Probabilistic Genetic Optimization for Text Categorization using Bayesian Models

Soumya Sangam Jha
Information Technology
National Institute of Technolgy Karnataka, Surathkal
211IT068

Vartika T. Rao
Information Technology
National Institute of Technolgy Karnataka, Surathkal
211IT077

Saatvik Krishna
Information Technology
National Institute of Technolgy Karnataka, Surathkal
211IT056

Subhojit Karmakar
Information Technology
National Institute of Technolgy Karnataka, Surathkal
211IT071

***Abstract*—Text categorization is the process of assigning text documents to predefined categories. It is a challenging task, especially when dealing with large volumes of unstructured text. One of the main challenges is feature selection. Features are the distinguishing characteristics of a text document, such as words, phrases, or topic-specific keywords. In large datasets, the number of features can be overwhelming, leading to the risk of overfitting.**

To address this challenge, we propose a Genetic Algorithm based feature selection technique that uses a probabilistic feature selection approach. Unlike usual implementations of genetic algorithm-based feature selection, our proposed method implements it as a filter method rather than a wrapper method, enhancing computational efficiency without compromising accuracy. We also implement a chi-square-based feature selection technique to compare our results with it.

To test our feature selection techniques, we use Multinomial Naive Bayes (MNB), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM) to do the binary classification task of categorizing news documents.

***Index Terms*—Text categorization (TC), Naive Bayes(NB), bayesian probabilistic models, feature selection, hypothesis testing, chi-square test, genetic algorithms, support vector machines (SVM), stochastic process optimization.**

I. Introduction

A huge amount of information is available in different types of fields like sports or science. These information are stored in unstructured documents. Thereby, it must be arranged and stored in such a way that users reach to those information. Text categorization (TC) can aid user to solve this problem. There are two basic types of rule-based methods to TC. The first rule-based approach represents the categorization procedures were physically generated by scientists in the range of the textual data. Despite of the rule-based method may realize high performance but it is costly in terms of work and time. The second method contains supervised learning mechanisms that categorization procedures are automatically generated via utilizing information from categorized (predefined-classified) documents. Supervised learning is considered as cost-economy due to it demands only categorized texts. From among different several techniques that are suggested for feature selection (FS), population-based optimization methods like genetic algorithm (GA)-based algorithm has been attracted much attention. Thus, these techniques try to realize the best solutions via applying of knowledge from prior repetitions. where genetic algorithms are considered as the optimization methods based on the mechanism of the naturalistic selection.

II. Background

A. Dataset Used

The dataset employed in this study consists of 200 datapoints, encompassing sports and sci-tech news articles, making it suitable for binary text classification. The dataset is curated to represent a balanced distribution of articles from both domains, ensuring a comprehensive evaluation of the classification model. Each datapoint includes relevant textual content and corresponding labels indicating whether the article falls under the sports or sci-tech category. This curated dataset serves as the foundation for training and evaluating the text classification model in the subsequent sections.

B. Literature Review

In the pursuit of a comprehensive understanding of the existing landscape within our research domain, we conducted a survey of pertinent literature, encompassing a diverse selection of 24 research papers. Their key findings and gaps are summarized as follows.

The paper [1] introduces a novel hybrid feature selection technique, combining chi-square and a genetic algorithm for text categorization. It evaluates the approach using Naïve Bayes and C4.5 classifiers on BBC sport and news datasets, demonstrating improved performance. While serving as a foundation, the paper solely explores wrapper-based feature selection,

prompting us to extend research by incorporating embedded methods and other Bayesian models for text categorization.

The paper [2] extensively reviews supervised machine learning applications in text categorization, emphasizing the shift from manual to automated processes. It covers techniques like k-NN, Naïve Bayes, and SVM, discussing their enhancements and combinations for improved performance. The exploration of external knowledge and meta-features reveals superior results, such as combining Naïve Bayes with external enriching. The paper underscores the importance of sentiment analysis and opinion mining in text categorization.

The paper [3] introduces a feature subset selection framework employing a genetic algorithm (GA) to enhance predictive accuracy and model comprehensibility. It stands out for its capacity to integrate various feature selection methods, catering to diverse criteria. Serving as a foundational work, it establishes the groundwork for utilizing GA in feature selection processes.

The study [4] investigates and compares multiple fitness functions in a genetic algorithm (GA) for feature selection, aiming to strike a balance between feature selection and classification accuracy. The paper delves into the impact of diverse fitness functions on the transformation process, highlighting variations in prioritizing classification accuracy maximization and feature subset size minimization. The user intends to experiment with these fitness functions to optimize their approach.

The paper [5] focuses on feature selection in text classification, utilizing Chi-square Statistics to extract feature words. A notable aspect is our simultaneous consideration of both single and double words as features. Through experiments with Naive Bayes and Support Vector Machine algorithms, we demonstrate the method's efficiency. However, we acknowledge a limitation: the Chi-Square method exclusively focuses on the frequency of feature words in documents, neglecting the impact of word frequency. We aim to address this limitation in our project.

The paper [6] investigates Multivariate Bernoulli and Multinomial Naïve Bayes Text Categorization for sentiment prediction in news articles. Comparing their performance, the study reveals that Multinomial Naïve Bayes surpasses Bernoulli Naïve Bayes, especially with a smaller dataset of 312 records. Emphasizing the challenge of achieving high accuracy with limited data, the paper suggests larger datasets could enhance accuracy for both methods. While Multinomial Naïve Bayes slightly outperforms, the difference is not substantial, as Bernoulli Naïve Bayes still achieves around 69 percent accuracy on the provided dataset.

Dawar and Kumar [7] explore using the Naïve Bayes algorithm for text categorization in this paper. They argue that the Naïve Bayes algorithm is a simple and effective method for text categorization and that it is particularly well-suited for tasks where computational resources are limited or where labelled data is scarce. Our research, primarily focusing on feature selection with GA and chi-square, benefits from acknowledging the strengths of the Naïve Bayes approach, a widely used method in text classification.

The paper [8] utilizes deep learning approaches such as BERT and RoBERTa for text categorization. It works around the limitations of manual and traditional machine-learning algorithms and exploits the benefits of automated, deep-learning-based techniques. The paper enriches the understanding and context of where machine learning algorithms lack and encourages us to think about how we can improve them compared to deep learning techniques.

The paper [9] explores using Random Forest and Naïve Bayes algorithms for news text categorization. It highlights the importance of feature extraction and selection and introduces the use of performance metrics such as accuracy and kappa statistics. The paper's findings imply that Random Forest is a more effective algorithm for news text categorization than Naïve Bayes. However, Naïve Bayes is a simpler algorithm and may be a better choice for tasks where computational resources are limited or labelled data is scarce. The paper aligns with our research, addressing news text categorization and feature selection.

In the paper [10], M. Baygin proposes a novel approach for classifying Turkish documents into five distinct categories using the Naive Bayes algorithm. This machine learning algorithm is known for its simplicity and effectiveness, enabling the proposed method to achieve an impressive success rate of 92% on a dataset of Turkish news articles. This approach is well-suited for tasks where large volumes of textual data must be classified efficiently and automatically.

The paper [11] investigates the influence of data preprocessing on the performance of the Naïve Bayes Classifier, particularly in the context of spam email identification. It highlights the significance of data preprocessing in handling unstructured text data and its impact on the accuracy of the classifier. By comparing the classifier's performance on preprocessed and non-preprocessed datasets, the study finds that proper data preprocessing enhances prediction accuracy, underlining the importance of this step in optimizing the Naïve Bayes Classifier's effectiveness in document classification tasks such as spam email identification. This provides valuable insights for improving classification processes.

The paper [12] by Shasha Wang, Liangxiao Jiang, and Chaoqun Li proposes a new algorithm called multinomial naive Bayes tree (MNBTree) for text classification. MNBTree is a hybrid algorithm that combines the advantages of decision trees and naive Bayes classifiers. It builds a binary decision tree with a multinomial naive Bayes classifier at each leaf node. The

multiclass version of MNBTree (MMNBTree) works by training a separate MNBTree classifier for each class. To classify a new text document, MMNBTree uses the MNBTree classifier for the class that predicts the highest probability. This approach has been shown to outperform other popular text classification algorithms on several standard datasets.

In the paper [13], a critical aspect is the enhancement of classification accuracy using the Naïve Bayes classifier. This study explores three approaches: the original Naïve Bayes, Naïve Bayes with Chi-Square feature selection, and Naïve Bayes with Information Gain. The Naïve Bayes classifier, despite its simplistic conditional independence assumption, has been a popular choice for text classification tasks. Feature selection methods such as Chi-Square and Information Gain(IG), are employed to identify and retain the most important features for sentiment classification, contributing to the overall objective of improving accuracy in sentiment analysis models.

In the paper [14], Deep Learning based approaches are implemented to classify Bangla text documents. Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) is used here for the classification task. Here we have implemented an advanced technique that encoded the documents at their character level. Documents from three different data sources are used to validate and test of the working models. The highest classification accuracy is 95.42% that is achieved on the Prothom Alo data set using LSTM. We presented a comparison between two models and explained how well the classification task can be carried out using our character level approach with higher accuracy.

The paper [15] presents a new hybrid two-layer feature selection approach that combines a wrapper and an embedded method in constructing an appropriate subset of predictors. In the first layer of the proposed method, the Genetic Algorithm(GA) has been adopted as a wrapper to search for the optimal subset of predictors, which aims to reduce the number of predictors and the prediction error. As one of the meta-heuristic approaches, GA is selected due to its computational efficiency; however, GAs do not guarantee the optimality. To address this issue, a second layer is added to the proposed method to eliminate any remaining redundant/irrelevant predictors to improve the prediction accuracy.

The research [16] focuses on news categorization, addressing intra-class classification within the sports category and inter-class classification across technology, business, sports, politics, and entertainment. Using the BBC news dataset, the study employs a multi-feature approach, involving pre-processing, and TF-IDF extraction of unigram, bigram, and trigram word tokens. This method is deemed valuable for tackling the complexity of news classification, considering the diversity and multi-class nature of news articles. In contrast to previous studies, this research offers a comprehensive and multi-feature perspective on news categorization, potentially enhancing the accuracy and informativeness of news classification systems.

In the paper [17], a two-stage feature selection and extraction approach is employed to enhance text categorization performance. In the initial stage, each term within the document is ranked based on its importance for classification using the information gain (IG) method. Subsequently, genetic algorithm (GA) and principal component analysis (PCA) feature selection and extraction methods are separately applied to the top-ranked terms, facilitating dimension reduction during text categorization. This process ignores less important terms, thereby reducing computational time and complexity. Experiments assessing the effectiveness of dimension reduction methods are conducted using the k-nearest neighbor (KNN) and C4.5 decision algorithms.

This study [18] focuses on text feature selection by introducing improved genetic algorithm (GA), addressing the limitations of traditional adaptive GAs. Feature selection is important as it aids in retaining informative features while reducing data dimensionality. By refining the fitness function, crossover probability, and mutation probability formulas, this study contributes to the ongoing efforts to enhance feature selection within text data. This optimization method is used for improving the accuracy and efficiency of text-based applications, ranging from sentiment analysis to document classification, ultimately advancing the field of natural language processing and information retrieval.

The paper [19] introduces a novel bi-objective genetic algorithm for feature selection in data mining. It combines rough set theory and multivariate mutual information to select precise and informative data while eliminating vagueness. Through parallel processing, an ensemble of feature selectors is generated, resulting in a more generalized feature subset. Validation on diverse datasets demonstrates its effectiveness in achieving improved classification accuracy and statistical measures compared to existing methods, offering a promising solution to the feature selection problem in data mining.

The article by Kruschke and Liddell [20] advocates for Bayesian statistical methods over traditional frequentist approaches. It critiques null hypothesis significance testing (NHST) limitations and highlights Bayesian benefits in hypothesis testing, estimation, meta-analysis, and power analysis. Emphasizing coherent frameworks and credible intervals, the authors encourage a transition to Bayesian methods for enhanced statistical rigor and understanding in scientific practices.

The paper [21] develops an intrusion detection system focusing on high accuracy and low false alarms. Employing an ensemble approach with classifiers like

SVM, modified Naive Bayes, and LPBoost, along with Chi-square feature selection, the model aims to enhance intrusion detection effectiveness while minimizing complexity. Experimental results show that the LPBoost ensemble outperforms individual classifiers, demonstrating high accuracy in identifying intrusions. The study underscores the significance of critical feature selection for an effective intrusion detection model and advocates for the benefits of a supervised learning approach in this context.

The paper by Abbas et al. [22] introduces a sentiment analysis model using Multinomial Naive Bayes. It employs preprocessing techniques such as stopword removal and stemming, representing text features through a bag-of-words approach. The model's effectiveness is demonstrated through performance metrics on diverse datasets. The study emphasizes the broad applications of sentiment analysis, including social media analysis, customer reviews, and market research, showcasing its versatility across domains.

Lei's (2012) paper [23] introduces a hybrid feature selection method, combining information gain and genetic algorithms in machine learning and data mining. The approach calculates information gain for feature relevance to class labels and employs genetic algorithms iteratively to optimize feature selection. Experimental results show its effectiveness in identifying informative features and improving classification performance. This hybrid approach balances the strengths of both methods, providing an efficient means to enhance classification accuracy and reduce computational complexity in various applications.

The study [24] introduces a novel multilabel neural network approach for functional genomics and text categorization. Using a multilayer perceptron (MLP) architecture with a sigmoidal activation function, the authors address the complexity of simultaneous assignment of multiple labels. They propose a feature selection technique based on genetic algorithms and ReliefF to handle high-dimensional genomics data, enhancing efficiency and performance. Experimental results on real-world datasets demonstrate the approach's superiority in accuracy and efficiency over existing methods, indicating its potential for solving complex multilabel classification challenges in functional genomics and text categorization.

C. Genetic Algorithm (GA)

A Genetic Algorithm (GA) is a search and optimization algorithm inspired by the process of natural selection and genetics. It is used to find approximate solutions to optimization and search problems. In a GA, a population of potential solutions evolves over generations, with individuals undergoing processes such as selection, crossover (recombination), and mutation. The fitness of individuals guides their likelihood of being selected for reproduction, allowing the algorithm to explore and exploit the solution space efficiently.

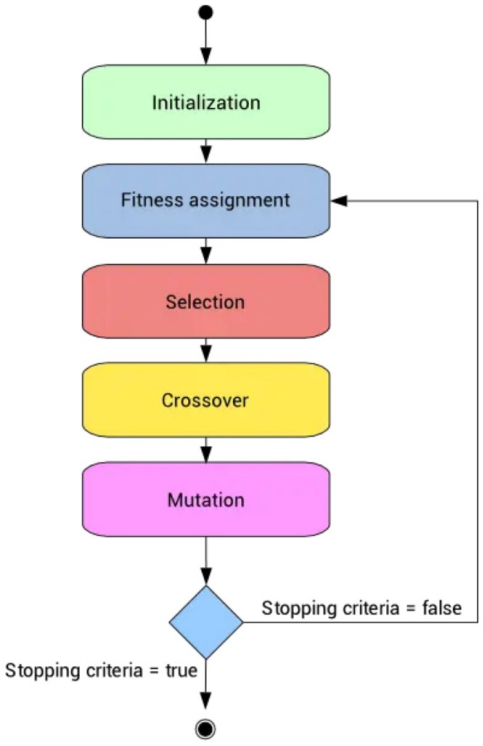

Fig. 1. Steps in Genetic Algorithm

GA can be broken down into eight simple steps:

- **Initialization**: Generate an initial population of potential solutions (individuals or chromosomes). Each individual represents a possible solution to the optimization problem.
- **Fitness Evaluation**: Evaluate the fitness of each individual in the population. The fitness function quantifies how well each solution solves the optimization problem. It is problem-specific and guides the algorithm toward better solutions.
- **Selection**:Select individuals from the population to serve as parents for the next generation. The probability of selection is often proportional to the fitness of the individuals; fitter individuals are more likely to be selected.
- **Crossover**: Perform crossover on pairs of parents. This involves exchanging genetic information between two parents to create one or more offspring. Crossover helps combine beneficial traits from different parents.
- **Mutation**: Introduce random changes in the offspring to maintain diversity in the population. Mutation helps explore the solution space and prevents premature convergence to a suboptimal solution.
- **Replacement**: Create a new population by combining parents and offspring. This can involve strategies like elitism, where the best individuals from the current generation are directly copied to

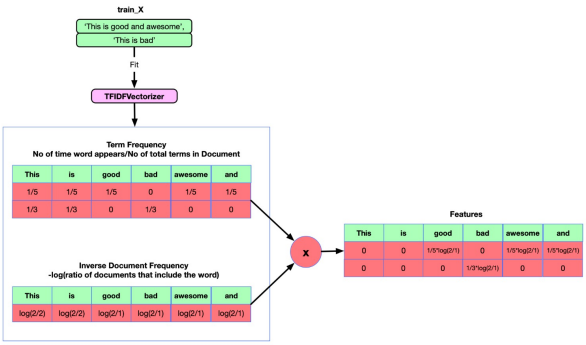

Fig. 2. TF-IDF

the next generation.

- **Termination Criteria**: Check if a termination condition is met. Common termination conditions include reaching a maximum number of generations, finding a satisfactory solution, or running the algorithm for a specified amount of time.
- **Repeat**: If the termination condition is not met, repeat steps 2-7 for the next generation.

D. Feature extraction using TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) calculates the importance of words within each document while considering their rarity across the entire dataset. The scikit-learn library was utilized for implementing the TF-IDF vectorization. The resulting feature matrix represents each document as a vector in a high-dimensional space, where each dimension corresponds to a unique term in the corpus. This transformed representation captures the distinctive features of each document, emphasizing terms that are indicative of the document's category. The TF-IDF vectors were then used as input features for training the text classification model, facilitating effective discrimination between sports and sci-tech news articles.

$$\text{TF}(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d} \tag{1}$$

The Inverse Document Frequency (IDF) is calculated using the formula:

$$\text{IDF}(t) = \log\left(\frac{N}{\text{df}(t)}\right) \tag{2}$$

where:

$\text{IDF}(t)$ is the IDF of term t,

N is the total number of documents in the corpus,

$\text{df}(t)$ is the number of documents in the corpus that contain the term t.

E. Feature selection using Chi-square

To improve upon the efficiency and interpretability of the text classification model feature selection was performed using the Chi-square test. This statistical method evaluates the independence between individual features and the binary target variable representing sports and sci-tech news articles. Features with significant chi-square statistics were retained, indicating their relevance in discriminating between the two classes. By reducing the dimensionality of the feature space, this process not only enhances the computational efficiency of the model but also aids in identifying the most discriminative terms for classification. The selected features were then used as input for training the binary text classification model.

The chi-square statistic is calculated using the formula:

$$\chi^2 = \sum \frac{(\text{Observed Frequency} - \text{Expected Frequency})^2}{\text{Expected Frequency}} \tag{3}$$

III. Methodology

The methodology for probabilistic genetic optimization for Text Categorization consists of the following steps:

A. Data Preprocessing

1) Removal of Punctuation and Numbers: In the initial stage of text preprocessing, the removal of special characters is crucial for enhancing the cleanliness and uniformity of textual data. Special characters such as punctuation marks, symbols, and other non-alphanumeric elements may introduce noise and hinder subsequent analysis. To address this, a regular expression-based approach is employed. Utilizing the `re` module in Python, the `remove_special_characters` function identifies and replaces any character not belonging to the set of letters (A-Z or a-z), digits (0-9), or whitespace (\s) with an empty string. This process ensures that the resulting text is devoid of unnecessary symbols, facilitating more accurate and meaningful analysis.

2) Removal of Stopwords and Shortwords: Stopwords, commonly occurring words in a language that contribute little semantic meaning, are often removed during text preprocessing to focus on more significant terms. In this phase, NLTK's predefined list of English stopwords is leveraged. The `remove_stopwords` function utilizes this list to filter out stopwords from the previously tokenized text. By eliminating these common words, the analysis can better concentrate on the content-rich terms, enhancing the accuracy and relevance of subsequent natural language processing tasks. Also words with length ¡ 3 are removed from the text since they're mostly stopwords or noise data.

3) Tokenization: Tokenization is a fundamental step in text preprocessing that involves breaking down a continuous text into individual units, known as tokens. These tokens serve as the basic building blocks for subsequent analysis, enabling a more granular understanding of the text. In this context, the Natural Language Toolkit's (`NLTK`) `word_tokenize` function

is applied. This function splits the text into words, providing a structured representation of the input text.

4) Lemmatization: Lemmatization is the process of reducing words to their base or root form, providing a standardized representation for variations of a word. In the context of text preprocessing, lemmatization aids in harmonizing different inflections of words, improving the overall consistency of the textual data. Employing NLTK's `WordNetLemmatizer`, the `lemmatize_tokens` function iterates through the tokenized text, transforming each word into its base form. This step contributes to a more nuanced understanding of the text by reducing words to their essential meanings and simplifying subsequent analyses, such as sentiment analysis or topic modeling.

These four sequential steps collectively form a comprehensive text preprocessing pipeline, transforming raw textual data into a refined and standardized format conducive to effective natural language processing and analysis.

B. Feature Selection

A filter-based Genetic Algorithm (GA) approach has been employed for feature selection to select the 100 best features. To assess our model outcomes, we have also implemented feature selection using chi-square for comparison.

1) Feature Selection using GA: The aim of feature selection is to select the best features which differentiate the two classes - sports (class 0) and sci-tech (class 1). Hence, a chromosome or an individual which represents the potential solution or the potential set of best features consists of two parts (say C1 and C2). C1 consists of the best set of features for class 0, while C2 consists of the best set of features for class 1. The feature selection is performed directly on the preprocessed text data.

The GA starts by randomly initialising a population of the given population size. Then, it evolves the population according to the number of generations set and returns the best individual found.

For each generation, the fitness of each individual in the population is calculated. The fitness is calculated as follows:

first_label_words represent the words associated with class 0 and the second_label_words represent the words associated with class 1. The fitness variable is initialized to zero and serves as a score accumulator. The algorithm iterates through the words in the sublists of individual and adjusts the fitness accordingly. For each word in C1, if it belongs to class 0, the fitness function is incremented by the words count in C1, else it is decremented by the same. Similarly, for each word in C2, if it belongs to class 1, the fitness function is incremented by the words count in C2, else it is decremented by the same.

Once the fitness of all the individuals in the population of the current generation is calculated, the current generation is evolved to produce the next generation. The evolution consists of the following steps:

```
Function find_fitness(individual, first_label_words,
second_label_words):
    fitness ← 0;
    for word in individual[0] do
        if word in first_label_words then
            fitness += first_label_words[word];
        end
        if word in second_label_words then
            fitness −=
              second_label_words[word];
        end
    end
    for word in individual[1] do
        if word in second_label_words then
            fitness +=
              second_label_words[word];
        end
        if word in first_label_words then
            fitness −= first_label_words[word];
        end
    end
    return fitness;
```

a) Finding elites: This step includes finding the most fittest individuals from the current generation and adding it to the next generation. 'elite_size' number of individuals are selected. For our implementation, we have set the 'elite_size' to 20, so from a population of 1000 individuals, the top 20 fittest individuals are selected to survive in the next generation.

b) Crossover: A single point crossover is performed by selecting random parents from the elite individuals list in the next generation. The crossover function starts by creating a child with two sets of traits, similar to the parents. Then, for each position or trait in the parents, it randomly decides whether to take that trait from parent1 or parent2, making this decision based on a probability of 50/50 (roughly like flipping a coin). If the random value is greater than 0.5, it picks the trait from parent1 for the first set of traits in the child and from parent2 for the second set. If the random value is less than or equal to 0.5, it does the opposite. Finally, it returns the newly created child, which is a blend of traits from both parents.

c) Mutation: This step performs mutation in the individuals so as to introduce diversity in the population. For each sublist (C1 and C2) within the individual, the function checks whether a mutation should occur. It does this by generating a random number for each sublist. If this random number is less than the mutation_rate, a mutation is triggered. When a mutation happens, it replaces the existing trait at that position with a randomly chosen trait from the vocab list. This process happens independently for

both sublists. Finally, it returns the mutated individual with potentially altered traits in either or both of its sublists. The mutated individual is then added to next generation, provided it already does not exist in it.

The above steps are performed until the number of individuals in the next generation is equal to the population size.

Throughout the genetic algorithm (GA) process, the algorithm maintains a record of the top-performing individual, which is the one with the highest fitness score. At the end of the GA process, this best individual is retrieved and returned as the output of the algorithm.

2) Feature Selection using chi-square: TF-IDF vectorizer is used to convert text data into numerical representations and then SelectKBest with chi2 is employed to select the most relevant features.

C. Training and evaluating models

Three distinct models—Multinomial Naive Bayes, Gaussian Naive Bayes, and Support Vector Machine (SVM)—were trained utilizing the selected features from each of the above approaches. Each set of models were evaluated and the results were compared.

IV. Results

Word cloud of sports related content as shown in Fig. 3 reveals a dynamic mix of terms, including specific sports names and action verbs. This concise representation captures the sports terms.

Word cloud of sci-tech related content as shown in Fig. 4 reveals a dynamic mix of terms, including specific sci-tech names and scientific verbs. This concise representation captures the scientific and technological terms.

Fig. 3. Words related to sports as obtained using wordcloud

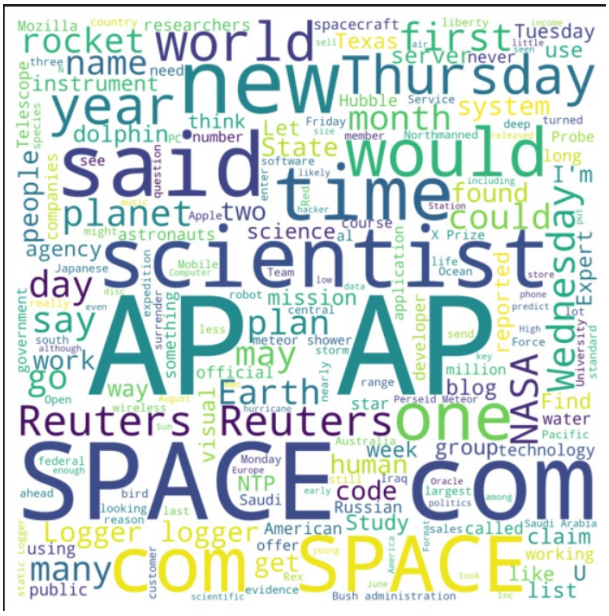

Fig. 4. Words related to sci-tech as obtained using wordcloud

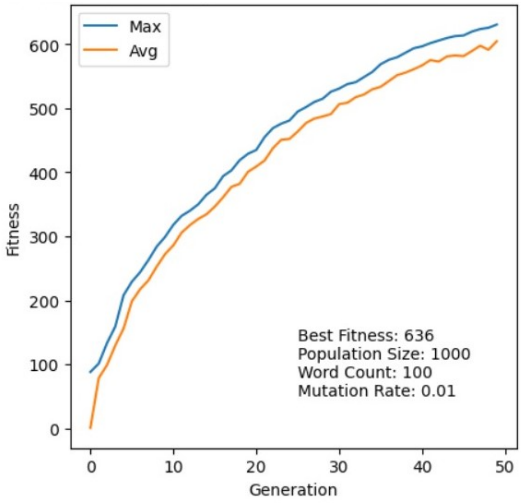

Fig. 5. Graph between fitness of an individual and the generation

A fitness-generation graph as depicted in Fig. 6 tracks how the fitness of individuals evolves over generations in an optimization algorithm. The x-axis represents generations, and the y-axis represents fitness scores. Trends in the graph indicate whether the algorithm is converging towards better solutions over time. This visual provides a quick assessment of the algorithm's effectiveness and convergence patterns. The results (Accuracy, Precision, Recall and F1-Score) are obtained after using Chi-Square for feature selection as shown in Fig. 7. The results show an improvement when using GA for feature selection as seen in Fig. 8.

Model	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	0.850	0.865857	0.850	0.849248
Gaussian Naive Bayes	0.775	0.795833	0.775	0.772879
SVM	0.850	0.865857	0.850	0.849248

Fig. 6. Results obtained with feature selection by chi-square

Model	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	0.925	0.935227	0.925	0.924859
Gaussian Naive Bayes	0.875	0.899038	0.875	0.872179
SVM	0.925	0.935227	0.925	0.924859

Fig. 7. Results obtained with feature selection by GA

V. Discussion

A. GA as a filter method

Employing GA as a filter method, rather than a wrapper method, decreases computational complexity as it avoids the need for repetitive model training. In our GA approach, dependence on model-specific metrics is eliminated, replaced by an evaluation through a Bayesian framework-based fitness function. This independence makes it adaptable to various models and less constrained by model-specific requirements.

B. GA before text-extraction

When applying GA directly on text data, you retain the richness and complexity of the original text. Each word and its arrangement contribute to the genetic makeup of individuals in the population. This approach preserves valuable information, capturing the intricacies and nuances of language that might be lost in feature extraction.

GA applied on raw text data allows the algorithm to freely evolve and explore combinations of words and phrases without predefined features.

GA applied on raw text allows the algorithm to evolve and discover domain-specific terms or lexicons that are crucial for the classification task. This can be especially valuable in domains with evolving language or specific jargon.

C. Probablistic and Bayesian framework based Fitness Function

The fitness function evaluates the occurrence of words within an individual in relation to two sets of labels. This approach to evaluating an individual's fitness based on word occurrences within predefined label sets aligns conceptually with the principles of conditional probability and Bayesian inference. It utilizes prior knowledge (associated words with labels) and observed data (word occurrences in the individual) to quantify the likelihood of the individual's alignment with these labels, reflecting a Bayesian approach to updating beliefs through evidence. The function's foundation in conditional probability enables a fine-grained assessment of word relevance within class contexts, pinpointing words that uniquely define particular classes. By integrating a Bayesian framework, it achieves adaptability. This approach ensures an information-rich evaluation that captures the intricate relationships between words and class distinctions.

D. Model results

Multinomial Naive Bayes and Support Vector Machines (SVM) outperform Gaussian Naive Bayes in text classification due to their well-suited characteristics for high-dimensional and sparse feature spaces inherent in natural language processing. Multinomial Naive Bayes leverages the frequency of word occurrences, assuming independence between features, making it particularly effective for document categorization. SVM, on the other hand, excels at finding optimal hyperplane boundaries in feature spaces, providing robust classification even in the presence of noise and irrelevant features. Unlike Gaussian Naive Bayes, these models are better adapted to the discrete and sparse nature of text data. SVM's ability to handle non-linear relationships further enhances its performance. Overall, Multinomial Naive Bayes and SVM offer more suitable frameworks for capturing the intricacies of text data, leading to superior text classification results compared to Gaussian Naive Bayes.

VI. Conclusion

Our project presents a comprehensive exploration of feature selection methodologies in the context of text categorization, with a specific focus on the sports and sci-tech domains. The utilization of a filter-based Genetic Algorithm (GA) approach, coupled with chi-square feature selection, demonstrated its efficacy in enhancing the categorization process. Comparative analyses between GA and chi-square feature selection revealed a substantial performance boost when employing GA, underscoring the significance of our proposed methodology. Guided by a Bayesian framework-based fitness function, this approach showcased adaptability across diverse models, mitigating the constraints associated with model-specific metrics. The application of GA directly on raw text data before extraction allowed for the preservation of linguistic nuances and the exploration of word combinations, providing a unique perspective on feature selection. The probabilistic and Bayesian framework-based fitness function emerged as a robust evaluation tool, contributing to the versatility of our feature selection approach. Model results indicated that Multinomial Naive Bayes and Support Vector Machines outperformed Gaussian Naive Bayes, aligning with their suitability for the intricacies of natural language processing tasks.

VII. Future Work

We aim to extend the application of the proposed Genetic Algorithm-based feature selection technique to multiclass data, broadening its utility in handling multiple categories. Additionally, we plan to explore

and experiment with more sophisticated fitness functions to further refine the feature selection process. An in-depth analysis of the impact of Genetic Algorithm parameters, including population size, mutation rate, and crossover rate, is crucial for optimizing algorithm performance. We also intend to investigate hybridizing the proposed technique with other established feature selection methods for a more robust solution. Real-world application and evaluation across diverse domains and scalability enhancements for large-scale datasets are vital aspects of our future research agenda.

REFERENCES

[1] A. I. Kadhim and A. A. Abdalhameed, "A hybrid feature selection technique using chi-square with genetic algorithm," 2022, pp. 212-217, doi: 10.1109/MICEST54286.2022.9790277

[2] Kadhim, Ammar. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review. 52. 10.1007/s10462-018-09677-1.

[3] G. Singh, B. Kumar, L. Gaur and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, UK, 2019, pp. 593-596, doi: 10.1109/ICACTM.2019.8776800.

[4] Jan Klusáček, Václav Jirsík, Comparing Fitness Functions for Genetic Feature Transformation, IFAC-PapersOnLine, Volume 49, Issue 25, 2016, Pages 299-304, ISSN 2405-8963, https://doi.org/10.1016/j.ifacol.2016.12.053.

[5] Y. Zhai, W. Song, X. Liu, L. Liu and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2018, pp. 160-163, doi: 10.1109/ICSESS.2018.8663882.

[6] G. Singh, B. Kumar, L. Gaur and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, UK, 2019, pp. 593-596, doi: 10.1109/ICACTM.2019.8776800.

[7] I. Dawar and N. Kumar, "Text Categorization By Content using Naïve Bayes Approach," 2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON), Jaipur, India, 2023, pp. 1-6, doi: 10.1109/IEMECON56962.2023.10092372.

[8] S. Rehman, A. Irtaza, M. Nawaz and H. Kibriya, "Text Document Classification Using Deep Learning Techniques," 2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Lahore, Pakistan, 2022, pp. 1-6, doi: 10.1109/ETECTE55893.2022.10007316.

[9] U. Parida, M. Nayak and A. K. Nayak, "News Text Categorization using Random Forest and Naïve Bayes," 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON), Bhubaneswar, India, 2021, pp. 1-4, doi: 10.1109/ODICON50556.2021.9428925.

[10] M. BAYGIN, "Classification of Text Documents based on Naive Bayes using N-Gram Features," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-5, doi: 10.1109/IDAP.2018.8620853.

[11] P. Chandrasekar and K. Qian, "The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier," 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC), Atlanta, GA, USA, 2016, pp. 618-619, doi: 10.1109/COMPSAC.2016.205.

[12] Wang, S., Jiang, L. & Li, C. Adapting naive Bayes tree for text classification. Knowl Inf Syst 44, 77–89 (2015). https://doi.org/10.1007/s10115-014-0746-y

[13] D. Kurniawan, M. Yasir and F. C. Venna, "Optimization of Sentiment Analysis using Naive Bayes with Features Selection Chi-Square and Information Gain for Accuracy Improvement," 2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Jakarta, Indonesia, 2022, pp. 153-160, doi: 10.23919/EECSI56542.2022.9946510.

[14] M. M. Rahman, R. Sadik and A. A. Biswas, "Bangla Document Classification using Character Level Deep Learning," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 2020, pp. 1-6, doi: 10.1109/ISMSIT50672.2020.9254416.

[15] Amini, F., & Hu, G. (2021). A two-layer feature selection method using genetic algorithm and elastic net. Expert Systems with Applications, 166, 114072.

[16] Singh, Dimpleveer & Malhotra, Sumit. (2018). Intra News Category Classification using N-gram TF- IDF Features and Decision Tree Classifier.

[17] A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, Knowledge-Based Systems, Volume 24, Issue 7, 2011, Pages 1024-1032, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2011.04.014.

[18] Su, Shuangquan & li, Lei & Zhao, Qing. (2011). Text feature selection based on improved adaptive GA. 169-172. 10.1109/NLPKE.2011.6138188.

[19] Das, A.K., Pati, S.K. & Ghosh, A. Relevant feature selection and ensemble classifier design using bi-objective genetic algorithm. Knowl Inf Syst 62, 423–455 (2020).

[20] Kruschke, J.K., Liddell, T.M. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev 25, 178–206 (2018). https://doi.org/10.3758/s13423-016-1221-4

[21] Thaseen, I.S., Kumar, C.A. & Ahmad, A. Integrated Intrusion Detection Model Using Chi-Square Feature Selection and Ensemble of Classifiers. Arab J Sci Eng 44, 3357–3368 (2019). https://doi.org/10.1007/s13369-018-3507-5

[22] Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), 62.

[23] Lei, S. (2012, March). A feature selection method based on information gain and genetic algorithm. In 2012 international conference on computer science and electronics engineering (Vol. 2, pp. 355-358). IEEE.

[24] Min-Ling Zhang and Zhi-Hua Zhou, "Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization," in IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, pp. 1338-1351, Oct. 2006, doi: 10.1109/TKDE.2006.162.