Int J Syst Assur Eng Manag (December 2024) 15(12):5637–5652
https://doi.org/10.1007/s13198-024-02555-w

ORIGINAL ARTICLE

Machine learning and rule-based embedding techniques for classifying text documents

Asmaa M. Aubaid[1] · **Alok Mishra**[2] · **Atul Mishra**[3]

Received: 3 May 2024 / Revised: 12 September 2024 / Accepted: 5 October 2024 / Published online: 24 October 2024

Abstract Rapid expansion of electronic document archives and the proliferation of online information have made it incredibly difficult to categorize text documents. Classification helps in information retrieval from a conceptual framework. This study addresses the challenge of efficiently categorizing text documents amidst the vast electronic document landscape. Employing machine learning models and a novel document categorization method, W2vRule, we compare its performance with traditional methods. Emphasizing the importance of tuning hyperparameters for optimal performance, the research recommends the W2vRule, a word-to-vector rule-based framework, for improved association-based text classification. The study used the Reuters Newswire dataset. Findings show that W2vRule and machine learning can effectively tell apart important categories. Rule-based approaches perform better than Naive Bayes, BayesNet, Decision Tables, and others in terms of performance metrics.

Keywords Bayesian · BayesNet · Lazy · IBK · Naïve Bayes · IBL · Reuters · Text analytics

1 Introduction

Text classification is a type of machine learning that puts open-ended text into a set of predefined categories. (Agrawal & Batra 2013). Text classifiers can organize, structure, and classify almost any text, including documents, medical studies, files, and text from the web. Classifying a large amount of text allows for platform standardization, more relevant and efficient searches, and an improved user experience. (Batrinca & Treleaven 2015). Thus, technologies like artificial intelligence (AI) and machine learning (ML) are becoming useful in many fields. Interpretable machine learning models facilitate making rational, data-driven conclusions. (Stiglic et al. 2020). Interpretability is how well people understand a model's decisions or how distinct features (inputs) generate an inevitable conclusion (output). The text is complex, with a wide range of topics and terms. Text classification can categorize emails, and text genres. (Onan 2018), topic sarcasm (Onan 2019), Sentiment analysis (Balli et al. 2022) and Web queries. New research utilizes cutting-edge machine learning techniques rather than ontology and rule-based approaches.

Text classification is essential in applications like information retrieval (Dwivedi & Arya 2016), e-government (Ku & Leroy 2014), data filtration (Melville et al. 2009), text archives (Tao et al. 2020), and digital libraries (Deng et al. 2019). It involves the categorization of text data into predefined classes based on its content. Machine learning algorithms, such as Support Vector Machines (SVMs), Naive Bayes, and Deep Neural Networks (DNNs), can be employed to perform text classification. Recent text classification

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13198-024-02555-w.

✉ Alok Mishra
alok.mishra@ntnu.no

Asmaa M. Aubaid
asalhmuh@gmail.com

Atul Mishra
atul.mishra@bmu.ed.in

[1] Ministry of Higher Education and Scientific Research/Science and Technology, Baghdad/Al-Jadriya, Iraq

[2] Faculty of Engineering, Norwegian University of Science and Technology, Trondheim, Norway

[3] BML Munjal University, Kapriwas, India

research has focused on categorizing news and related stories. Recent research on text categorization has concentrated mainly on classifying the information and other relevant items. This study compares the novel document classification method "W2vRule," which leverages Word2vec for efficient feature selection in various ways. The study attempts to classify news items into distinct verticals using machine learning and a rule-based approach for text classification.

The study used the Reuters-21578 dataset, a collection of news articles (Mimaroglu 2020). The original corpus has 10,369 documents and a vocabulary of 29,930 words. The experiments use this standard dataset and focus on several important issues related to establishing feature selection metrics. The primary objective of choosing a feature metric is to assess the importance of words based on several criteria that show which keywords should be kept intelligible and which should be eliminated. The categorization process uses these keywords to determine the most effective rule-based machine learning (RBML) strategies and the most accurate metrics to evaluate the system's efficiency (Onan et al., 2018). It is expected that a high level of accuracy and F-measures will be attained by employing the similarity technique to determine the most suitable feature selection or vocabulary, which is a significant step in the classification process. While various schemes have been proposed to tackle the challenges of categorizing text documents, it is imperative to acknowledge the existing limitations in the literature. Prior approaches often grapple with issues such as scalability, adaptability to diverse datasets, and the need for extensive parameter tuning. Furthermore, the evolving landscape of electronic document archives demands robust methods that can efficiently handle the growing volume and complexity of textual information (Mao et al. 2024)(Ranjan & Prasad 2023). This study aims to address these gaps by introducing a novel document categorization method, W2vRule, and systematically comparing its performance against conventional machine learning approaches. Efficient and accurate document classification holds paramount importance across various industries, impacting information retrieval, organizational efficiency, and decision-making processes. The proposed W2vRule framework not only aims to overcome the shortcomings of current approaches but also positions itself as a valuable tool for industries grappling with vast and dynamic datasets.

We used Weka (Witten et al. 2002)(Hall et al. 2009), a suite of machine learning algorithms used for data mining tasks. The study used Naïve Bayes, Naive Bayes Updateable (*"Class for a Naive Bayes classifier using estimator classes"*), Decision Tables, Lazy IBL, Lazy IBK, and rule-based models to discover an excellent approach to classification strategies. The study contributes to finding the most effective methods for text categorization by examining the performance metrics and undertaking a comprehensive machine-learning approach. Naive Bayes and NaiveBayes-Updateable classifiers, based on Bayes' probability theory, provide basic yet effective text classification models (Myaeng et al. 2006). Classifiers based on decision tables are also preferred for text classification since they are similar to the decision tree and neural network (Hmeidi et al. 2015). The research also used a word-to-vector rule-based framework, referred to as W2vRule, as a potential method for more accurate text classification based on association rules. The study evaluates the performance of Bayesian classifiers, lazy classifiers, W2vRule, and Random Forest to determine the optimal classification algorithm. Instance-Based Learning (IBL) finds the training instance closest to the test cases and predicts the class. The K-Nearest Neighbour (IBK) search strategy has a configurable distance function. Everything else is just like IBL (Pereira et al. 2021). The system architecture of the research work and the detailed analysis of these classifiers have been discussed in the subsequent section.

The rest of this paper is organized as follows: The second section reviews the literature on text classification using word embedding and machine learning techniques. The third section introduces the resources and procedures for text categorization and problem formulation. Section 4 describes the proposed methodology and introduces rule-based processing. Section 5 discusses the results and empirical findings. This section provides a comprehensive analysis of the evaluation findings from all approaches. Section 6 outlines the discussion. Finally, Sect. 7 concludes with recommendations for future research.

2 Literature review

Text classification is essential in machine learning, data mining, information retrieval, and natural language processing. Many research studies address different aspects and provide distinct solutions. This section focuses on text classification models based on embedding and rule-based vs. machine learning. Table 1 shows the significance of text classification over the past five years.

The study compares existing techniques where the primary focus is on Machine Learning and Rule-based embedding techniques. A recent study compared the performance of various machine learning algorithms. (Shahi et al. 2022), such as Support Vector Machines (SVM), Naive Bayes (NB), and Decision Trees (DT), with rule-based embedding techniques for text document classification. Another study proposed that feature selection techniques can significantly enhance the performance of both machine learning and rule-based techniques for text document classification (Pintas et al. 2021). A third study compared the performance of various deep learning algorithms, such as Convolutional Neural

Table 1 Baseline work on text classification

Title	Description	Remarks
"*Improved Word Segmentation System for Chinese Criminal Judgment Documents*" Zhang, (2024)	A hybrid model for automatically segmenting Chinese criminal judgment documents, combining BERT, BiLSTM, and CRF. Results show significant accuracy improvement, achieving a 94.82% F1 score	The hybrid model's integration and rule-based system demonstrate promise in enhancing accuracy for processing Chinese criminal judgment documents
"*Text Classification for Records Management*" Franks, (2022)	Classic text categorization approaches were compared to modern natural language processing systems using actual records data	Transformer language models demonstrated higher categorization ability
"*Supervised Text Classification using Text Search*" Mondal & Lohia, (2020)	The study describes a set of industrial standard algorithms that can accurately (86%) predict the categorization of any text given before labelled text data	These strategies automate ticket report distribution. This is specifically true for ticketing systems
"*A Feature Selection Method for Multi-Label Text Based on Feature Importance*" Zhang & Duan, (2019)	A feature selection method for multi-label text based on feature importance	The inter and intra-category contributions to each category are determined. The importance of the features is combined
"*Text Classification Using Word Embeddings*" Helaskar & Sonawane, (2019)	Measure text classification using word embeddings	Bag-of-Words is a popular model used to represent text

Networks (CNN) and Recurrent Neural Networks (RNN) (Banerjee et al. 2019) with rule-based embedding techniques for text document classification. A fourth research study evaluated the effectiveness of transfer learning in improving the performance of machine learning algorithms for text document classification compared to rule-based embedding techniques (Zhang & El-Gohary 2021). The results showed that transfer learning significantly improved the performance of machine learning algorithms, whereas rule-based embedding techniques did not show significant improvement. A fifth study compared the performance of ensemble learning techniques, such as Bagging and Boosting, with rule-based embedding techniques for text document classification (Mohsen et al. 2021).

Text classification approaches include rule-based, word embedding, and machine learning. Rule-based systems (also called "generation systems" or "expert systems") are classified as artificial intelligence systems, with growth beginning in the 1960s but becoming more common in the 1970s and 1980s (Levy & Goldberg 2014) and addressing concurrent processing and the activation of rules in production systems in the 1980s and 1990s. The rules-based system is used to express categorization system information (Ligeza 2006). The expert system affects rule-based systems and mimics human experience logic to explain a data-intensive challenge. Machine learning has advanced quickly (Pong et al. 2008). Text categorization approaches include decision trees, nearest neighbour classifiers, neural networks, regression (Mendel 2017; Sebastiani 2002) and semantic rule-based information extraction. (Cui et al. 2024).

In recent years, Support Vector Machines (SVMs) for text classification have been re-examined and implemented in several studies with encouraging outcomes. This is supported by an empirical comparison of twelve feature selection methods. (Forman 2003). Wibowo and Williams (Wibowo & Williams 2002) show that using pre-categorized training papers to select a few features helps to build hierarchical machine learning-based classification. Because a sizeable unlabeled corpus triggers it, embedding is a rewarding implementation in unsupervised machine learning and learning transfer. The core data collected during embedding can be used for tasks requiring a few datasets. Word embedding methods include Embedding Layer, Word to Vector (Word2Vec), and Global Vectors (GloVe).

The dataset has 12,902 articles from the newswire and 90 subject groups. This study used Word2Vec, a word embedding method (Aubaid & Mishra 2018) to classify the documents. Similar studies used bag-of-n-grams (Joulin et al. 2017), character n-grams, feature hashing (Weinberger et al. 2009)(Mikolov et al. 2011), and a machine-learning-based system (Martinelli et al. 2018) for text classification. Tailor and Patel (Tailor & Patel 2019) used statistical, unsupervised machine learning

with a rule-based approach to tokenizing Gujarati running texts. A passage is a section of text that defines the entire work, regardless of length (one word, one sentence, etc.). Some other embedding algorithm studies are Word2Vec (Çano & Morisio 2019), GloVe (Pennington et al. 2014), and Fast Text (Bojanowski et al. 2017)(Ducoffe et al. 2016). The following section will outline the theoretical techniques required to construct the rule-based approach using the word2vec embedding model to produce our W2vRule.

There have been several feature selection algorithms proposed and used in the field of text classification in the past decade. Some of the popular feature selection algorithms are Chi-Square, Information Gain, Correlation-based Feature Selection (CFS), Recursive Feature Elimination (RFE), Wrapper methods, and Embedded methods. For instance, in (Corrales et al. 2018), the authors compared the performance of Chi-Square, Information Gain, Correlation-based Feature Selection (CFS), and Recursive Feature Elimination (RFE) for text classification. The results showed that the Recursive Feature Elimination (RFE) method performed better than the other three methods. In other studies (Liu et al. 2022) (Wankhade et al. 2022), the authors compared the performance of Wrapper methods and Embedded methods for feature selection in text classification. The results showed that the Embedded method (Lasso) performed better than the Wrapper method (Backward Selection). Basu et.al (Basu & Murthy 2016) proposed a supervised term selection technique for dimensionality reduction. These studies demonstrate that the choice of feature selection algorithm can significantly impact the performance of text classification models and it is important to carefully evaluate different algorithms on the specific problem and dataset at hand.

3 Methods

The rule-based method considers several factors and is used for related tasks like classifying text. The rule-based categorization method uses models like word2vec and includes the essential steps of preparing the data for training, testing, and validation. Preprocessing steps include Tokenization, Stopword removal, Stemming, and Vectorization using TF-IDF representation.

Figure 1 presents an entire cycle of text classification using the machine learning technique. A machine learns from past outcomes to enhance its performance. Machine learning document classification improves a machine's capacity to categorize documents based on prior performance.

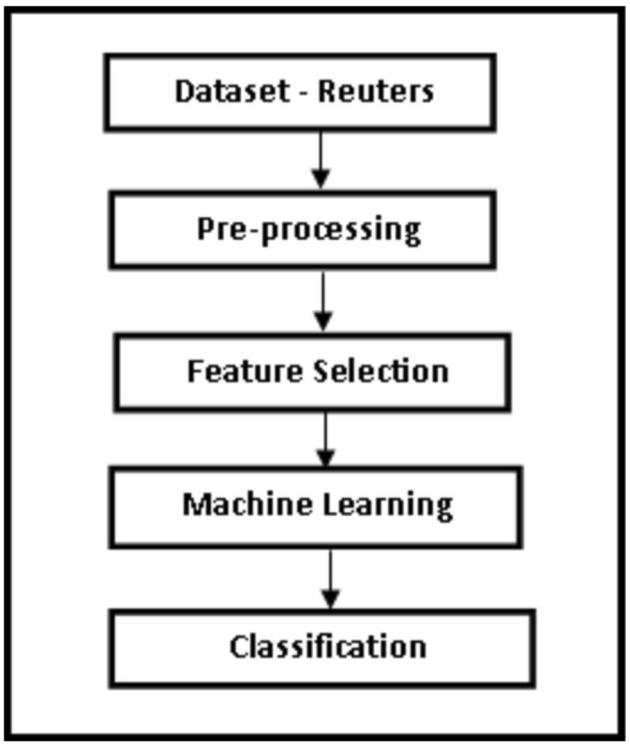

Fig. 1 Block diagram of the Machine Learning technique

3.1 Word embedding

Word embeddings equally represent words with similar meanings. This includes vocabulary as well as documents. In word embedding techniques, real-valued vectors represent single words. Each word is transformed into a vector before being input into a neural network. Then it is classified as deep learning. Each phrase is shown prominently and distractingly (Aubaid & Mishra 2020). Real-valued vectors can describe each term's tens or hundreds of measurements. This vector is multi-dimensional. Two-word embedding strategies are discussed below.

3.1.1 Term frequency-inverse document frequency (TF-IDF)

This technique also depends on the frequency method, yet it differs from count vectorization as it considers the occurrence of a word in a single document and the entire text. The idea of the TF-IDF feature selection is to choose the words with the highest scores, and TF-IDF provides the highest scores for terms in a few high-frequency documents. Alternatively, it is more discriminatory when a term occurs more often in a document. In contrast, most documents seem less discriminative about the content.

$$t_f - idf = \log\left(N/df_{ij}\right), \quad (1)$$

subscriptindicates the number of documents in which a term appears in all the documents.

N: Refers to the total number of documents.

3.1.2 Word2Vec

Word2vec is a statistical method used effectively to learn an independent word embedded in the body of a text. Google developed it in 2013 in response to making neural-network-based embedding training more efficient. It has since become the objective standard for developing ready-made word placement. In addition, the study included the analysis of the learned vectors and the study of vector mathematics on representations of words. The fundamental word2vec models are Skip-Gram and CBOW. In the Continuous Bag of Words model, a word given a context (a context can be something like a sentence) is estimated. Skip-Gram is the opposite: it calculates the context given an input word. The models are described below.

3.2 Models

In this section, we discuss the necessary steps to build a model.

3.2.1 Rule-Based

Here, we introduce the main principles of creating a rule-based method using a word embedding techniques and other criteria.

3.2.2 Rules–Preliminary

Rules: The symbolic representations of knowledge are derived using data. These symbolic representations are the most popular in the classification approach:

Its representation is taken in the easy and natural form, and its inspection is possible by depending on interpretations of humans (Aubaid & Mishra 2020).

A Standard Form of Rules:

A rule has a more standard form related to creating a rule, such as.

IF “instance” THEN “instance.”

In the first instance, “condition” is inserted, but “class” is inserted in the second instance.

Other forms: Class IF Conditions.

3.2.3 Machine learning approach (ML)

The machine learning approach can be defined as the machine’s capability to enhance performance based on previous results acquired earlier. The definition of the classification of documents in machine learning is the ability to improve a machine’s performance in document classification by depending on previous results. The goal of categorization in machine learning techniques is to collect items into sets that have the same feature values. Standard models of machine learning will be discussed in the following sections.

3.2.3.1 Naïve-Bayes The Naïve Bayes technique is one of the simplest models of classification. It assumes that all attributes of the examples are independent of each other, given the class context. This assumption is called the " Naïve Bayes assumption”. The "qualitative scales" are referred to as the nominal scales, and the measurements that comprise these qualitative scales are referred to as "qualitative data. However, the emergence of qualitative research has made using small scales confusing. Finally, the Uni-Gram language model, which has integer word counts, is called a multinomial model. The Naive Bayes classifier is a classification method derived from the Bayes Theorem. The main feature of NBC is that it assumes the independence of each condition (Boyles et al. 2007).

Class: It is a nominal class, a binary class, and missing class values.

Attributes: It is a binary attribute, empty nominal attributes, nominal attributes, Numeric attributes, missing values, and unary attributes.

3.2.3.2 Bayes Net A Bayesian network is defined as a type of probabilistic graphical model used for displaying uncertain domain information, where each node corresponds to a random variable, and each edge reflects the conditional probability for those same random variables. Accordingly, Bayesian belief networks provide an intermediate approach less restrictive than the naive Bayes classifier’s universal conditional independence assumption. A Bayesian Network supplies a simple method of applying the Bayes Theorem to complex problems. A Bayesian network captures the joint probabilities of the events represented by the model. A Bayesian belief network describes the joint probability distribution for variables (Maindonald 2007). A Bayesian network is a type of probabilistic graphic model. A probabilistic graphical model (PGM) represents a probabilistic model with a graphed structure. The nodes in the graph represent random variables, and the nodes’ edges represent the relationships between the random variables.

3.2.3.3 Decision table classifiers The decision table is an accurate method that predicts a numeric form from decision trees. It is defined as an ordered set of if–then rules that can be more compact and, therefore, more understandable than the decision trees themselves. It is less computationally intensive since it is a simpler algo-

rithm than the decision-tree-based approach. Decision tables are one of the most straightforward hypothesis spaces possible and are usually easy to understand. Most of the classifiers are built using a decision table, and they evaluate feature subsets using best-first search, which can use cross-validation for evaluation. One option is to use the KNN method to determine the class for each instance that is not covered by a decision table entry instead of the table's global majority based on the same features (Kalmegh 2014)(Othman & Yau, 2007).

3.2.3.4 Lazy classifier Lazy learners store training instances but do not perform any actual work until classification time. Lazy learning is a method in which the training data and generalization are delayed until a query is made for the system. In contrast, the system attempts to generalize the training data before receiving questions. Lazy learning is divided into two categories:

3.2.3.5 Instance-based learning (IBL) IBL is an instance-based learner, predicted by the training instance closest to Euclidean distance according to the training distance. The extensional concept descriptions are not built by the IBL algorithms. Instead, the current set of saved spaces is used by the similarity and classification functions selected by the IBL algorithms to determine concept descriptions (Othman & Yau, 2007). In the following framework that describes every IBL algorithm, where the classification functions and similarity, which represent two of the three components,

Similarity Function: Similarities are numeric values used to calculate the similarity between instances in the concept depiction and training examples.

Classification Function: A classification can be obtained by recording the results of the instances.

Concept Description Updater: The function of Updater is to keep records of the performance of the type and then decide which models to include in the concept description.

3.2.3.6 K-Nearest Neighbor (IBK) The same distance is used in both cases. The object editor computes the nearest neighbours using leave-one-out and cross-validation, focusing on the upper limit given by a certain value. Various search techniques are used to determine the nearest neighbours quickly. The linear search is the default, but other options include ball trees, so-called "cover trees," and KD trees (K-Dimensional). The distance function employed is a search algorithm parameter. The last option, the Euclidean distance, is similar to the IBL; other distances include the Murkowski, Manhattan, and Chebyshev distances (Eminagaoglu 2022). Several neighbours' expectations can be weighted based on their distance from the test instance using two different formulae (Ghosh et al. 2012)(Shazmeen et al. 2013).

3.2.3.7 Trees random forest Random Forest is a well-known and efficient group learning technique. Data in a database table or spreadsheet are often used to identify predictive regression issues. The random forest constructs several decision trees using bootstrap samples from the training dataset (Vijayarani & Sudha 2013).

3.3 Evaluation measurements

The evaluation measurements of a selection of features are computed depending on the following metrics (Brownlee 2016) (Table 2).

4 Proposed methodology

This section outlines the proposed Rule-Based Method (W2vRule) and discusses the Vocabulary, process, Algorithm and Application. The overall process is depicted in Fig. 2, which shows the flow of the machine learning approach and the proposed W2vRule method.

The proposed methodology can be condensed into the following three verticals:

4.1 Data preparation

1. Load the Reuters 21,578-Apte-90 dataset.
2. Filter out documents not in the top ten categories (acq, corn, crude, earn, grain, interest, money-fx, wheat, ship, and trade).
3. Divide the remaining documents into training and testing sets.
4. For each document, apply any necessary pre-processing steps such as removing stopword, stemming, or lemmatization.

4.2 Word embedding

1. Choose a word embedding method, such as skip-gram or CBOW.
2. Train the word embedding model using the training set.

Table 2 Evaluation parameters

Parameters	Formula
Precision	$= \frac{TP}{TP+FP} = \frac{\text{RetrievedRelevent}}{\text{Retrieved}}$,
Recall	$= \frac{TP}{TP+FN} = \frac{\text{RetrievedRelevent}}{\text{Relevent}}$,
F Measure	$= \frac{2(TP)}{FP+FN+2(TP)} = \frac{2(\text{RetrievedRelevent})}{\text{Relevent+Retrieved+2(Retrieved Relevent)}}$,
Accuracy	$= \frac{TP+TN}{TP+TN+FP+FN}$,
Error Rate Inverse of Accuracy	$= \frac{FP+FN}{TP+TN+FP+FN}$

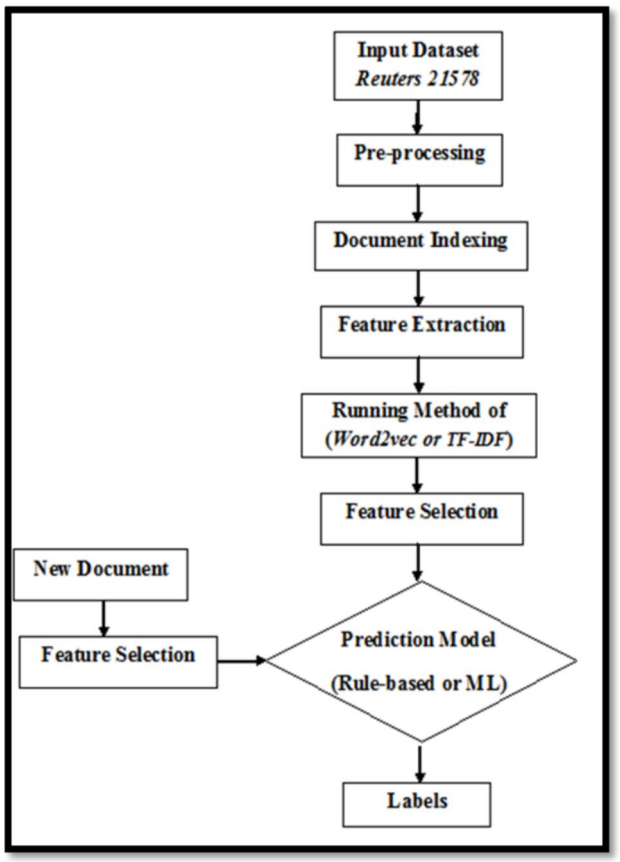

Fig. 2 Steps of rule-based method

3. Use the trained model to generate embeddings for each word in the vocabulary.
4. Rule-Based Approach:
5. Define a set of rules for classifying documents based on their word embeddings.
6. Use the word embeddings for each document in the testing set to classify them according to the defined rules.

4.3 Machine learning approach

1. Compute the term frequency-inverse document frequency (TF-IDF) weights for each term in the training set using C++.
2. Use the TF-IDF weights to represent each document as a vector of term weights.
3. Train six different machine learning models (Naive Bayes, NaiveBayesUpdateable, Decision Tables, Lazy IBL, Lazy IBK, and Random Forest) on the training set using Weka7.
4. Use the trained models to classify documents in the testing set based on their vector representations.

Table 3 Definition of Algorithm's parameters

Item	Parameters	Definitions
1	C	Set of classes
2	Di	Set of document
3	D	Document
4	W	Word
5	N	Number of documents
6	I	Number of a word in sentences
7	S	Sentences (w_1 〖,w〗 _2,….., w_i)
8	M	Number of sentences in document
9	Ci	class {D_1, D_2,….,D_N}

This study uses skip-gram to generate embeddings for each word in the Reuters 21,578-Apte-90 dataset, which is used to train a subroutine and develop a rule-based approach to classify documents. Table 3 defines some algorithm parameters.

To construct the feature selection or vocabulary (important word) in the next steps of the text classification system, the algorithm does certain pre-processing, as shown below.

5 Algorithm 1: Preprocessing

Input: Document D
Procedure:
for i,j =0
for each sent in D
if Word not in stopword
Stemming word
T[j] = word
j=j+1
end
Output: Term T

5.1 Rule-based method processing (W2vRule)

Figure 3 illustrates the pre-processing processes for the rule-based method. This figure also presents the strategy for the rule-based system used for the dataset (training documents and testing documents).

5.2 Preparing the vocabulary

It is critical to include commonly used words in a document's vocabulary to improve its vocabulary. We used Word-2Vec to analyse the unique characteristics of each document class to achieve this goal. These Word2Vec categories were

used to train the vocabularies, reducing phrase repetition by identifying similarities in definition, terminology, or both. Furthermore, considering synonyms aided in the selection of more appropriate word choices. The model's similarity comparisons were performed in every sentence to ensure consistency throughout the document. The algorithm used to build the W2vRule model is described in the following section.

6 Algorithm 2: Creation of the W2vRule model

Input: T: Set of term (for each class), N: Number of Documents in corpus, i= $term_i$, j= document t_j
Procedure:

for each term in T
compute term frequency (tf_{ij})
compute inverse document (idf)
compute $weight_{ij} = tf_{ij} * idf$
vector [ ] = $weight_{ij}$
end
for all classes in corpus
build model (word2vec)
end

Output: Vector of terms, Model

The information gain (IG) (Sanderson 2010)(Lee & Lee 2006) is used to choose a feature (informative word). Some researchers, such as (Shang et al. 2013), consider a feature selection problem to be an estimation problem. In contrast, others (Li & Zhang 2008), pick these features or terms during the training of a clustering model using the Expectation–Maximization (EM) technique. According to the Guozhong Feng et al. principle, the unsupervised version of the method proposed by (Feng et al. 2012) was constructed through term frequency. This study requires selecting informative words (vocabularies) for developing a rule-based classification system (W2vRule) for text documents from the Reuter corpus. The cosine approach determines those vocabularies in the text documents from the Reuter corpus. Algorithm 3 depicts the algorithm developed to select features and use them later in a rule-based environment.

7 Algorithm 3: Feature selection

Input: Vocabulary
Procedure:

for each vocabulary in Vector (features)
compute cosine similarity
end
sort (vector similarity in ascending)
for each vocabulary in vector(sorting)
compute normalisation

$$\mathrm{X}normalised = \frac{\mathrm{x} - \mathrm{x}min}{\mathrm{x}max - \mathrm{x}min}$$

end
for each vocabulary in vector(normalisation)
compute threshold value(vocabulary)
end
sort (keyword ascending)
end

Output: Keywords (Selected features by informative words)

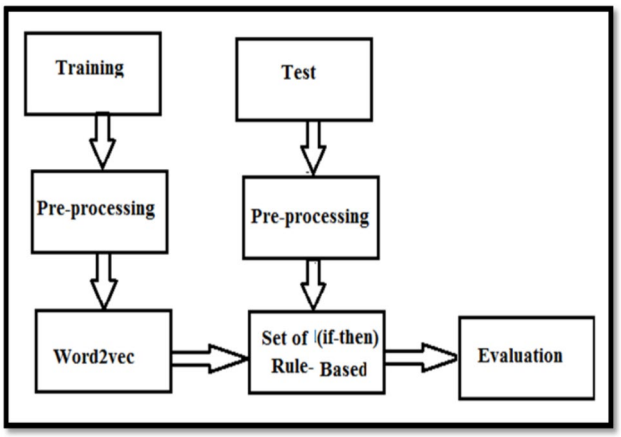

Fig. 3 Steps of Rule-based techniques applied to the dataset

Table 4 The vocabulary of categories with similarities

Item	Category	Terms	Similarity value
1	Acq	shareholder	0.998768
2	Corn	farmer	1.0
3	Crude	oil	0.9985847

7.1 Rule-based process structure

The following steps illustrate the rule-based process's structure:

1. After extracting the terms, similarity values are computed. Values of similarity must be larger than the criterion. The point at which a program is executed changes at a threshold value. The threshold value, which is utilized to identify the important words in these papers, is indicated by how similar the terms are in the documents.
2. A feature selection metric's main objective is to assign valid words using several metrics, such as similarity, weight, term frequency, and so on while rejecting non-informative terms, improving the task of the text classification technique. This indicates that the selected key terms (keywords) appear in the documents most frequently.
3. The documents for the exam component are picked.
4. The documents in the test portion of the dataset have been pre-processed.
5. The classification process starts when the training part terms are defined, such as one word for each document in the test section.

7.2 Algorithm

The following instructions are representative of rule-based instructions: If ("maiz" in doc, "shipment" in doc, "ton" in doc, or "corn" in doc), category = "corn". The rule-based categories are checked according to classification rules and categories of the dataset, followed by an evaluation of the measurements being computed. Finally, Fig. 6 illustrates the procedure of the rule-based algorithm for one category out of the ten top categories in the dataset. Algorithm 4 introduces the functions of the Word2vec classifier for one category.

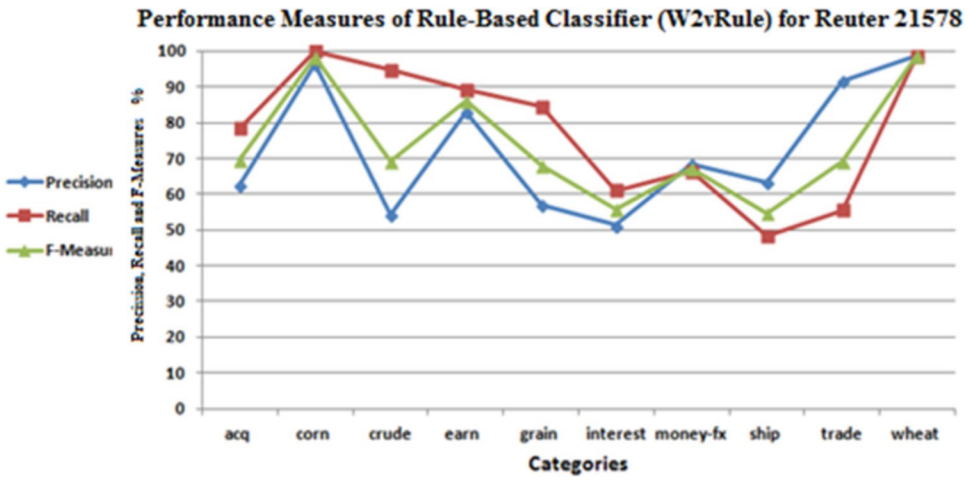

Fig. 4 Rule-Based Metric Measurements for W2vRule model on Reuters 21,578

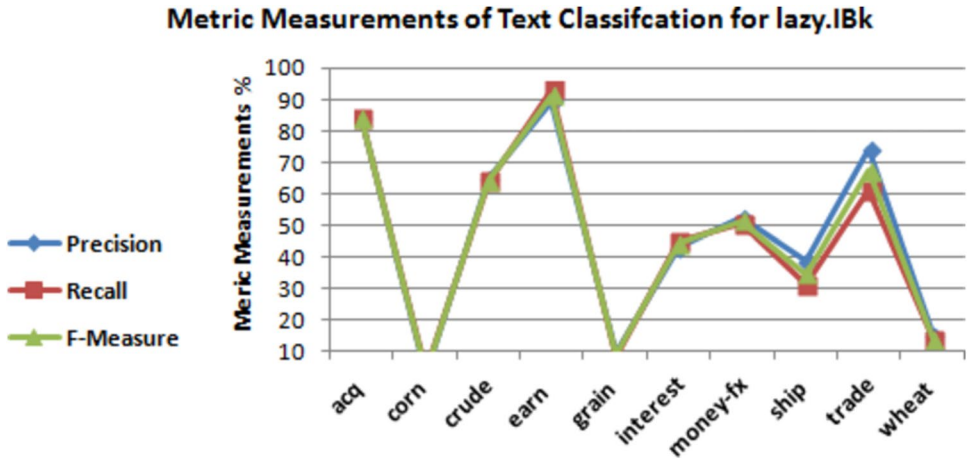

Fig. 5 Lazy IBK Metric Measurements on Reuters 21,578

8 Algorithm 4: Rule-based category classifier with Word2vec

Input: Class of test corpus, Keyword (Vocabulary)
Procedure:

1. *For each category in Categories, create a list of keywords based on the Keyword Vocabulary.*
2. *For each document in the Test corpus, apply the following rule-based instructions:*
3. *For each category in Categories, check if the document contains any of the keywords in the corresponding list. If the document matches, assign the category to the document.*
4. *For each category in Categories, compute the weight of each term in the corresponding documents using tf-idf weighting.*
5. *For each category in Categories, train a Word2Vec model on the pre-processed documents (i.e., cleaned documents with stop words removed, stemmed, and converted to lowercase) of that category.*
6. *For each document in the Test corpus, apply the following procedure:*
   - *If the document has already been assigned to a category in step 2, skip stepping 6. Otherwise, compute the cosine similarity between the pre-processed document and each category's Word2Vec model.*
   - *Assign the document to the category with the highest cosine similarity. If there is a tie, assign the document to the category with the highest tf-idf weighted term for that document in that category.*
7. *Output the classified documents with their assigned categories.*

Output: Classified documents with their assigned categories

Example: If "maiz" in doc or "shipment" in doc or "ton" in doc or "corn" in doc, category = "corn". If "wheat" is in doc or "shipment" in doc or "ton" in doc or "grain" in doc, category = "wheat".

8.1 Application of rule-based instructions

In the following, an example from the dataset will be taken, such as the acq, corn, and crude categories, and the rule-based instructions will be built according to the following steps:

1. A selection of vocabulary from the acq, corn, and crude types is used, such as shareholder, farmer, and oil.
2. In Table 4, the vocabularies for the acq, corn, and crude categories are ordered according to their similarity to terms in the train and test data sets using word2vec. This is a critical step because it establishes the category's essential vocabulary.

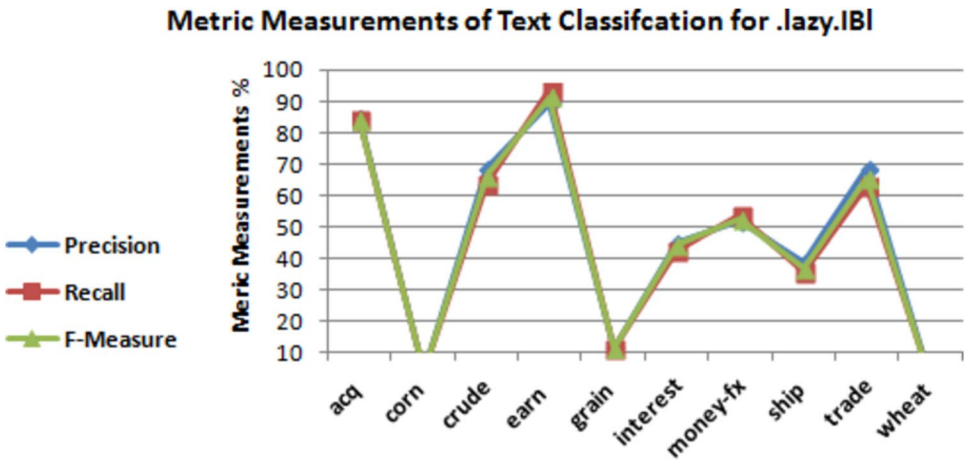

Fig. 6 Lazy IBL Metric Measurements on Reuters 21,578

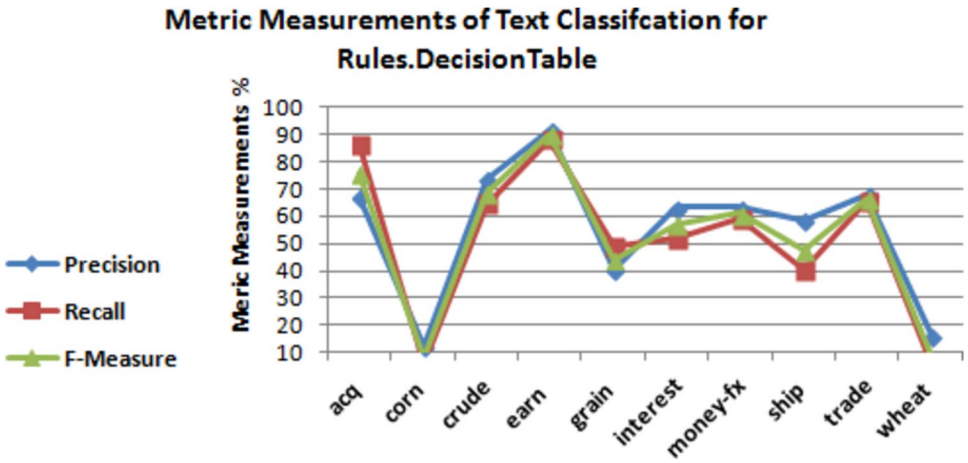

Fig. 7 Rules-Decision Table Metric Measurements on Reuters 21,578

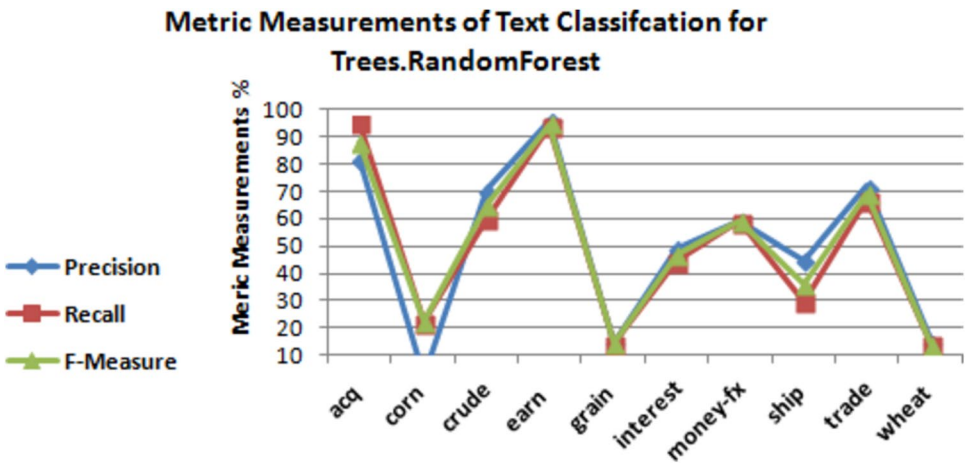

Fig. 8 Trees.RandomForest Metric Measurements on Reuters 21,578

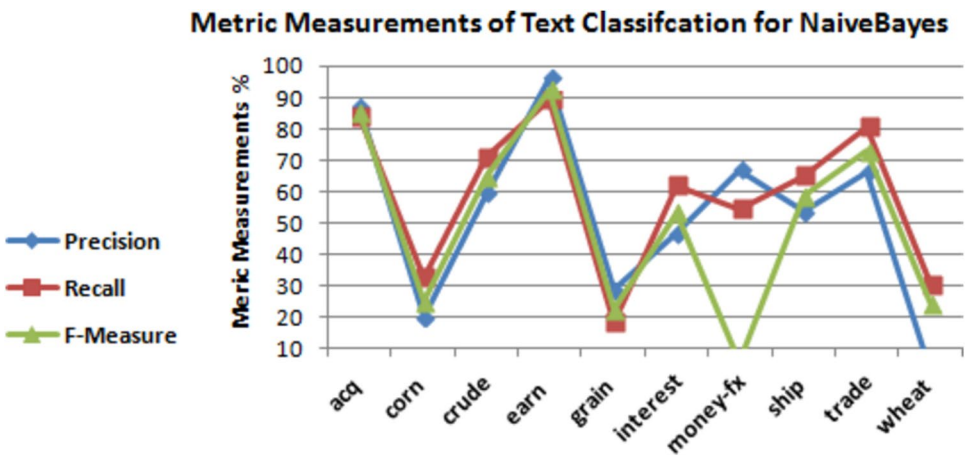

Fig. 9 Naive Bayes's Metric Measurements on Reuters-21578

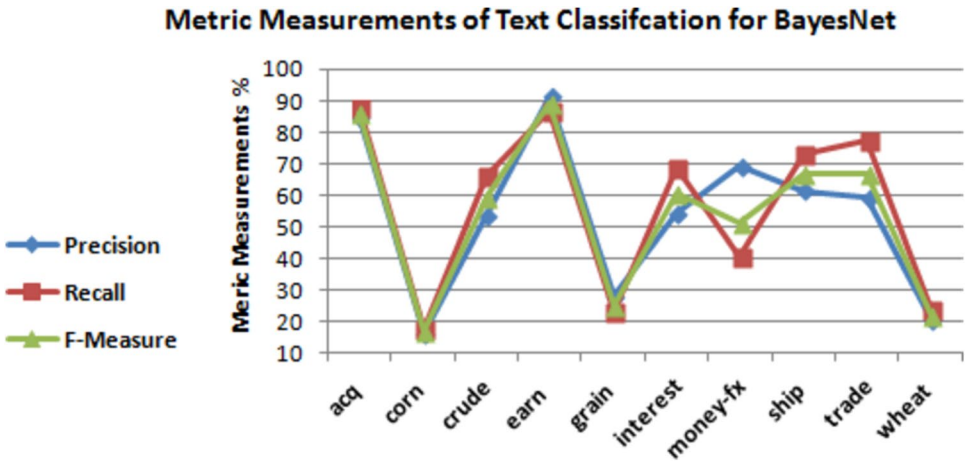

Fig. 10 Bayes Net's Metric Measurements on Reuters-21578

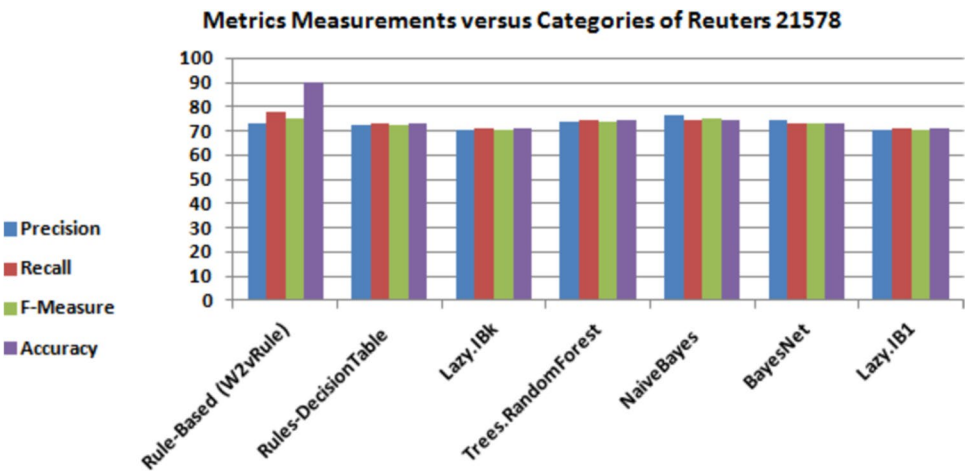

Fig. 11 Variation in in metric measurements of the text classification model across various categories

Table 5 Experimental results of different methodologies for Reuter 21,578

Methodologies	Precision	Recall	F-Measure	Accuracy
Rule-Based (W2vRule)	73%	77.71%	75.09%	89.91%
Rules DecisionTable	72.1%	73.2%	72.1%	73.1962%
Lazy IBK	70.3%	70.8%	70.5%	70.7771%
Lazy IBL	70.4%	71.1%	70.7%	70.0691%
Trees.RandomForest	73.7%	74.3%	73.8%	74.2666%
Naïve Bayes	76.7%	74.6%	75.3%	74.6281%
Bayes Net	74.2%	73.4%	73.3%	73.3769%

3. Providing documents for tests, such as acquisition documents.
4. Execution of rule-based commands.
5. Comparing the document to the test (rule-based) and classifying it.

Finally, the results are evaluated using precision and recall parameters. The preceding Fig. 3 illustrates the rule-based process using a flow chart.

9 Results

This section investigates the effectiveness of the proposed approach and the Performance of Different Models Across All Classes in the Reuters 21,578 Dataset. Using our rule-based methodology, this study investigates the categories, including acquisition, crude, corn, earn, interest, grain, money-fx, wheat, ship, and trade.

The goal of this study was to evaluate the performance of different models for categorizing text into various classes based on the Reuters 21,578 dataset. To do this, a rule-based technique was used to categorize the text into different classes, and the metric measurements were analyzed. The frequency of words (number of features in documents) was set to 50, which is close to the value used in machine learning models. The metric evaluations of the W2vRule model are presented in Fig. 4. The Lazy-IBk and Lazy-IBL metric measurements are compared in Figs. 5 and 6, while the Rules-Decision Table metric Measurements are compared in Fig. 7. The Trees Random Forest metric measurements are compared to those shown in Fig. 8, but in Fig. 9 depicts the metric measurements of the Naive Bayes algorithm, and Fig. 10 depicts the Bayes Net method's measurements. It's important to note that these figures present the distinction between these models. Finally, Fig. 11 shows the range of possible values for the metric sizes used in sorting text into various categories.

The experimental results of the different methodologies for the Reuter 21,578 of the top ten categories are illustrated in Table 5.

Figure 11 presents the breakdown of the range of possible values for the metric sizes employed in sorting text into its many categories. This information is useful in understanding the distribution of the metric values and can help determine which values are optimal for a given task. By examining the range of possible values, one can make informed decisions about which metrics to use and how to adjust the parameters of the models to achieve the best performance.

10 Discussion

It is crucial for rule-based and autonomous learning techniques to improve the efficiency and effectiveness of information retrieval technology. Numerous methods, including artificial neural networks (Minaee et al. 2021; Sinha et al. 2021), Bayesian learning (Ying & Mursitama 2021), and rule-based learning (Liu & Beldona 2021), allows computers to decide on an appropriate option in decision-making. This section examines the Naive Bayes, Bayes Net, and Naive Bayes Rules DecisionTables, RandomForest, Lazy IBK, Lazy IBL, and Lazy IBL methods for problem-solving. In contrast, numerous techniques for performing decision analysis are available (Avasthi et al. 2021; Liang & Yi 2021).

Each method has its advantages and limitations. Bayesian learning is one of the most widely used techniques in machine learning for classification decisions in pattern recognition. The rule-based method is one of the most adaptable because it allows for visualizing the text categorization method's black box. The framework uses open-source tools and technologies such as C++, Weka, and Python. The word embedding method is applied to represent text documents, which proved to be more appropriate when preparing a data set. As a result, rule-based and machine-learning processes would aid text categorization. This is done by optimizing the classification metrics by adjusting the term frequency (tf), defined as the number of times essential words appear in documents, to 50. An accuracy (more significant than others for machine learning) is associated with a rule-based approach to categorizing test data sets. The other metrics are greater or equal to their values for machine learning models.

On the other hand, it demonstrated that the recall, precision, error rate, and accuracy of a rule-based approach for classifying text documents are more significant than their values, particularly in the case of machine learning models. For the following reasons, a rule-based approach is the best

technique for text classification: Independence is a defining feature of rule-based systems. Second is the naturalness of expression, which may incorporate an expert's knowledge as a guide's norm. Thirdly, the restricted syntax generates the rules, and additional programs check for consistency. Fourthly, a problem can be represented using a compact generic knowledge representation. Finally, the explanation presentation presents critical aspects of rule-based systems (Meelen et al. 2021).

We conducted a comprehensive investigation to facilitate discussion and comparison of our findings with those of other studies. Ligeza (Ligeza 2006) discussed symbolic rules, some of the most well-known techniques for representing and inferring knowledge (Mishra et al. 2018). Associative classification techniques have been applied to a variety of categorization tasks due to their simplicity and high accuracy of between (80 and 86%) for minimum confidence (min. conf, equals to 70%) (Yoon & Lee 2013). It has been shown, by applying massive volumes of text as experimental data, that the proposed algorithms improve the viability of using associative categorization for large-scale challenges. The search method (SFS), which is not dependent on the bagging algorithms, achieved good results of (89.60 per cent) and appeared to be superior to other techniques (Panthong & Srivihok 2015).

Finally, the classification experiments were conducted using several sets of word unigrams and different machine learning methods, with acceptable results for all six measures: accuracy, precision, recall, F1, PRC-area, and ROC-area. It is evident that the proposed method achieved an accuracy of 89.91 per cent for ten categories, which is higher than the results obtained for seven categories with fewer documents (L. Zhang & Duan 2019). Based on these factors, it can be concluded that rule-based (W2vRule) text classification is superior to other methods when it comes to text classification approaches when compared to other techniques.

11 Conclusions and future research

This study used six different algorithms, including Nave Bayes, NaiveBayesUpdateable, Decision Tables, Lazy IBL, Lazy IBK, and rule-based models. These algorithms are used to categorize the text documents in the datasets. These were also incorporated into our rule-based methodology (Word2vec) for evaluations based on metrics such as recall, precision, F-measures, and accuracy. As documents proliferate, so do categories. This study discovered that our results for the ten types is significantly higher, at 89.91 percent than the results for seven classes. Based on these considerations, rule-based (Word2Vec) text classification outperforms other approaches.

As a future research direction, the study intends to improve the rules-based approach with the other dataset by selecting new knowledge-based methods. A new system may be developed by combining rule-based and knowledge-based techniques. Including SVM and LSTM, along with other popular algorithms, would provide a more comprehensive comparison and give a better idea of the performance of the proposed method in comparison to the state-of-the-art algorithms for text classification. It would also provide valuable insights into the strengths and weaknesses of the proposed method in comparison to these algorithms.

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital). This research received no external funding.

Data availability The Reuters-21578 collection, for example, is a publicly accessible version of the well-known Reuters-21578 "Apte-Mod" corpus for text categorization. Then, Reuters Ltd. (S. Weinstein, S. Dobbins, and M. Topliss) and the Carnegie Group, Inc. (M. Cellio, P. Andersen, P. Hayes, Ir. Nirenburg, and L. Knecht) collected and indexed these documents according to specific categories. The collection is distributed into 22 files. Each of the first 21 files (reut2-000.sgm through reut2-020.sgm) contains 1000 documents, while the last (reut2-021.sgm) contains 578 documents.

Declarations

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

References

Agrawal R, Batra M (2013) A detailed study on text mining techniques. Int J Soft Comput Eng 2(26):2231–2307

Aubaid AM, Mishra A (2018) Text classification using word embedding in Rule-based methodologies: a systematic mapping. TEM J 7(4):902–914. https://doi.org/10.18421/TEM74-31

Aubaid AM, Mishra A (2020) A rule-based approach to embedding techniques for text document classification. Appl Sci (Switzerland) 10(11):4009. https://doi.org/10.3390/app10114009

Avasthi S, Chauhan R, Acharjya DP (2021) Techniques, applications, and issues in mining large-scale text databases. Advances in information communication technology and computing. Springer, Singapore, pp 385–396

Balli C, Guzel MS, Bostanci E, Mishra A (2022) Sentimental analysis of twitter users from turkish content with natural language processing. Comput Intell Neurosci. https://doi.org/10.1155/2022/2455160

Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, Chapman B, Amrhein T, Mong D, Rubin DL, Farri O, Lungren MP (2019) Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med 97:79–88. https://doi.org/10.1016/j.artmed.2018.11.004

Basu T, Murthy CA (2016) A supervised term selection technique for effective text categorization. Int J Mach Learn Cybern 7(5):877–892. https://doi.org/10.1007/s13042-015-0421-y

Batrinca B, Treleaven PC (2015) Social media analytics: a survey of techniques, tools and platforms. AI & Soc 30(1):89–116. https://doi.org/10.1007/s00146-014-0549-4

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146. https://doi.org/10.1162/tacl_a_00051

Boyles S, Fajardo D, Waller ST (2007) Naive bayesian classifier for incident duration prediction. Transportation Research Board 86th Annual Meeting, 253(07–1801). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.526.3396&rep=rep1&type=pdf

Brownlee J (2016) Machine learning mastery with python: understand your data, create accurate models, and work projects end-to-end. Machine learning mastery.

Çano E, Morisio M (2019) Word embeddings for sentiment analysis: a comprehensive empirical survey. *ArXiv Preprint* ArXiv:1902.00753.

Corrales DC, Lasso E, Ledezma A, Corrales JC (2018) Feature selection for classification tasks: expert knowledge or traditional methods? J Intell Fuzzy Syst 34(5):2825–2835. https://doi.org/10.3233/JIFS-169470

Cui M, Huang R, Hu Z, Xia F, Xu X, Qi L (2024) Semantic rule-based information extraction for meteorological reports. Int J Mach Learn Cybern 15(1):177–188. https://doi.org/10.1007/s13042-023-01885-8

Deng X, Li Y, Weng J, Zhang J (2019) Feature selection for text classification: a review. Multimed Tool Appl 78(3):3797–3816. https://doi.org/10.1007/s11042-018-6083-5

Ducoffe M, Mayaffre D, Precioso F, Lavigne F, Vanni L (2016) Machine Learning under the light of Phraseology expertise : use case of presidential speeches De Gaulle-Hollande. In: *JADT 2016-Statistical Analysis of Textual Data*, *1*, 157–168.

Dwivedi SK, Arya C (2016) Automatic text classification in information retrieval: a survey. ACM Int Conf Proc Ser. https://doi.org/10.1145/2905055.2905191

Eminagaoglu M (2022) A new similarity measure for vector space models in text classification and information retrieval. J Inf Sci 48(4):463–476. https://doi.org/10.1177/0165551520968055

Feng G, Guo J, Jing B-Y, Hao L (2012) A Bayesian feature selection paradigm for text classification. Inf Process Manage 48(2):283–302

Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3(Mar):1289–1305

Franks J (2022) Text classification for records management. J Comput Cultural Heritage 15(3):1–19. https://doi.org/10.1145/3485846

Ghosh S, Roy S, Bandyopadhyay SK (2012) A tutorial review on text mining algorithms. Int J Adv Res Comput Commun Engneering 1(4):7

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software. ACM SIGKDD Explorations Newsl 11(1):10–18. https://doi.org/10.1145/1656274.1656278

Helaskar MN, Sonawane SS (2019) Text classification using word embeddings. *Proceedings - 2019 5th International Conference on Computing, Communication Control and Automation, ICCUBEA 2019*, 1–4. https://doi.org/10.1109/ICCUBEA47591.2019.9129565

Hmeidi I, Al-Ayyoub M, Abdulla NA, Almodawar AA, Abooraig R, Mahyoub NA (2015) Automatic Arabic text categorization: a comprehensive comparative study. J Inf Sci 41(1):114–124. https://doi.org/10.1177/0165551514558172

Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, *2*, 427–431. https://doi.org/10.18653/v1/e17-2068

Kalmegh SR (2014) Effective evaluation of classification of indigenous news using decision table and OneR algorithm. Int J Adv Inform Sci Technol (IJAIST) 26(26):6–11

Ku CH, Leroy G (2014) A decision support system: Automated crime report analysis and classification for e-government. Gov Inf Q 31(4):534–544. https://doi.org/10.1016/j.giq.2014.08.003

Lee C, Lee GG (2006) Information gain and divergence-based feature selection for machine learning-based text categorization. Inf Process Manage 42(1):155–165

Levy O, Goldberg Y (2014) Linguistic regularities in sparse and explicit word representations. *CoNLL 2014 - 18th Conference on Computational Natural Language Learning, Proceedings*, 171–180. https://doi.org/10.3115/v1/w14-1618

Li M, Zhang L (2008) Multinomial mixture model with feature selection for text clustering. Knowl-Based Syst 21(7):704–708

Liang D, Yi B (2021) Two-stage three-way enhanced technique for ensemble learning in inclusive policy text classification. Inf Sci 547:271–288

Ligeza A (2006) Logical foundations for rule-based systems. In: *Logical foundations for rule-based systems*. Springer: Berlin

Liu X, Tang H, Ding Y, Yan D (2022) Investigating the performance of machine learning models combined with different feature selection methods to estimate the energy consumption of buildings. Energy Build 273:112408. https://doi.org/10.1016/j.enbuild.2022.112408

Liu Y, Beldona S (2021) Extracting revisit intentions from social media big data: a rule-based classification model. Int J Contemp Hospital Manage 33(6):2176–2193

Maindonald J (2007) Pattern recognition and machine learning. J Stat Softw https://doi.org/10.18637/jss.v017.b05

Mao R, He K, Zhang X, Chen G, Ni J, Yang Z, Cambria E (2024) A survey on semantic processing techniques. Inform Fus 101:101988. https://doi.org/10.1016/j.inffus.2023.101988

Martinelli F, Mercaldo F, Nardone V, Santone A, Vaglini G (2018) Real-time driver behaviour characterization through rule-based machine learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11094 LNCS*, 374–386. https://doi.org/10.1007/978-3-319-99229-7_32

Meelen M, Roux É, Hill N (2021) Optimisation of the largest annotated Tibetan corpus combining rule-based, memory-based, and deep-learning methods. ACM Trans Asian Low-Resour Lang Inform Process (TALLIP) 20(1):1–11

Melville P, Gryc W, Lawrence RD (2009) Sentiment analysis of blogs by combining lexical knowledge with text classification. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1275–1283. https://doi.org/10.1145/1557019.1557156

Mendel JM (2017) Uncertain rule-based fuzzy systems. *Introduction and New Directions*, 684.

Mikolov T, Deoras A, Povey D, Burget L, Černocký J (2011) Strategies for training large scale neural network language models. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings*, 196–201. https://doi.org/10.1109/ASRU.2011.6163930

Mimaroglu DS (2020). Reuters-21578 text categorization collection. In *College of Science and Mathematics*. https://www.cs.umb.edu/~smimarog/textmining/datasets/

Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J (2021) Deep learning–based text classification: a comprehensive review. ACM Comput Surv (CSUR) 54(3):1–40

Mishra D, Aydin S, Mishra A, Ostrovska S (2018) Knowledge management in requirement elicitation: situational methods view. Comput Stand Interfaces 56:49–61. https://doi.org/10.1016/j.csi.2017.09.004

Mohsen A, Ali Y, Al-Sorori W, Maqtary NA, Al-Fuhaidi B, Altabeeb AM (2021) A performance comparison of machine learning classifiers for Covid-19 Arabic Quarantine tweets sentiment analysis. *2021 1st International Conference on Emerging Smart Technologies and Applications, ESmarTA 2021*, *16*(2), e0245909. https://doi.org/10.1109/eSmarTA52612.2021.9515749

Mondal N, Lohia M (2020) *Supervised text classification using text search*. http://arxiv.org/abs/2011.13832

Myaeng SH, Han KS, Rim HC (2006) Some effective techniques for naive Bayes text classification. IEEE Trans Knowl Data Eng 18(11):1457–1466. https://doi.org/10.1109/TKDE.2006.180

Onan A (2018) An ensemble scheme based on language function analysis and feature engineering for text genre classification. J Inf Sci 44(1):28–47. https://doi.org/10.1177/0165551516677911

Onan A (2019) Topic-enriched word embeddings for sarcasm identification. Adv Intell Syst Comput 984:293–304. https://doi.org/10.1007/978-3-030-19807-7_29

OthmanBin Yau MFTMS (2007) Comparison of different classification techniques using WEKA for breast cancer. IFMBE Proceedings 15:520–523. https://doi.org/10.1007/978-3-540-68017-8_131

Panthong R, Srivihok A (2015) Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. Procedia Comput Sci 72:162–169

Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543. https://doi.org/10.3115/v1/d14-1162

Pereira RB, Plastino A, Zadrozny B, Merschmann LHC (2021) A lazy feature selection method for multi-label classification. Intell Data Anal 25(1):21–34. https://doi.org/10.3233/IDA-194878

Pintas JT, Fernandes LAF, Garcia ACB (2021) Feature selection methods for text classification: a systematic literature review. Artif Intell Rev 54(8):6149–6200. https://doi.org/10.1007/s10462-021-09970-6

Pong JYH, Kwok RCW, Lau RYK, Hao JX, Wong PCC (2008) A comparative study of two automatic document classification methods in a library setting. J Inf Sci 34(2):213–230. https://doi.org/10.1177/0165551507082592

Ranjan NM, Prasad RS (2023) A brief survey of text document classification algorithms and processes. J Data Min Manage 8(1):6–11

Sanderson M (2010) Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, introduction to information retrieval Cambridge university press 2008. ISBN-13 978-0-521-86571-5, xxi + 482 pages. Nat Lang Eng 16(1):100–103

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47. https://doi.org/10.1145/505282.505283

Shahi TB, Sitaula C, Paudel N (2022) A hybrid feature extraction method for Nepali COVID-19-related tweets classification. Comput Intell Neurosci. https://doi.org/10.1155/2022/5681574

Shang C, Li M, Feng S, Jiang Q, Fan J (2013) Feature selection via maximizing global information gain for text classification. Knowl-Based Syst 54:298–309

Shazmeen SF, Baig MMA, Pawar MR (2013) Performance evaluation of different data mining classification algorithm and predictive analysis. J Comput Eng 10(6):1–6

Sinha S, Ghosh I, Satapathy SC (2021) A study for ANN model for spam classification. Intelligent data engineering and analytics. Springer, Singapore, pp 331–343

Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L (2020) Interpretability of machine learning-based prediction models in healthcare. Wiley Interdisciplinary Rev: Data Min Knowl Discover 10(5):e1379. https://doi.org/10.1002/widm.1379

Tailor C, Patel B (2019) Sentence tokenization using statistical unsupervised machine learning and rule-based approach for running text in gujarati language. Advances in intelligent systems and computing. Springer, Singapore, pp 319–326

Tao D, Yang P, Feng H (2020) Utilization of text mining as a big data analysis tool for food science and nutrition. Compr Rev Food Sci Food Saf 19(2):875–894. https://doi.org/10.1111/1541-4337.12540

Vijayarani S, Sudha S (2013) Comparative analysis of classification function techniques for heart disease prediction. Int J Innov Res Comput Commun Eng 1(3):735–741

Wankhade M, Rao ACS, Kulkarni C (2022) A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55(7):5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Weinberger K, Dasgupta A, Langford J, Smola A, Attenberg J (2009) Feature hashing for large scale multitask learning. *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, 1113–1120.

Wibowo W, Williams HE (2002) Simple and accurate feature selection for hierarchical categorisation. *Proceedings of the 2002 ACM symposium on document engineering*, 111–118. https://doi.org/10.1145/585058.585079

Witten IH, Frank E, Geller J (2002) Data mining: practical machine learning tools and techniques with java implementations. SIGMOD Record 31(1):76–77. https://doi.org/10.1145/507338.507355

Ying Y, Mursitama TN (2021) Effectiveness of the news text classification test using the naïve Bayes' classification text mining method. J Phys: Conf Ser 1764(1):12105

Yoon Y, Lee GG (2013) Two scalable algorithms for associative text classification. Inf Process Manage 49(2):484–496

Zhang C (2024) Improved word segmentation system for Chinese criminal judgment documents. Appl Artif Intell 38(1):2297524. https://doi.org/10.1080/08839514.2023.2297524

Zhang L, Duan Q (2019) A feature selection method for multi-label text based on feature importance. Appl Sci (Switzerland) 9(4):665. https://doi.org/10.3390/app9040665

Zhang R, El-Gohary N (2021) A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking. Autom Constr 132:103834