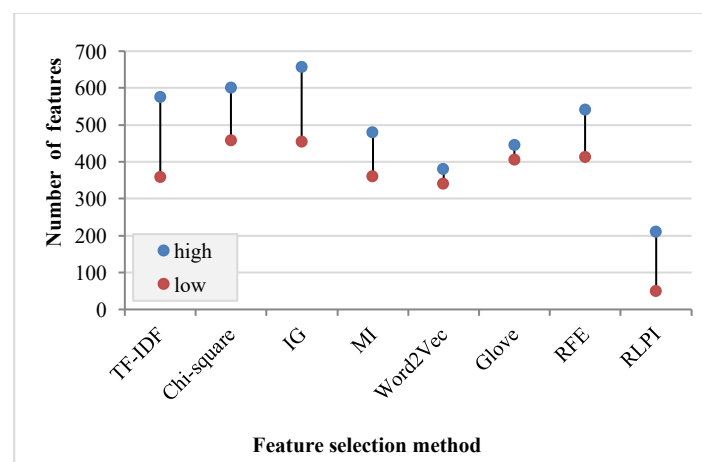


Figure 3. Max and min number of features selected by each method

Upon selecting the minimum set of features among each trial of each feature selection methods, the selected features sets are then considered as inputs to the classification models that are discussed in section 2.4. Table 2 presents the classification results of various feature selection methods and neural network-based classifiers. The results are tabulated for the dataset divided with 50:50 ratios for training and testing respectively. Table 3 presents the results experimented on same set of feature selection and classification models while the results are tabulated for the dataset divided with 60:40 ratio for training and testing respectively.

Table 2. Classification performance for 50:50 ratio of dataset partition

Classification method	Feature selection methods and their feature count							
	TF-IDF 357 features	Chi-square 457 features	IG 453 features	MI 360 features	W2V 340 features	Glove 405 features	RFE 412 features	RLPI 49 features
FNN	83.42	84.92	84.18	84.96	86.23	85.85	85.42	86.98
RNN	84.62	84.45	84.06	85.28	86.66	85.31	85.25	87.58
RBF-NN	83.62	83.28	84.94	85.46	86.10	86.26	85.16	86.72
GRU	86.03	85.69	85.85	87.17	86.86	86.80	86.50	87.17
LSTM	87.43	86.40	87.24	86.00	88.26	87.99	87.93	88.89

Table 3. Classification performance for 60:40 ratio of dataset partition

Classification method	Feature selection methods and their feature count							
	TF-IDF 357 features	Chi-square 457 features	IG 453 features	MI 360 features	W2V 340 features	Glove 405 features	RFE 412 features	RLPI 49 features
FNN	85.85	85.13	86.28	86.19	87.61	87.28	86.31	88.74
RNN	86.12	86.38	86.46	87.11	88.66	87.53	86.99	88.93
RBF-NN	85.25	86.57	84.62	84.12	87.94	87.16	86.77	88.59
GRU	86.48	88.30	86.81	88.65	90.31	89.36	88.06	90.91
LSTM	88.22	91.06	90.78	90.17	91.97	91.92	91.81	92.43

Firstly, the observations in Tables 2 and 3 show the behavior of each classification model with various set of features from different feature selection methods. It can be seen that the performance of the classification models is better when paired with RLPI, despite selecting least number of features in a set. It demonstrates that the RLPI can choose the most distinctive and pertinent features, while keeping the feature count low.

It can also be seen from the above observations that irrespective of number of features, LSTM is consistently performing better than other classification models. Finally, from the observation, it can be noted that the RLPI and LSTM combination is outperforming other combinations irrespective of train-test split ratios. The results also confirm the fact that in order to handle sequential data such as text as in this case, LSTM is best suited option.

4. CONCLUSION

In this work, we analyze the influence of pre-processing techniques. Mainly, the feature selection stage which is intended for reducing the dimensionality, on the overall classification performance. During the experiment, RLPI was incorporated along with various feature selection techniques in order to obtain the least number of most relevant and distinctive set of features. The classification performances of neural network-based models are evaluated against minimum feature sets, which are obtained by different feature selection methods. Results show that the combination of RLPI in its simplest form and LSTM outperform all the other combinations in both feature selection and sentiment classification respectively. The results once again affirm the fact that the LSTM is one among the best suited models for handling sequential data. It was observed that, the variance between minimum and maximum number of features was almost same in each feature selection approaches. Sentiment classification would benefit more from an enhanced method for obtaining the ideal number of features while keeping the most relevant terms. A better dimensionality reduction method is also needed, which can lower the final dimensionality of features while maintaining context.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Gowrav Ramesh Babu Kishore	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Bukahally Somashekar Harish		✓	✓	✓	✓				✓	✓		✓	✓	
Chaluvegowda Kanakalakshmi Roopa				✓	✓	✓	✓			✓		✓	✓	

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflicts of interest.

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created in this study.




REFERENCES

- [1] V. Dogra, A. Singh, S. Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib, "Understanding of data preprocessing for dimensionality reduction using feature selection techniques in text classification," in *Intelligent Computing and Innovation on Data Science*, Singapore: Springer, 2021, pp. 455–464, doi: 10.1007/978-981-16-3153-5_48.
- [2] M. Anandarajan, C. Hill, and T. Nolan, *Practical text analytics: maximizing the value of text data*. Cham: Springer, 2019, doi: 10.1007/978-3-319-95663-3.
- [3] S. Vijayarani, J. Ilamathi, and Nithya, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.
- [4] Y. D. Kirana and S. Al Faraby, "Sentiment analysis of beauty product reviews using the K-nearest neighbor (KNN) and TF-IDF methods with chi-square feature selection," *Journal of Data Science and Its Applications*, vol. 4, no. 1, pp. 31–42, 2021, doi: 10.34818/JDSA.2021.4.71.
- [5] L. H. Patil and M. Atique, "A novel approach for feature selection method TF-IDF in document clustering," in *2013 3rd IEEE International Advance Computing Conference (IACC)*, 2013, pp. 858–862, doi: 10.1109/IAdCC.2013.6514339.
- [6] S. Qu, S. Wang, and Y. Zou, "Improvement of text feature selection method based on TFIDF," in *2008 International Seminar on Future Information Technology and Management Engineering*, 2008, pp. 79–81, doi: 10.1109/FITME.2008.25.
- [7] X. Li, H. Dai, and M. Wang, "Two-stage feature selection method for text classification," in *2009 International Conference on Multimedia Information Networking and Security*, 2009, pp. 234–238, doi: 10.1109/MINES.2009.127.
- [8] B. Wang and S. Zhang, "A novel feature selection algorithm for text classification based on TFIDF-weight and KL-divergence," in *Proceedings of the 11th Joint International Computer Conference*, 2005, pp. 438–441, doi: 10.1142/9789812701534_0099.
- [9] J. Song, M. Xu, and C. Fan, "A text feature selection method using TFIDF based on entropy," in *Computational Intelligence*, 2010, pp. 962–967, doi: 10.1142/9789814324700_0147.
- [10] N. S. M. Nafis and S. Awang, "An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification," *IEEE Access*, vol. 9, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.
- [11] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Information Processing and Management*, vol. 42, no. 1, pp. 155–165, 2006, doi: 10.1016/j.ipm.2004.08.006.
- [12] S. Lei, "A feature selection method based on information gain and genetic algorithm," in *2012 International Conference on Computer Science and Electronics Engineering*, 2012, pp. 355–358, doi: 10.1109/ICCSEE.2012.97.
- [13] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, "Feature selection via maximizing global information gain for text classification," *Knowledge-Based Systems*, vol. 54, pp. 298–309, 2013, doi: 10.1016/j.knosys.2013.09.019.
- [14] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann, "Information gain feature selection for multi-label classification," *Journal of Information and Data Management*, vol. 6, no. 1, pp. 48–48, 2015.
- [15] E. O. Omuya, G. O. Okeyo, and M. W. Kimwele, "Feature selection for classification using principal component analysis and information gain," *Expert Systems with Applications*, vol. 174, 2021, doi: 10.1016/j.eswa.2021.114765.
- [16] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A chi-square statistics based feature selection method in text classification," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, 2018, pp. 160–163, doi: 10.1109/ICSESS.2018.8663882.
- [17] C. Jin et al., "Chi-square statistics feature selection based on term frequency and distribution for text categorization," *IETE Journal of Research*, vol. 61, no. 4, pp. 351–362, 2015, doi: 10.1080/03772063.2015.1021385.




- [18] Y. H. Li, "Text feature selection algorithm based on chi-square rank correlation factorization," *Journal of Interdisciplinary Mathematics*, vol. 20, no. 1, pp. 153–160, 2017, doi: 10.1080/09720502.2016.1259769.
- [19] A. W. Haryanto, E. K. Mawardi, and Muljono, "Influence of word normalization and chi-squared feature selection on support vector machine (SVM) text classification," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 229–233, doi: 10.1109/ISEMANTIC.2018.8549748.
- [20] İ. Sel, A. Karci, and D. Hanbay, "Feature selection for text classification using mutual information," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1–4, doi: 10.1109/IDAP.2019.8875927.
- [21] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, no. 7, pp. 1330–1339, 2009, doi: 10.1016/j.patcog.2008.10.028.
- [22] D. Agnihotri, K. Verma, and P. Tripathi, "Mutual information using sample variance for text feature selection," in *Proceedings of the 3rd International Conference on Communication and Information Processing*, 2017, pp. 39–44, doi: 10.1145/3162957.3163054.
- [23] X. Ding and Y. Tang, "Improved mutual information method for text feature selection," in *2013 8th International Conference on Computer Science & Education*, 2013, pp. 163–166, doi: 10.1109/ICCSE.2013.6553903.
- [24] H. K. Darshan, A. R. Shankar, B. S. Harish, and K. H. M. Kumar, "Exploiting RLPI for sentiment analysis on movie reviews," *Journal of Advances in Information Technology*, vol. 10, no. 1, pp. 14–19, 2019, doi: 10.12720/jait.10.1.14-19.
- [25] M. B. Revanasiddappa, B. S. Harish, and S. V. A. Kumar, "Meta-cognitive neural network based sequential learning framework for text categorization," *Procedia Computer Science*, vol. 132, pp. 1503–1511, 2018, doi: 10.1016/j.procs.2018.05.104.
- [26] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721–735, 2009, doi: 10.1109/TPAMI.2008.110.
- [27] D. Cai, X. He, W. V. Zhang, and J. Han, "Regularized locality preserving indexing via spectral regression," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 741–750, doi: 10.1145/1321440.1321544.
- [28] I. Mokriš and L. Skovajsová, "Feed-forward and self-organizing neural networks for text document retrieval," *Acta Electrotechnica et Informatica*, vol. 8, no. 2, pp. 3–10, 2008.
- [29] Z. Wang, Y. He, and M. Jiang, "A comparison among three neural networks for text classification," in *2006 8th international Conference on Signal Processing*, 2006, doi: 10.1109/ICOSP.2006.345923.

BIOGRAPHIES OF AUTHORS






Gowrav Ramesh Babu Kishore    received his B.E. degree in information science and engineering from Maharaja Institute of Technology, Mysuru, India. and M.Tech. degree in data science from the Department of Information Science and Engineering, JSS Science and Technology University, India. Presently he is a research scholar in the Department of Information Science and Engineering, JSS Science and Technology University, India. He can be contacted at email: kkishorkumar12@gmail.com or kishore_gr@jssstuniv.in.



Bukahally Somashekar Harish    obtained his Ph.D. in computer science from University of Mysore, India. Presently he is working as a Professor in the Department of Information Science and Engineering, JSS Science and Technology University, India. He was a visiting researcher at DIBRIS - Department of Informatics, Bio Engineering, Robotics and System Engineering, University of Genova, Italy. He has been invited as a resource person to deliver various technical talks on data mining, image processing, pattern recognition, and soft computing. He is serving as a reviewer for international conferences and journals. He has published articles in more than 100+ international reputed peer reviewed journals and conferences proceedings. He successfully executed AICTE-RPS project, which was sanctioned by AICTE, Government of India. His area of interest includes machine learning, text mining, and computational intelligence. He can be contacted at email: bharish@jssstuniv.in.



Chaluvegowda Kanakalakshmi Roopa    received her B.E. degree in information science and engineering and M.Tech. degree in computer engineering from Visvesvaraya Technological University, Belagavi, Karnataka, India. She completed her Ph.D. from University of Mysore, India. She is currently working as an associate professor at JSS Science and Technology University. She is serving as reviewer for many conferences and journals. She is a lifetime member of ISTE and CSI. Her area of research includes medical image analysis, biometrics, and text mining. She can be contacted at email: ckr@jssstuniv.in.