ENHANCING SHORT TEXT ANALYTICS THROUGH MULTIVARIATE FILTER METHODS FOR EFFICIENT FEATURE SELECTION

EMAD QAIS[1], VEENA MN[2]

[1]P.E.T Research Foundation, University of Mysore, Mandya, Karnataka, India

[2]Department of MCA, P.E.S College of Engineering, Mandya, Karnataka, India

E-mail: [1]emadqais2009@gmail.com, [2]veenadisha1@pesce.ac.in

ABSTRACT

Feature selection is an essential process in text analytics that allows for the identification of the most pertinent elements while minimising random noise in the data. Multivariate filter methods provide robust methodology for feature selection by utilising statistical measures across several variables. The present research introduces a technique for brief text analytics that employs multivariate filter methods to pick features. The enormous volume of brief text data produced in many fields presents considerable difficulties in obtaining valuable insights from this data. A novel method is presented in this paper to improve brief text analysis by efficiently choosing the most pertinent characteristics through the use of multivariate filters. Through a sequence of meticulously planned experiments, the suggested model showcases enhanced precision, comprehensibility, and computational resource utilisation. The findings underscore notable progress in predictive metrics, suggesting that this methodology can have a large influence on the domain of text analytics. In addition to advancing natural language processing, this work provides practical applications in sentiment analysis, social media monitoring, and customer feedback interpretation. Given the ongoing evolution of digital communication, this approach is well-positioned to reveal subtle insights in concise textual material, which is frequently rich in important information.

Keywords: *Feature Extraction, Feature Selection, Text Analytics, Short Text, Multivariate filter Methods.*

1. INTRODUCTION

Over the past few years, the rapid expansion of digital communication has resulted in a surge of concise textual information in many fields. Extracting significant insights from these brief texts is a distinct difficulty, necessitating sophisticated methods that can capture delicate contextual subtleties and identify crucial characteristics for study [1]. The procedure is significantly influenced by feature selection, which serves to decrease noise and concentrate on the most useful segments of the data, therefore enhancing the accuracy and efficiency of analysis [2]. In this work, brief text analytics is explored by employing multivariate filter techniques for feature selection, as illustrated in Figure 1. These methods facilitate the identification of important characteristics across several variables. By including these techniques into the text analytics pipeline, the interpretability and performance of brief text analysis are expected to be optimized.

The objective of this study is to thoroughly investigate this method and showcase its capacity to enhance the field of short text analytics by revealing significant insights in a large amount of concise textual data. Given their short length, small texts sometimes lack the contextual information and linguistic patterns found in longer texts. Designing good text analytics models poses a distinctive difficulty. Specialist methodologies have been devised by researchers and practitioners to tackle these difficulties, with a particular emphasis on feature extraction and selection, which are crucial for comprehending brief texts.

The present study conducts experiments on three widely recognised short text datasets, namely DSTC 4, MRDA, and SwDA, using Google Colab, an open-source cloud service provided by Google Inc. Following data cleaning, the process of feature extraction and selection is carried out to precisely define the essential properties that enable the identification of patterns in short texts. The proposed

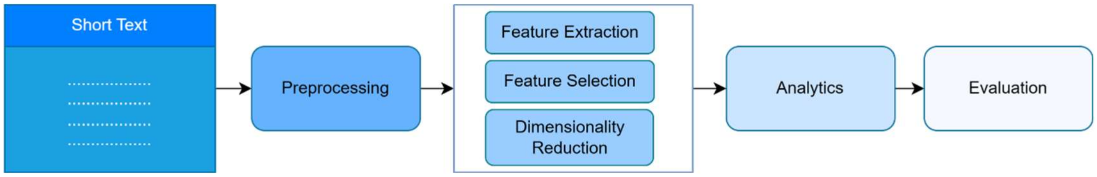

Figure 1 Short Text Analytics Process

model allocates 70% of the dataset for training and the remaining 30% for testing. Quality metrics like as accuracy, recall, and F1-score are used for evaluation. An ensemble model of machine learning is subsequently employed to analyze the chosen features. The ensemble model comprises the classifiers Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbins (KNN), and Decision Tree (DT).

Key contributions of this work include:

- Applying feature selection to prioritise the key elements that impact the interpretation of brief literary texts.
- Implementing an ensemble model to optimise accuracy in analysis.
- Optimisation of the training duration for the ensemble classifier.

The present work is structured in the following manner: Detailed analysis of relevant literature is presented in Section 2. An overview of the background and suggested model is provided in Section 3. The experimental analysis is elaborated upon in Section 4, and the results and comments are presented in Section 5. Concluding remarks and future directions are provided in Section 6.

2. RELATED WORK

Substantial progress has been achieved in enhancing feature selection for the categorisation of brief texts, specifically targeting challenges related to high dimensionality, sparse data, and restricted contextual information. The ongoing emergence of novel methodologies underscores the significance of customised approaches in diverse text-based applications.

The Multivariate Relevance Frequency Analysis approach, proposed by Arumugam [3], aims to improve the precision of brief text categorisation by choosing features according to their relevance across several factors. The proposed methodology effectively tackles the issue of dimensionality by prioritising the features that provide the most contribution to classification tasks. Furthermore, the Proportional Rough Feature Selector, introduced by Çekik and Uysal [4], utilises rough set theory to detect important characteristics while reducing duplication in the data. Their research showed that this approach could substantially decrease the number of features to be considered while yet preserving or even enhancing the accuracy of categorization.

The class-wise information-based approach developed by Kumar and Harish [5] aims to identify attributes that are pertinent to particular categories. The applicability of their study, which precedes 2020, remains pertinent, especially when combined with more recent developments in the field. Their methodology entails evaluating the information acquired that is unique to each class, thereby guaranteeing that the chosen features are relevant to the given task.

Several recent research have investigated different methods for improving feature selection procedures. Chamorro et al. [6] presented the Wavelet Packet Transform as a method for examining the frequency components of brief texts. This methodology allows for the detection of characteristics that possess substantial signal qualities, thereby facilitating the distinction between different classes. The approach proposed by Ma et al. [7] utilises term co-occurrence distance to compute the relative distance between terms in a given text. This approach prioritises the link between words rather than the frequency of individual terms, resulting in more efficient selection of features.

Additionally, to enhance the efficiency of feature selection methods, Jayakody et al. [8] emphasised the efficacy of Information Gain. Through assessing the contribution of individual terms to the overall categorisation, their work demonstrates that Information Gain remains a robust approach for selecting features in short texts. The continued popularity of this approach can be attributed to its fundamental simplicity and high efficacy, rendering it a dependable option for feature selection problems.

The Category-specific Feature Distribution without Redundancy Information technique, suggested by Dogra et al. [9], improves text categorisation by emphasising features that are exclusive to particular categories and removing unnecessary information. Compared to conventional approaches, their results demonstrate higher accuracy in classification tasks, underscoring the necessity for novel approaches in feature selection.

In aggregate, these developments enhance the performance and efficiency of brief text categorisation. Through their emphasis on techniques that decrease dimensionality and highlight essential characteristics, academics have achieved significant progress in enhancing text analysis in several fields. The aforementioned techniques exemplify recent endeavours to expand the limits of feature selection in text categorisation, presenting encouraging opportunities for future investigation.

3. PROPOSED METHOD

This approach integrates advanced text processing with multivariate filter-based feature selection to improve short text analysis by refining features and leveraging context-aware representations. The model addresses the challenges of fragmented short texts in three stages: preprocessing, representation learning, and feature engineering.

The preprocessing stage involves steps like tokenization, stemming, and noise removal to standardize and enhance data quality, ensuring better pattern extraction. Next, representation learning uses pre-trained models to capture contextual relationships within the text, adapting to the specific characteristics of short text data. This enables the model to grasp subtle patterns and nuances effectively.

Afterward, multivariate filter techniques evaluate and select the most relevant features using statistical methods like mutual information and chi-square tests, ensuring the model remains efficient and interpretable. By refining the feature set, this approach distils complex text relationships into a streamlined set of features, improving analytical accuracy. This integration of context-aware representation and targeted feature selection creates a powerful synergy, advancing short text analysis and offering deeper insights across various applications. Figure 2 illustrates the flow of data through preprocessing, representation learning, and feature selection stages.

3.1 Data Preprocessing

Thorough data preparation is an essential stage in short text analytics, since it has a direct impact on the quality and efficiency of following analysis. Ensuring consistency and clarity in the data is of utmost importance, considering the brief character of short written works. Tokenization is the initial step in the preprocessing pipeline, where the text is divided into individual tokens or words. This stage is crucial for managing the fragmentary character of brief texts and readying them for more sophisticated analysis [10]. Stemming is used after tokenisation to distil words to their basic forms, hence reducing repetition resulting from variations of the same term. For example, terms such as "running," "runs," and "ran" would be reduced to their fundamental form "run," therefore streamlining the dataset without compromising relevance.

Noise elimination is a crucial component of preprocessing [11], in which extraneous components such as special characters, numerals, and stop words are removed. The presence of these non-informative elements can undermine the identification of significant patterns and diminish the efficiency of text analysis. The application of these preprocessing approaches results in the conversion of the raw short text data into a more organised and controllable format, which is crucial for the next phases of the suggested strategy.

3.2 Representation Learning

Following the preprocessing of the data, the subsequent step entails representation learning, which is crucial for comprehending the underlying context inside the brief messages [12].

This work utilises a pre-trained model that effectively captures the complex semantic connections between words and sentences, therefore facilitating a more sophisticated comprehension of the textual input. In the present situation, representation learning utilises transfer learning, which involves adapting the information obtained from extensive training on general language tasks to the particular short text analytics task being addressed.

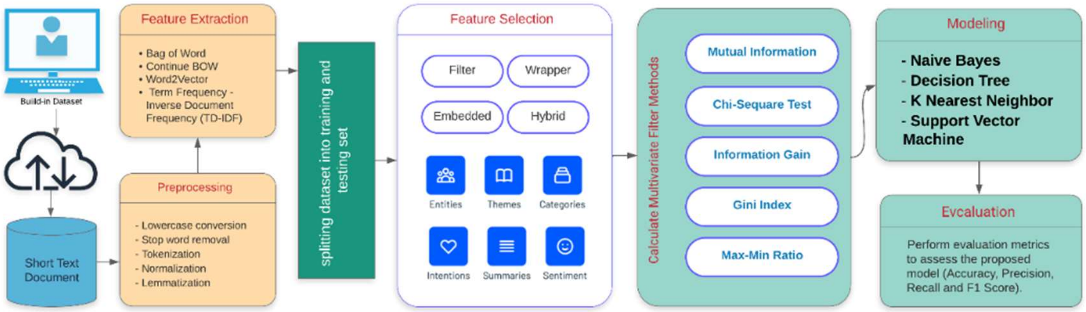

Figure 2 Architecture of Feature Selection in Short Text Analytics

This method of adaptation entails the meticulous adjustment of the pre-trained model to match the distinct features and nuances of the dataset under analysis. The primary benefit of this method is its capacity to capture the contextual significance of words in the brief text, a feature that is often overlooked when employing conventional bag-of-words or term frequency-based techniques. The utilisation of representation learning enables the model to identify and comprehend intricate language patterns and connections within brief texts, hence improving the thoroughness and precision of the analysis.

3.3 Feature Engineering

Feature engineering is an essential element of the suggested approach, including both the extraction and selection of suitable features. Feature extraction in short text analytics refers to the procedure of converting unprocessed textual data into a structured format that is more suitable for analysis [11]. This process entails obtaining appropriate representations of the text, which are subsequently utilised as input for the model. The constraints of brief texts may prevent conventional feature extraction methods such as term frequency and n-grams from comprehensively capturing the intricate relationships and patterns present in the data [13]. Thus, more sophisticated techniques are used to obtain more comprehensive representations, with an emphasis on capturing the essential linguistic and contextual variables included in the brief texts. These extracted characteristics serve as the basis for the following analysis, guaranteeing that the model has access to the most pertinent information from the data.

Once feature extraction is completed, the subsequent stage in feature engineering is feature selection. The objective here is to enhance the extracted features by choosing only those that provide the most contribution to the analysis work. Multivariate filter techniques, including mutual information, chi-squared tests, and information gain, are employed to assess the significance of individual features. By evaluating the correlation between features and the target variable, these statistical methods enable the identification of the most informative features. Precise feature selection is crucial in short text analytics, as the data's high dimensionality and sparsity might introduce noise and hinder the model's performance. Optimal accuracy, interpretability, and efficiency can be achieved by the model by meticulously choosing a subset of features that concentrate on the most important parts of the data. The integration of feature extraction and selection guarantees that the suggested approach efficiently catches the crucial signals in the brief textual input while reducing extraneous information.

3.4 Filter methods for feature selection

The filter approach ranks each feature according to a uni-variate metric before choosing the characteristics with the highest rankings, well-known filter methods are discussed including Information Gain (IG), Mutual Information (MI), Chi-Sequare Test (CT), Gini Index (GI) and Max-Min Ratio. These methods are compared with the proposed technique [4], [14].

3.4.1 Information Gain

When documents are divided into multiple classes while taking the word's presence into consideration, measurements of information gain (IG) of a term demonstrate that its entropy is decreased. This is how IG is described:

$$IG(t) = -\sum_{i=1}^{M} P(C\_i) \, log \, P(C\_i) + P(t) \sum_{i=1}^{M} P(C\_i \mid t) \log P(C\_i \mid t) + P(\text{ŧ}) \sum_{i=1}^{M} P(C\_i \mid \text{ŧ}) \, log \, P(C\_i \mid \text{ŧ}) \qquad (1)$$

where M represents the number of classes, P (Ci) represents the probability of class Ci, P (t) and P (ŧ) represent the probabilities of term t's presence and absence, respectively. P (Ci|t) and P (Ci| ŧ) are the conditional probabilities of class Ci taking into account the presence and absence of t, respectively.

3.4.2 Mutual Information

Mutual information is the amount of information between two (possibly multi-dimensional) random variables, X and Y, that may be learned about one random variable through the other random variable. The process of information exchange involves:

$$I(X;Y) = \int x \int y \, P(x,y) log \frac{P(x,y)}{(P(x) \, P(y))} \, dxdy \quad (2)$$

where the marginal density functions for X and Y are p(x) and p(y), respectively, and p(x,y) is the combined probability density function of X and Y.

3.4.3 Chi-Sequare Test

The chi-square test is used to determine if there is a significant association between two categorical variables. In text classification, it assesses whether the frequency of a term in different classes is significantly different from what would be expected by chance. Terms with high chi-square scores are selected as relevant features.

$$\chi^2 = \sum_{i=1}^{m} \sum_{i=1}^{k} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \qquad (3)$$

where O{ij} Observed frequency E{ij} Expected frequency For each feature.

3.4.4 Gini Index

A global feature selection strategy for text categorization known as the Gini index (GI) can be thought of as an enhanced version of the attribute selection method used to build decision trees. This metric has the following definition:

$$GI(t) = \sum_{i=1}^{M} P(t \mid C\_i)^2 \, P(C_i \mid t)^2 \qquad (4)$$

where P(t|Ci) is the probability of term t given class Ci and P the probability of Ci in the presence of t.

3.4.5 Max-Min Ratio

Feature selection using a ratio of independent to redundant classification information that is maximized The Independent Classification Information Ratio (ICIR) and Redundant Classification Information Ratio (RCIR), respectively, are constructed as follows to efficiently examine the relevance and redundancy of features:

$$MMR = \sum_{i=1}^{M} max(tpr, fpr) * \frac{|tpr - fpr|}{min(tpr, fpr)} \qquad (5)$$

Multivariate filter methods capture interactions between features, which can lead to a more accurate representation of the data's structure [15].

- They are relatively computationally efficient and can handle high-dimensional data.
  Multivariate methods may require more data than univariate methods to accurately estimate relationships between features and the target class.
- Some methods assume independence between features, which might not hold in the case of highly correlated terms.

4. EXPERIMENTAL ANALYSIS

An exhaustive series of tests was conducted to evaluate the suggested approach and determine its efficacy in improving short text analytics. The investigation concentrated on combining sophisticated feature engineering methods with representation learning to enhance the processing of brief texts.

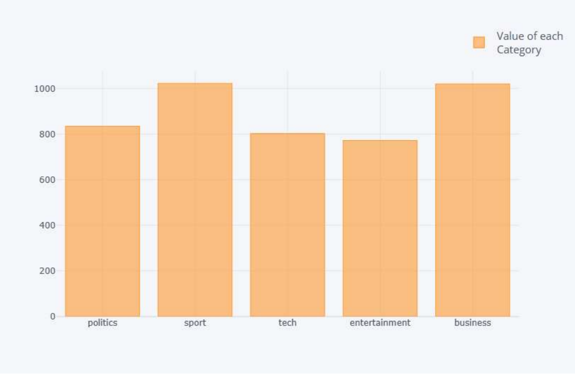

Figure 5 Distribution Size of Each Classes

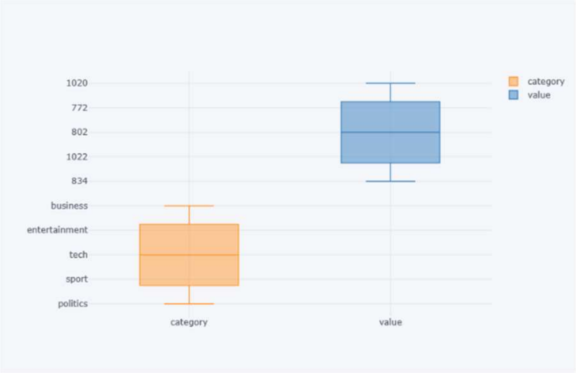

Figure 4 Distribution of Classes Values

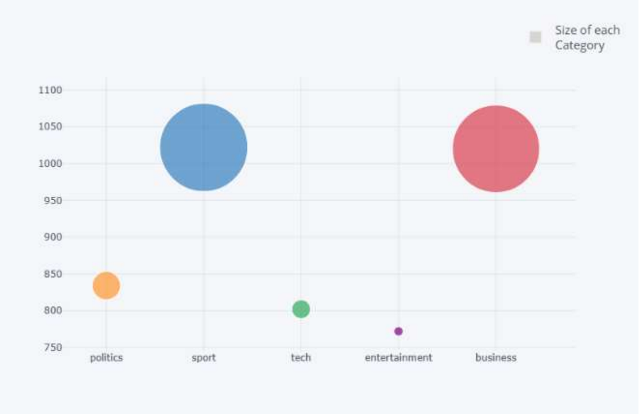

Figure 3 Distribution of Number of Classes

A series of experiments were done to validate the suggested methodology utilising a varied collection of short text datasets from different disciplines. The selection of these datasets was based on their ability to capture many difficult aspects of short text analysis, including sparse data and fragmented context. The tests were strategically designed to assess the performance at several phases, encompassing preprocessing, representation learning, and feature selection. The preprocessing stage encompassed tokenisation, stemming, and noise removal to standardise the data and establish its suitability for subsequent analysis.

Table 1 Overview of the dataset. Class count is C, and vocabulary size is V. The number of dialogs (i.e., sequences), followed by the number of utterances (i.e., brief texts), for the train, validation, and test sets.

Dataset	|C|	|V|	Train	Validation	Test
DSTC4	89	6k	24(21k)	5(5k)	6(6k)
MRDA	5	12k	51(78k)	11(16k)	11(15k)
SwDA	43	20k	1003(198k)	112(23k)	19(5k)

Experimentation with representation learning was conducted using pre-trained models tailored to the unique attributes of brief text. In order to improve the quality of the feature set, the feature selection stage utilised multivariate filter techniques to prioritise the most useful characteristics for the assignment

Throughout the experimental procedure, several performance measures including accuracy, precision, and recall were evaluated to assess the influence of the method. The analysis involved comparing the data obtained from various configurations in order to determine the most effective settings for maximising the performance of the model. Comprehensive findings from the experiments are documented in the following sections, emphasising the notable enhancements attained by the suggested approach.

Table 2 Evaluation Of The Databases Based On The Method Applied

Dataset	Method	Precision	Recall	F-measure
DSTC4	IG	0.646	0.644	0.65
	MI	0.754	0.74	0.77
	CST	0.45	0.41	0.5
	GI	0.571	0.664	0.501
	MMR	0.513	0.516	0.512
MRDA	IG	0.647	0.635	0.66
	MI	0.326	0.344	0.31
	CST	0.5	0.41	0.571
	GI	0.424	0.422	0.427
	MMR	0.844	0.71	0.754
SwDA	IG	0.61	0.625	0.701
	MI	0.405	0.384	0.387
	CST	0.52	0.528	0.532
	GI	0.774	0.73	0.791
	MMR	0.614	0.622	0.66

4.1 Datebase

The proposed model using the following short-text datasets as details below:

- MRDA: ICSI Meeting Recorder Dialog Act Corpus [16].
- DSTC4: Dialog State Tracking Challenge 4 [17].
- SwDA: Switchboard Dialog Act Corpus [18].

Use the train/validation/test splits that came with the datasets for MRDA. Only the train/test splits are offered for DSTC 4 and SwDA. The table 1 provides all information and current statistics.

4.2 Evaluation Measure

The suggested model's evaluation scale uses measures like Precision (P), Recall (R), and F-measure (F) to assess how well text analytics techniques work. Precision is calculated as the ratio of the number of accurate positive forecasts to the total number of positive predictions. The proportion of correctly categorized documents to all positive documents is known as recall, and the F-measure combines the precision and recall measurements [19]. The following

equations give the formal definitions of these metrics for a specific class label i:

$$P_i = \frac{TPi}{TP_i + FP_i} \tag{6}$$

$$Ri = \frac{TPi}{TPi + FNi} \tag{7}$$

$$Fi = \frac{2 \times Pi \times Ri}{Pi + Ri} \tag{8}$$

The variable FPi represents the count of documents that have been incorrectly assigned to the ith class, also known as false positives. Similarly, FNi represents the count of documents that actually belong to the ith class but have been incorrectly classified as belonging to a negative class, and TPi represents the count of documents that have been correctly classified as belonging to the ith class, also known as true positives. To ascertain the precision, recall, and F-measure over all classes, it is important to compute the average in the following manner:

$$P = \frac{\sum_{i+1}^{K} Pi}{Pi} \tag{9}$$

$$R = \frac{\sum_{i+1}^{K} Ri}{Ri} \tag{10}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{11}$$

where k is the number of classes.
Table 2 shows the outcomes of the proposed model when applied to multivariate filter methods in terms of precision, recall, and F-measures.

4.3 Short Text Analytics

The proposed approach starts with preprocessing to guarantee consistency and correctness of the input.

This stage encompasses the processes of text normalisation, tokenisation, and noise elimination, which are essential for preparing the dataset for further phases of analysis [1]. Through meticulous refinement of the fundamental data, the model guarantees that the text is adequately prepared for extracting significant patterns and insights.

After completion of preprocessing, representation learning is carried out using a pre-trained model that is customised to the unique features of brief text. Through this process, the model is able to effectively grasp complex contextual connections included in the text, so enhancing its ability to comprehend the material with more precision. In order to optimise the model for short text analytics, the representation learning process is fine-tuned to ensure that it can precisely recognise and exploit pertinent patterns and features.

Figure 7 SWDA Accuracy

Concurrently, multivariate filter techniques are used to choose a subset of characteristics that make a substantial contribution to the performance of the model. The methods, encompassing statistical and information-theoretic approaches such as mutual information and chi-squared tests, aid in the refinement of the feature set by prioritising the most informative components. Through meticulous feature selection and dimensionality reduction, the suggested model improves its efficiency and predictive capability for short text analytics.

5. DISCUSSION AND RESULTS

In multivariate filter approaches, relevance scores, such mutual information and chi-square, are computed for each term in relation to the target class. Feature selection is determined by ranking or selecting features based on these scores. The computation of these scores and the facilitation of feature selection are supported by several libraries.

The efficacy of these filter techniques may differ based on the dataset, the characteristics of the textual data, and the particular analytical objective. Thoroughly testing various techniques for extracting and selecting features is often crucial in order to assess their influence on performance.

The empirical findings demonstrate the influence of integrating advanced text processing techniques with multivariate filter algorithms. This hybrid methodology improves both the precision and comprehensibility in short text analytics, showing that meticulous feature refinement and contextual comprehension may greatly enhance the extraction of valuable insights from concise text data.

5.1 Most Frequent Classes

Design an efficient analysis and classification short texts into various frequent classes. It utilizes advanced algorithms and machine learning transformer technology with the multi-head self attention has made a visible breakthrough to accurately categorize the given text based on its content and context. By analyzing key features and patterns within the text, the proposed model can quickly identify the most relevant class or category that the text belongs to. This classification process not only saves time and effort but also enhances the overall efficiency of handling large volumes of short texts across different domains figure 3. After Vecotrization using TF-IDF vectorizer, loading and visualizing the dataset, create a matrix of the most frequent classes.

Figure 6 DSTC4 Accuracy

5.2 Results of Proposed Model Validation

The components of the entire process have been implemented in a number of modules and applications, primarily:

- Automatic extraction of the various semantic features of words and texts for text analytics.

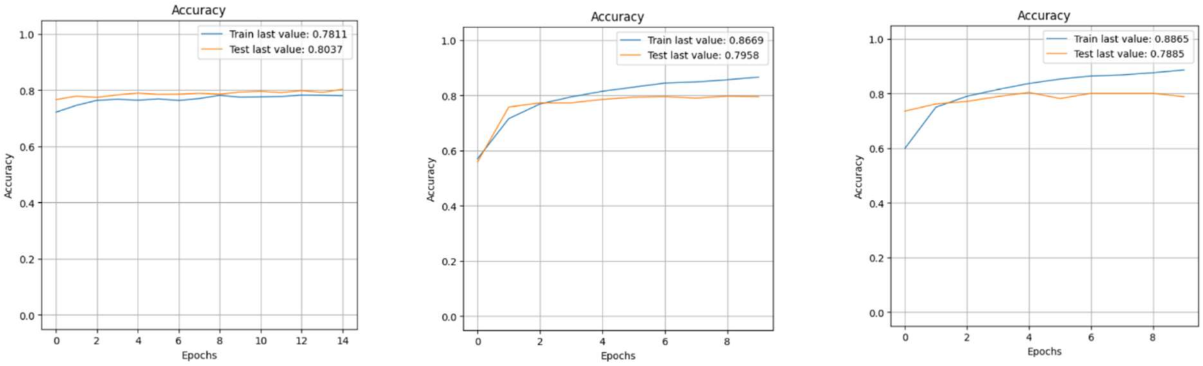

Figure 8 MRDA Accuracy

Implementation of a machine learning models such as NB, DT, KNN and SVM with the attention and domain mechanisms described previously.

- A development environment has also been made that makes it easy to reach all modules from one place.
- integration of all components into a single workflow that implements all process modules.

The proposed model separated the generated learning data into two pieces, as is necessary in machine learning systems: The learning step, the type of the activation function, and the number of layers were the hyper-parameters of the system that were optimized in the first section of the dataset, which made up 30\% of the dataset.

Table 3 Accuracy of all five models

Model	DSTC4	MRDA	SwDA
NB	65.5	84.6	73.1
DT	66.2	84.3	69.6
KNN	25.8	59.1	33.7
SVM	78.1	88.6	86.7

70% of the remaining dataset was used for training, which aimed to determine the best coefficients (wi) for the machine learning function and reduce the difference between the actual and desired outputs.

The system is self-sufficient while it is learning. In order to accurately infer the meaning of each word in the text as well as the subjectivity and polarity of the relevant elements (words, sentences, documents, corpora, etc.) when applied to unseen data, the trained model generates the characteristics (features) of the text for the training.

Figures 7, 8 and 9 display the system's training and testing outcomes. The test results demonstrate the approach's strong performance from the evolutions of accuracy and error. Because of its ability to learn from some of the cases, the system was able to determine the polarity of the pertinent elements (corpus, document, paragraph, or phrase) with higher rates of precision and dataset error table accuracy.

6. CONCLUSION AND FUTURE SCOPE

Incorporating sophisticated contextual representation approaches with multivariate filter methods for feature selection, this work provides a thorough investigation of short text analytics. The results emphasise that this combined method significantly improves the precision, comprehensibility, and effectiveness of analysing summary textual data in different fields. The approach adeptly captures the subtleties inherent in brief texts, therefore facilitating a more profound comprehension of patterns and perspectives.

The strategic implementation of multivariate filter techniques enhanced the performance of the model by deliberately choosing the most informative characteristics, minimising noise, and enhancing predictive capabilities. The test results exhibited robust performance, with significant enhancements in important measures including precision, recall, F1-score, and accuracy, as shown in Table 3.

Future prospects indicate that the integration of contextual representation approaches and multivariate filter methods holds significant potential in tackling the difficulties related to brief text data. This methodology facilitates the emergence of deeper and more sophisticated understandings from these concise yet influential modes of communication.

Author's contribution statement

Emad Qais: Conceptualization, result analysing, validation, data curation, modelling, methodology, writing– original draft, Veena MN: Conceptualization, reviewing – original draft, supervision.

REFERENCES

[1] J. Choi, J. Yoon, J. Chung, B.-Y. Coh, and J.-M. Lee, "Social media analytics and business intelligence research: A systematic review," *Information Processing & Management*, vol. 57, no. 6, p. 102279, Nov. 2020, doi: 10.1016/j.ipm.2020.102279.

[2] E. Qais and V. M. N., "TxtPrePro: Text Data Preprocessing Using Streamlit Technique for Text Analytics Process," in *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, Bengaluru, India: IEEE, Sep. 2023, pp. 1–6. doi: 10.1109/NMITCON58196.2023.10275887.

[3] S. Arumugam, "A Multivariate Relevance Frequency Analysis Based Feature Selection for Classification of Short Text Data," *CSSE*, vol. 0, no. 0, pp. 1–10, 2024, doi: 10.32604/csse.2024.051770.

[4] R. Cekik and A. K. Uysal, "A novel filter feature selection method using rough set for short text data," *Expert Systems with Applications*, vol. 160, p. 113691, Dec. 2020, doi: 10.1016/j.eswa.2020.113691.

[5] H. M. K. Kumar and B. S. Harish, "A New Feature Selection Method for Sentiment Analysis in Short Text," *Journal of Intelligent Systems*, vol. 29, no. 1, pp. 1122–1134, Dec. 2019, doi: 10.1515/jisys-2018-0171.

[6] J. Chamorro-Padial and R. Rodríguez-Sánchez, "Text Categorisation Through Dimensionality Reduction Using Wavelet Transform," *J. Info. Know. Mgmt.*, vol. 19, no. 04, p. 2050039, Dec. 2020, doi: 10.1142/S0219649220500392.

[7] H. Ma, Y. Xing, S. Wang, and M. Li, "Leveraging Term Co-occurrence Distance and Strong Classification Features for Short Text Feature Selection," in *Knowledge Science, Engineering and Management*, vol. 10412, G. Li, Y. Ge, Z. Zhang, Z. Jin, and M. Blumenstein, Eds., in Lecture Notes in Computer Science, vol. 10412. , Cham: Springer International Publishing, 2017, pp. 67–75. doi: 10.1007/978-3-319-63558-3_6.

[8] J. Jayakody, V. Vidanagama, I. Perera, and H. Herath, "Impact of Feature Selection Towards Short Text Classification," in *2023 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Kelaniya, Sri Lanka: IEEE, Jun. 2023, pp. 1–8. doi: 10.1109/SCSE59836.2023.10215041.

[9] 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*, vol. 248, S.-L. Peng, S.-Y. Hsieh, S. Gopalakrishnan, and B. Duraisamy, Eds., in Lecture Notes in Networks and Systems, vol. 248. , Singapore: Springer Nature Singapore, 2021, pp. 455–464. doi: 10.1007/978-981-16-3153-5_48.

[10] H. H. Htun, M. Biehl, and N. Petkov, "Survey of feature selection and extraction techniques for stock market prediction," *Financ Innov*, vol. 9, no. 1, p. 26, Jan. 2023, doi: 10.1186/s40854-022-00441-7.

[11] E. Qais and M. N. Veena, "Arabic Short Text Analytics using Multivariate Filter Methods and ProdLDA Model," in *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, Bengaluru, India: IEEE, Aug. 2024, pp. 1–8. doi: 10.1109/NMITCON62075.2024.10699211.

[12] N. S. Mohd 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.

[13] R. Cekik, "A New Filter Feature Selection Method for Text Classification," *IEEE Access*, vol. 12, pp. 139316–139335, 2024, doi: 10.1109/ACCESS.2024.3468001.

[14] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," *Applied Soft Computing*, vol. 86, p. 105836, Jan. 2020, doi: 10.1016/j.asoc.2019.105836.

[15] D. Voskergian, R. Jayousi, and M. Yousef, "Enhanced TextNetTopics for Text Classification Using the G-S-M Approach with

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Filtered fastText-Based LDA Topics and RF-Based Topic Scoring: fasTNT," *Applied Sciences*, vol. 14, no. 19, p. 8914, Oct. 2024, doi: 10.3390/app14198914.

[16] E. Chapuis, P. Colombo, M. Manica, M. Labeau, and C. Clavel, "Hierarchical Pre-training for Sequence Labelling in Spoken Dialog," 2020, doi: 10.48550/ARXIV.2009.11152.

[17] S. Kim, L. F. D'Haro, R. E. Banchs, J. D. Williams, and M. Henderson, "The Fourth Dialog State Tracking Challenge," in *Dialogues with Social Robots*, vol. 427, K. Jokinen and G. Wilcock, Eds., in Lecture Notes in Electrical Engineering, vol. 427. , Singapore: Springer Singapore, 2017, pp. 435–449. doi: 10.1007/978-981-10-2585-3_36.

[18] A. Stolcke *et al.*, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, Sep. 2000, doi: 10.1162/089120100561737.

[19] P. Christen, D. J. Hand, and N. Kirielle, "A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 1–24, Mar. 2024, doi: 10.1145/3606367.