information

*Article*

# Reimagining Literary Analysis: Utilizing Artificial Intelligence to Classify Modernist French Poetry

**Liu Yang [1,*], Gang Wang [2] and Hongjun Wang [2]**

[1] School of Foreign Studies, Zhongnan University of Economics and Law, Wuhan 430073, China
[2] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China; 2022wg@my.swjtu.edu.cn (G.W.); wanghongjun@swjtu.edu.cn (H.W.)
* Correspondence: yarres@zuel.edu.cn

**Abstract:** Aligned with global Sustainable Development Goals (SDGs) and multidisciplinary approaches integrating AI with sustainability, this research introduces an innovative AI framework for analyzing Modern French Poetry. It applies feature extraction techniques (TF-IDF and Doc2Vec) and machine learning algorithms (especially SVM) to create a model that objectively classifies poems by their stylistic and thematic attributes, transcending traditional subjective analyses. This work demonstrates AI's potential in literary analysis and cultural exchange, highlighting the model's capacity to facilitate cross-cultural understanding and enhance poetry education. The efficiency of the AI model, compared to traditional methods, shows promise in optimizing resources and reducing the environmental impact of education. Future research will refine the model's technical aspects, ensuring effectiveness, equity, and personalization in education. Expanding the model's scope to various poetic styles and genres will enhance its accuracy and generalizability. Additionally, efforts will focus on an equitable AI tool implementation for quality education access. This research offers insights into AI's role in advancing poetry education and contributing to sustainability goals. By overcoming the outlined limitations and integrating the model into educational platforms, it sets a path for impactful developments in computational poetry and educational technology.

**Keywords:** AI in literary analysis; machine learning; modern french poetry classification; feature extraction techniques; SVM in poetry analysis



**Citation:** Yang, L.; Wang, G.; Wang, H. Reimagining Literary Analysis: Utilizing Artificial Intelligence to Classify Modernist French Poetry. *Information* **2024**, *15*, 70. https://doi.org/10.3390/info15020070

Academic Editor: Katsuhide Fujita

Received: 30 November 2023
Revised: 15 January 2024
Accepted: 15 January 2024
Published: 24 January 2024



## 1. Introduction

In digital humanities, AI's role is transformative, especially in analyzing complex literary genres like Modernist French poetry [1]. This study uses AI to classify Modernist French poetry, aligning with the United Nations Sustainable Development Goals, particularly in education and reducing inequalities [2]. We aim to bridge cultural and linguistic divides through AI-driven analysis, opening new pathways for appreciating diverse literary traditions [3].

Modernist French poetry, with its audacious departure from traditional forms, offers a unique challenge for AI with its symbolism, unconventional syntax, and thematic depth [4]. Its distinct characteristics make it well-suited for AI analysis, providing a rich dataset for machine learning algorithms [5]. Despite its richness, the AI analysis of this genre is still emerging, with a focus on traditional forms, leaving a gap in Modernist French poetry exploration [4].

This study develops an AI framework for classifying Modernist French poetry, enhancing digital humanities and understanding this literary movement. We apply AI to delve into linguistic patterns, stylistic innovations, and thematic elements of the genre [6]. The defining features of Modernist French poetry—fragmentation, symbolism, subconscious exploration, and reflections of urbanization—present a compelling AI analysis subject [7].

An AI-driven analysis can preserve and disseminate Modernist French poetry, making it accessible and fostering global appreciation [8]. It also promises educational applications,

enhancing critical thinking and diverse perspectives in literary studies [9]. AI serves as a complementary tool in literary criticism, enriching traditional methods with new insights [10].

Our research, focusing on modern French poetry, blends AI with poetic expression and establishes new benchmarks in literary research. We prioritize a narrative style over complex mathematics, aiming for accessibility in an interdisciplinary context. This approach highlights AI's role in augmenting literary analysis and comprehension [11].

This research marks an interdisciplinary breakthrough by applying AI to analyze and classify modern French poetry, integrating AI with a literary analysis. We highlight three major contributions:

- **Innovative AI in Literary Research:** This study introduces AI, particularly SVM and advanced feature extraction like TF-IDF and Doc2Vec, into modern French poetry analysis. This approach not only brings new analytical tools to literary research but also enriches education and cross-cultural understanding.
- **Enhancing Literary Comprehension:** Our work advances literary understanding through AI, dissecting stylistic elements of modern French poetry. This enriches educational methodologies and opens new avenues for cross-cultural studies, deepening the appreciation of literary diversity.
- **Technical Enrichment:** We focus on efficient SVM deployment and diverse feature extraction methods, ensuring data integrity for machine learning classification. Our research employs various classifiers, enhancing the accuracy and applicability of AI in a literary analysis.

In summary, our study extends existing AI methodologies to literary analysis, particularly focusing on modern French poetry. We blend AI analysis with a deep understanding of poetic expression, setting new benchmarks for literary AI application. The study avoids complex mathematics for accessibility and clarity, emphasizing AI's role in augmenting literary comprehension and analysis.

## 2. Related Works

Research in computer-generated poetry and literature analysis is extensive [12]. Key advancements include the Quality-Aware Masked Language Model for high-quality quatrain generation and cross-cultural poetry generation research [13]. Educational applications in poetry writing and computational tools for analysis have also been explored [14,15]. Notably, AI's role in Arabic poetry classification and analysis is emerging [16].

### *2.1. Challenges of Text Classification in Poetry*

Text classification in poetry faces challenges like ambiguity, subjectivity, and varied styles [17–19]. Additional issues include feature reduction, sentiment analysis, dataset domain and sources [20], bilingual text recognition challenges [21], and benchmarking difficulties [22,23]. Other limitations are semantic label information neglect, modeling restrictions in image-text concatenation [24], and cross-domain text classification issues [25]. Challenges in short text classification [26], incremental few-shot learning [27], human-designed feature biases [28], and the inadequacy of "bag of words" models [29] are also significant.

In summary, text classification in poetry requires a multifaceted approach integrating machine learning, NLP, and domain-specific knowledge to address its diverse challenges.

### *2.2. Categorizing Existing Works Based on Approaches and Contributions*

Poetry classification research divides into rule-based and data-driven methodologies. Rule-based methods, utilizing predefined rules often grounded in linguistic analysis, are praised for their interpretability and domain expertise integration. Abel and Lantow's 2019 framework [30] and Wang and Hong's study on rule-based feature selection [31] illustrate these methods' structured approach. Additionally, associative the rule-based classification [32] and association rule mining [33] demonstrate rule-based methods' diversity and applicability.

Conversely, data-driven approaches employ machine learning algorithms to autonomously learn from text data that are adaptable for handling large datasets. Liu et al.'s 2020 work [34] and Fragoso's 2019 evaluation [35] showcase these methods' flexibility. The use of hierarchical models for offensive content classification [36] and genetic programming [34] exemplifies advancements in data-driven techniques. Additionally, privacy concerns in data-driven classification [37] highlight the multifaceted nature of these approaches.

These references collectively provide an overview of poetry classification research, outlining the varied methodologies and contributions of both rule-based and data-driven approaches to the field.

### *2.3. Critically Evaluating Strengths and Weaknesses of Each Work*

The evaluation of text classification methods, particularly in poetry, reveals distinct strengths and weaknesses. Rule-based methods, as shown in Abel and Lantow's study [30], offer interpretability and expert knowledge integration but may lack flexibility and struggle with poetic ambiguity. Data-driven approaches excel in adaptability and handling large datasets, but their reliance on extensive training and sometimes opaque processes can be limitations. Their ability to capture complex patterns, as discussed in recent studies [38,39], is countered by concerns over interpretability and overfitting risks. Hybrid models, combining rule-based and data-driven methods, enhance effectiveness but add complexity, as highlighted by Wang et al. [40]. Emerging techniques, such as graph neural networks and weakly supervised learning [41,42], offer innovative solutions but face challenges like data requirements and computational demands. In summary, this section critically assesses various methodologies in text classifications within poetry, underscoring the balance between strengths and limitations and the need for innovative approaches to address the unique complexities of poetic texts.

### *2.4. Beyond Boundaries: Recent AI Advancements Illuminate Modernist French Poetry Analysis*

This comprehensive review of text classification literature, especially in poetry, synthesizes key advancements and future directions. Recent studies [43] emphasize advancements in data processing, showcasing the evolution of text classification methods. The integration of NLP techniques [44] highlights the effectiveness of combined methodologies in enhancing model accuracy and applicability. The applications of text classification across various domains, including news and job market analysis [45,46], demonstrate its versatility. Challenges such as decision jumps, data imbalance, and algorithm complexity are outlined in recent research [47], underscoring the importance of feature selection and model establishment in developing accurate classification models. Our review extends beyond traditional boundaries, incorporating AI advancements to enrich the Modernist French poetry analysis. For example, the PF-BiGRU-TSAM method [48] adeptly handles complex data, an approach adaptable to poetic language analysis. Challenges in poetry classification, like ambiguity and complex structures, are informed by studies on feature reduction and sentiment analysis [20–22]. Recent developments, such as contrastive learning and adversarial training [40], offer novel methods to tackle classification challenges relevant to Modernist French poetry. By integrating these advanced AI techniques, we aim to develop a more nuanced model for the poetry analysis.

In conclusion, the text classification in poetry is evolving towards more sophisticated, integrated approaches that continuously adapt to address challenges and explore new analytical possibilities.

### *2.5. Positioning the Present Work within the Broader Context*

In the evolving landscape of text classification, particularly within the nuanced domain of poetry, our present work aims to carve out a distinct position by building upon and extending the existing body of research. Inspired by Yongqi Li and Wenjie Li [43], our research incorporates advanced data processing techniques to enhance the accuracy and

efficiency of poetry classification, leveraging the latest advancements in data handling. Additionally, drawing on the approach outlined by Amisha Shingala, this work integrates sophisticated NLP techniques to deepen the understanding of poetic texts, aiming to improve the classification accuracy significantly.

Our approach also builds upon the empirical comparisons made by Yindalon et al. in their 2014 study, incorporating the most effective elements from various classification methods to create a robust model specifically for poetry classification. Furthermore, echoing the versatility demonstrated in studies like [45,46], our work adapts text classification techniques to meet the unique challenges presented by poetic texts, acknowledging the diverse applications of these methodologies.

In addition, our research seeks to address and overcome specific challenges of ambiguity and stylistic variation inherent in poetry, informed by insights from the paper [47]. We also recognize the importance of feature selection and model establishment, as discussed in paper [49] and the text emotion classification model patent [50], and incorporate these critical aspects into our work to enhance the classification model's performance.

By situating our current work within this broader context, we not only acknowledge the contributions of previous studies but also strive to advance the field of poetry classification. Our approach combines proven techniques with innovative methods to address the unique challenges of poetry, setting a new benchmark in the realm of text classification. This paper discusses the challenges and limitations of short text classification and literary text classification. Short text classification often faces problems such as data sparsity, noise and ambiguity, and contextual limitations. Literary text classification faces challenges such as subjectivity and ambiguity, contextual complexity, and linguistic diversity. This paper also discusses the progress made in both areas and highlights the need for further research to address the limitations of existing methods and develop more accurate, efficient, and versatile classification algorithms.

## 3. Methodology

This study primarily focuses on employing traditional machine learning techniques for the novel task of style classification in modernist French poetry. Our attention is directed towards several key areas:

**Text Preprocessing and Vectorization:** The text preprocessing phase for classifying modern French poetry is a pivotal process, comprising two critical steps: the identification of stopwords and the tokenization of text. These procedures are fundamental in efficiently extracting pertinent features and ensuring precise text classification. Prior to implementing machine learning techniques, it is imperative to preprocess the dataset of French poetry to confirm its suitability for model training and assessment. Furthermore, in exploring the role of AI in analyzing modernist French poetry, AI's capacity to unveil unique linguistic patterns and stylistic nuances has been invaluable. However, it is crucial to recognize that poetry is not an isolated construct but is shaped by its historical, literary, and cultural milieu. Overlooking these aspects may result in oversimplified and misleading interpretations. For machine learning classification, it is essential to convert textual data into numerical feature vectors that align with the classification models.

**Feature Extraction Methods:** Our research includes an exploration of various feature extraction techniques, notably TF-IDF, to extract salient stylistic features from the text. These features play a pivotal role in understanding and categorizing the styles of poetry. Our study conducts a comparative analysis of two text feature extraction methodologies: Term Frequency-Inverse Document Frequency (TF-IDF) and Document to Vector (Doc2Vec). This comparison underscores the importance of feature extraction as a vital step in transforming textual content into vectors apt for classification models. The transformation of raw textual data into numerical representations that are compatible with machine learning algorithms is pivotal for the efficacious classification of text. This section is dedicated to a comprehensive exploration of the crucial processes involved in text processing and vectorization that is specifically tailored for the classification of poetry.

**Comparison of classification algorithms:** To achieve a precise classification of the styles in modern French poetry, our study will test and evaluate eight different machine learning classification algorithms. Our aim is to identify which algorithm is most adept at handling the complexity and uniqueness of poetic texts. As a central part of our research, we have implemented SVM models for classifying poetry styles, integrating traditional feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF). In order to adapt to the unique features of poetic texts, we have carefully adjusted and optimized the SVM parameters. In addition to SVM, we have investigated several other machine learning classifiers, including Logistic Regression (LR), Bagging, Random Forest (RF), AdaBoost, Gradient Boosted Decision Trees (GBDT), XGBoost, and LightGBM. The parameters of each algorithm have been meticulously calibrated to achieve an optimal classification performance. Through extensive testing of these classification algorithms and evaluating their performance based on accuracy, F1 scores, and other relevant metrics, we will select the algorithm that best classifies modern French poetry. This process involves not only a technical comparison but also an in-depth analysis of each algorithm's ability to handle the specific challenges presented by literary texts. Ultimately, our goal is to determine an optimized model that can accurately reflect the stylistic nuances of poetry and provide new perspectives and tools for literary analysis.

In our research, we consciously refrained from deploying deep learning techniques, opting instead for traditional machine learning algorithms to categorize literary texts. This choice was driven by the desire for simplicity and transparency in our models, factors that greatly facilitate their direct interpretation and understanding. This is particularly pertinent in the cross-disciplinary realm of literary studies. Our approach highlights the innovative integration of machine learning in literary analysis that is particularly evident in our work with modernist French poetry, which is known for its rich expression and stylistic diversity. This methodology is not just a technological breakthrough; it also carves out new pathways for the analysis and interpretation of literary texts. Moreover, our method is deeply anchored in computational linguistics and literary theory. It not only fosters technological advancements but also ensures that our models meet the intricate demands of literary interpretation. We harnessed the strengths of a suite of algorithms, including Support Vector Machine (SVM), Logistic Regression, Bagging, Random Forest, AdaBoost, Gradient Boosted Decision Trees, and XGBoost. This amalgamation is tailored to capture the complex patterns and nuances of poetic language. As a result, our models demonstrate exceptional efficacy in classifying modernist French poetry in Figure 1, a claim that will be substantiated in the detailed evaluations presented in the subsequent chapters. Ultimately, our study paves the way for a new theoretical paradigm in applying machine learning to poetry classification, providing a conceptual foundation that bridges the understanding of poetic language with classification accuracy.

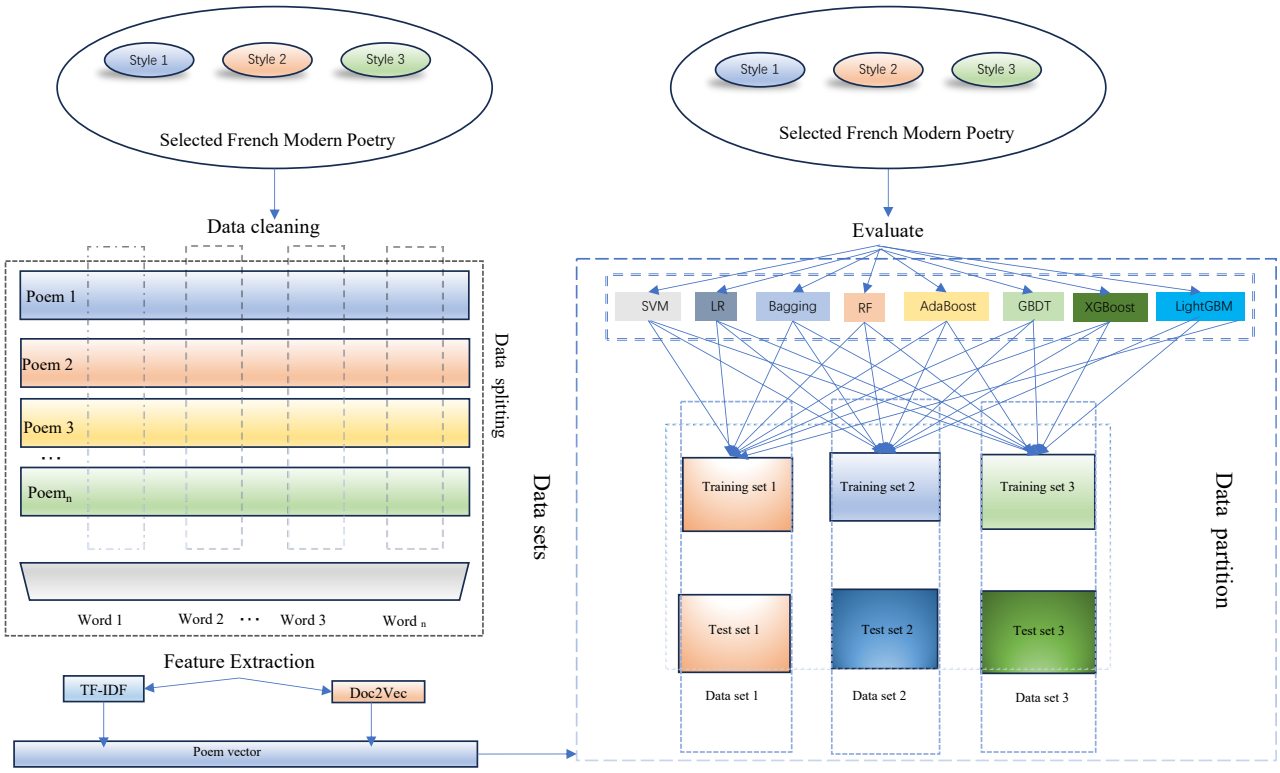


**Figure 1.** Classification Training for Modern French Poetry.

### *3.1. Text Preprocessing for Categorization*

In the realm of Modern French poetry classification, text preparation is a pivotal phase, encompassing two integral steps: the identification of stop words and the process of tokenization. These steps are vital for the effective extraction of features and the accurate classification of text. Prior to employing machine learning techniques, preprocessing the dataset of French poetry is crucial to render it apt for model training and evaluation.

The text preparation process involves two key stages: (1) **Data Cleaning**: As the initial step in the text classification trajectory, this involves the removal of punctuation and the conversion of all letters to lowercase. This simplification is essential to mitigate unnecessary complexities that might arise from variations in punctuation and letter case; (2) **Data Segmentation**: In alignment with standard machine learning experimental protocols, the dataset is bifurcated into a training set and a testing set, following a 7:3 ratio. Specifically, 70% of the data (amounting to 1152 samples) is allocated for training purposes, while the remaining 30% (comprising 494 samples) is reserved for testing. This segmentation is crucial for detecting the potential overfitting of the model on test data and for assessing its overall efficacy. The unique linguistic attributes of Modern French poetry make preprocessing especially significant. By eliminating semantically non-essential auxiliary words, the preprocessing step ensures that the machine learning model focuses on learning from the most crucial vocabulary, thereby enhancing both the efficiency and accuracy of the model.

Additionally, when we explore the application of AI technology in the analysis of Modernist French poetry, an AI analysis proves invaluable in uncovering linguistic patterns and stylistic novelties. However, it is crucial to recognize that poetry is not an isolated entity but is shaped by its historical, literary, and cultural milieu. Overlooking these contexts could oversimplify and mislead interpretations.

We advocate for an AI analytical approach that encompasses the wider literary ecosystem. This encompasses: (1) Understanding the historical evolution of Modernist French poetry; (2) Investigating its connections with literary traditions like Symbolism and Surrealism; (3) Examining how these poems' receptions and interpretations have evolved over time. Incorporating these contextual elements enables a more thorough and nuanced

comprehension of Modernist French poetry. AI technologies are adept at detecting inherent patterns and links in poetry, but it is the contextual knowledge that frames these findings. Such a methodology enhances not only the precision and pertinence of AI analysis but also enriches our overall grasp of Modernist French poetry. Practical examples of this methodology include: (1) Utilizing AI to discern links between poetry and historical events; (2) Examining the stylistic features of Modernist French poetry relative to preceding literary movements; (3) Monitoring the evolving reception and interpretations of these poems through different historical critiques.

Finally, for machine learning classification purposes, textual data must be converted into numerical feature vectors. This study contrasts two text feature extraction techniques, TF-IDF and Doc2Vec, which highlights that feature extraction is a crucial phase in converting the text into feature vectors suitable for classification models.

#### 3.1.1. Text Processing and Vector Construction

The process of transforming raw text data into numerical representations suitable for machine learning algorithms is crucial for effective text classification. In this section, we delve into the essential steps involved in text processing and vector construction for poetry classification.

#### 3.1.2. Text Preprocessing: Laying the Foundation

Text preprocessing is the initial stage where the raw text, often riddled with inconsistencies and noise, is transformed into a cleaner and more structured format. This preparatory step involves several crucial steps:

1. **Normalization:** Normalization ensures uniformity in the text representation by converting all characters to lowercase, eliminating punctuation marks and special characters, and handling inconsistencies in whitespace and encoding. This standardization helps in reducing the vocabulary size and unifying the text representation, making it easier for machine learning algorithms to process and analyze.
2. **Tokenization:** Tokenization breaks down the text into individual tokens, which can be words, phrases, punctuation marks, or any other meaningful units of language. This step is essential for identifying and extracting relevant features from the text. The choice of tokenization strategy depends on the specific task and language, and in the context of poetry classification, it is crucial to consider the unique characteristics of poetic language, such as wordplay, neologisms, and unconventional syntax.
3. **Stop Word Removal:** Stop words are common words that carry little or no meaning in a given context, such as "le," "un," "une," and "à." Removing stop words can reduce the dimensionality of the vector space and improve classification performance by eliminating noise and focusing on more meaningful terms. However, the decision of whether or not to remove stop words requires careful consideration, as some stop words may carry contextual significance in poetry, especially in cases where they contribute to the rhythm, rhyme, or overall structure of the poem.

### *3.2. Feature Extraction Methods: TF-IDF and Doc2Vec*

Our research investigates the role of feature extraction in poetry classification, focusing on techniques like TF-IDF. We compare TF-IDF and Doc2Vec to analyze their effectiveness in transforming text into vectors suitable for machine learning models. This exploration highlights the significance of converting raw text into numerical formats for successful text classification, emphasizing the process's importance in poetry analysis.

#### 3.2.1. TF-IDF

Feature extraction, which is essential in our study, involves transforming preprocessed text into a numerical format, thereby capturing key linguistic characteristics crucial for classification. This enables machine learning algorithms to discern patterns and formulate insights. Our approach utilized two predominant feature extraction methods:

- **Term Frequency (TF)**: This quantifies the frequency of a term within a document, signifying its role in encapsulating the document's essence. A term's elevated TF value denotes its recurrent presence, suggesting its pertinence to the theme or genre of the document.
- **Inverse Document Frequency (IDF)**: This assesses a term's rarity across a document collection, granting higher significance to rarer terms. This methodology accentuates the value of distinctive terms that contribute to differentiating documents.

Integrating TF and IDF, we generated a TF-IDF vector for each document, encapsulating both the term frequency within and its rarity across the corpus. This comprehensive representation enriches the characterization of document content for classification.

Yet, when analyzing Modernist French poetry using TF-IDF, we encounter limitations in capturing the nuanced styles and themes. Our innovation lies not in discarding TF-IDF but in complementing it with Doc2Vec and various machine learning algorithms, including SVM. This combination creates a detailed and sophisticated toolkit for probing Modernist French poetry. This methodology addresses initial limitations and establishes new benchmarks in AI-driven literary analysis. We aim to continually investigate and integrate emerging machine learning techniques, enhancing our model's analytical capabilities and ensuring its sustained relevance in literary studies.

In addition, we must pay more attention to these major concerns:

1. **Lemmatization and Stemming: Preserving Linguistic Integrity**: Lemmatization and stemming are essential in maintaining the integrity of poetic language for classification. Lemmatization is particularly valuable in poetry for preserving grammatical structure and meaning, while stemming reduces words to their root forms. The choice between these methods hinges on balancing grammatical accuracy with computational simplicity.
2. **Balancing Simplicity and Effectiveness: Striking the Right Chord**: This involves finding a balance in text processing techniques for poetry classification. The combination of TF-IDF vectorization, Doc2Vec, and lemmatization optimizes this balance, capturing essential linguistic features while preserving the unique characteristics of poetry. This balance is crucial for maintaining the richness of poetic language without oversimplification, ensuring effective classification through machine learning.
3. **Enhancing Poetic Language Analysis in Machine Learning-Driven Classification**: The future of poetry classification in AI hinges on addressing challenges like subjective language and poetic diversity. The research should focus on developing methods to handle these complexities, incorporating comprehensive poetic knowledge, and leveraging advancements in NLP and machine learning. This will foster a deeper understanding and appreciation of poetic language, which contributes significantly to the field's advancement.

Here, a word's occurrence in a given document is quantified by a metric called Term Frequency (TF). Conversely, Inverse Document Frequency (IDF) assesses the significance of the term throughout the whole corpus. Multiplying a word's TF value by its IDF value yields its TF-IDF composite score. A higher composite score signifies the word's elevated importance within the document in question.

Subsequent sections will delve into the mathematical intricacies and variations of the TF-IDF algorithm.

Let $M$ stand for a collection of papers, represented as:

$$M = \left\{ d_1, d_2, \ldots, d_j, \ldots d_{|M|} \right\}$$

where $d_j$ represents the $j$th document in $M$.

Allow $M$ to include a collection of words, $K$, represented as:

$$K = \left\{ t_1, t_2, \ldots, t_j, \ldots t_{|K|} \right\}$$

where $t_i$ stands for the $i$ th in the element of $K$.

The occurrence of a term in a text is quantified using the term frequency (TF) metric. For instance, the formula for determining the frequency with which occurrences of the sequence $t_i$ occur in sequence $d_j$ is as follows:

$$TF_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}},$$

where $n_{i,j}$ is the total number of occurrences of word $t_i$ in document $d_j$, and $\sum_k n_{k,j}$ is the total occurrences of all terms $K$ in document $d_j$.

Overall, the TF-IDF combines the word's frequency in the text with its scarcity in the corpus in order to assess the word's relevance inside. Word frequency in a document serves to determine a word's TF by dividing the number of occurrences of the word by the overall number of words in the content of the document.

A word's scarcity in a corpus can be assessed with the inverted document frequency (IDF) method. Take $IDF_i$ as an example. The logarithm of the ratio of documents in the corpus that includes the term to the total number of documents in the corpus is the formula for the IDF. Here is the formula:

$$IDF_i = \log \frac{|M|}{|d \in M : t_i \in d|} \tag{1}$$

where, for a large number of documents, $|M|$, the corpus is considered. The total number of occurrences of $t_i$ in the corpus is represented by $|d \in M : t_i \in d|$.

Word $t_i$ 's inverse document frequency (IDF) is found by using the formula:

$$IDF_i = \log \frac{|M|}{|d \in M : t_i \in d|}$$

where the total number of documents in the corpus is $|M|$. The number of documents in the corpus that include the term $t_i$ is given by $|d \in M : t_i \in d|$. The larger the $IDF_i$ value, the fewer are the documents which contain $t_i$, indicating that $t_i$ has a significant capacity to discriminate between documents.

TF-IDF stands for Term Frequency-Inverse Document Frequency. A word's TF-IDF in a document is determined by multiplying its term frequency (TF) by its inverse document frequency (IDF). The TF-IDF is calculated by multiplying $TF_{ij}$ and $IDF_i$. This is how it is defined:

$$TF - IDF = TF_{ij} \times IDF_i.$$

Finally, in text mining and natural language processing, the TF-IDF is a prominent measure of word value. It is utilized for tasks like document classification, text summarization, and information retrieval. In the context of our research, TF-IDF is a text analysis approach that considers the frequency of a word $t_i$ in a document as well as its frequency in the corpus to determine its relevance. The significance of a term $t_i$ is proportional to its frequency in the corpus. Figure 2 depicts the poetry vector training procedure.

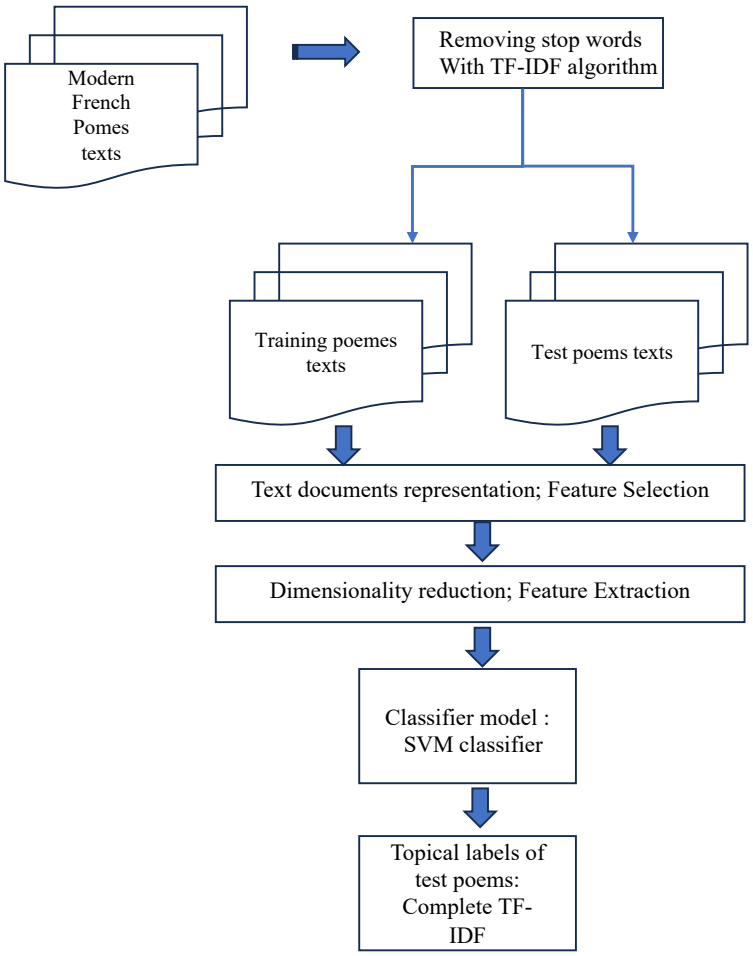


**Figure 2.** TF-IDF processing for Modern French Poetry texts.

#### 3.2.2. Doc2Vec

One-hot encoding, a rudimentary technique for converting words into fixed-size numerical vectors, suffers from two significant limitations. First of all, it necessitates an immense quantity of memory, as the dimensionality of the word vector corresponds to the total number of unique words in the corpus, which might fluctuate between hundreds of thousands. Second, it fails to capture semantic relationships between words, leaving all words as mutually independent, as evidenced by their zero inner products. To address these shortcomings, Mikolov et al. introduced Word2Vec, a neural network-based word representation technique [51]. The key objective of Word2Vec is to maximize the log-probability of a given sequence of training words $w_1, w_2, \ldots, w_T$

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t \mid w_{t-k}, \ldots, w_{t+k}),$$

where $k$ is the dimension of the window utilized to maintain contextual information. In general, the prediction is performed using a multiclass classification with the softmax function, which is illustrated below:

$$p(w_t \mid w_{t-k}, \ldots, w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}},$$

where $y_i$ suggests the $i$-th output value of a feedforward neural network created in the following way:

$$y = b + Uh(w_{t-k}, \ldots, w_{t+k}; W)$$

where variables $b, U, h$, and $W$ are the bias terms between the hidden and output layer, the weight matrix between the hidden and output layer, the average or concatenation of context words, and the word embedding matrix.

Word2Vec's successor, Doc2Vec, aims to maintain the semantic connections between documents by searching for a continuous vector representation of a paragraph or bigger text. In the same way that Word2Vec does, we represent each word as a continuous vector of length $d$, where $d$ is significantly less than the size of the vocabulary in the corpus ($d << |V|$). Also, in the same vector space as words, the document itself is represented as a continuous vector.

Each document in the Doc2Vec model has its own unique vector, which is a row in the matrix $D$. In a similar way, each word is associated with a particular vector, represented by a separate row in the matrix $W$. Introducing $D$ to Eq. merely partially alters the mathematical formulation of the network.

$$y = b + Uh(w_{t-k}, \ldots, w_{t+k}; W, D)$$

Distributed representations of each document in the corpus could be obtained once the network was properly trained. Text mining methods such as clustering and classification benefit greatly from these representations. Both the PV-DM and PV-DBOW designs, which are cornerstones of the Doc2Vec framework, were utilized in the present research. By integrating the vectors generated by these two topologies, we improved classification efficiency.

To accommodate texts of varying lengths, from single words to whole documents, Doc2Vec is an unconstrained framework. Figure 3 depicts the poetry vector generation training method. The model's input layer consists of both the poetry vector and the word vectors in the poem. The tokenized French text is scrolled across a window throughout training to extract lines of poetry. This process generates several training iterations for each poem, all of which involve the poetry vector. In each training iteration, the PoetryVector is used to represent the poem's central idea or theme.

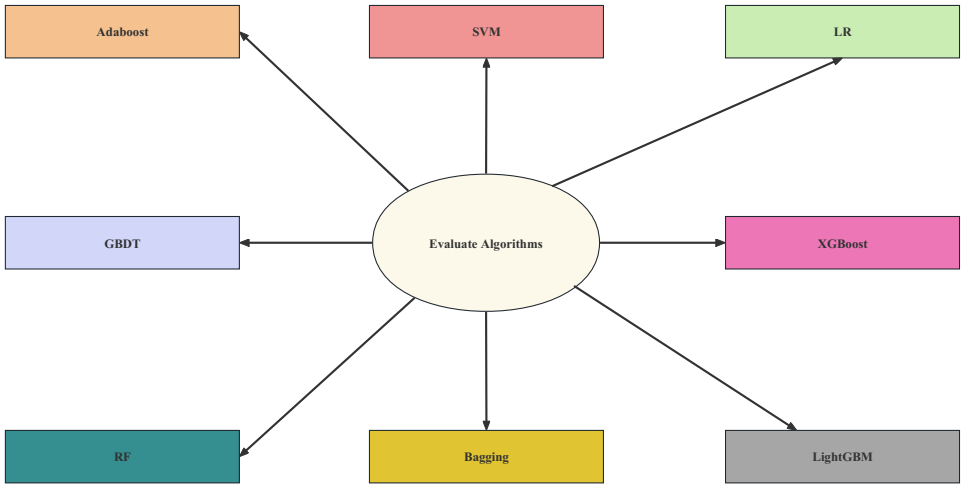


**Figure 3.** Modern French Poetry vector training system.

A frequently used poetry vector is obtained in addition to word vectors throughout training. Each training process employs this vector as part of its input layer. As the window passes across more words, this method obtains more and more efficient results. When training poetry vectors with Doc2Vec, a unique vector can be generated for each poem, which captures the most significant possible connections between the poem's words.

During the training phase, model parameters are refined based on the computed gradient of the neural network's error. Upon reaching convergence, the fidelity of the vectors undergoes continual enhancement. The poetry vector becomes increasingly adept at encapsulating the poem's thematic essence. Concurrently, word vectors are generated, which can be transformed into a format compatible with machine learning classification algorithms.

### *3.3. Comparison of Classification Algorithms*

To adapt to the unique characteristics of poetic texts, we carefully adjusted and optimized the SVM parameters, including regularization terms and kernel function types. Besides SVM, we investigated several other machine learning classifiers, including Logistic Regression (LR), Bagging, Random Forest (RF), AdaBoost, Gradient Boosted Decision Trees (GBDT), XGBoost, and LightGBM. Each algorithm's parameters were meticulously calibrated to achieve the best classification performance. In summary, through extensive testing of these classification algorithms and evaluating their performance based on accuracy, F1 scores, and other relevant metrics, we aim to select the algorithm that best classifies modern French poetry. This process involves not only a technical comparison but also a thorough analysis of each algorithm's ability to address the specific challenges of literary texts. Ultimately, our goal is to determine an optimized model that can accurately reflect the stylistic differences in poetry and provide new perspectives and tools for literary analysis.

### *3.4. Significance of Automated Poetry Classification*

Expert-driven traditional poetry classification methods, prone to subjective biases and inconsistency, are often unsuitable for contemporary poetry and large-scale tasks. Our research introduces an automated classification model using AI and natural language processing, aiming to standardize and objectify poetry classification. This approach, leveraging AI's ability to identify patterns in extensive poetry datasets, provides a more objective, consistent, scalable, and adaptable alternative. Particularly beneficial for new or evolving poetry styles, AI-driven classification can significantly advance poetry understanding and categorization, as shown in Figure 4.

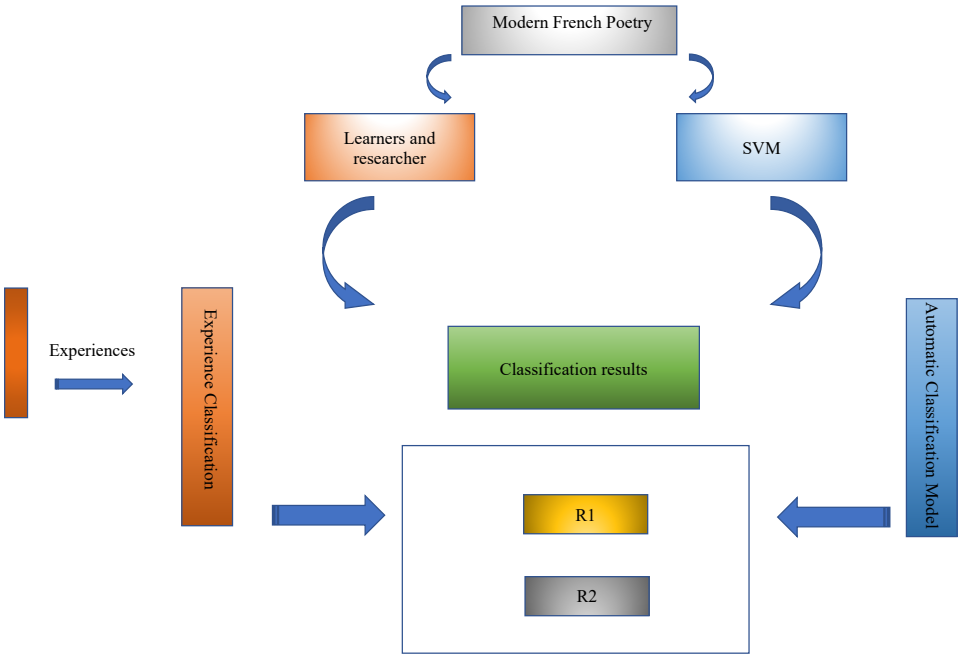


**Figure 4.** SVM Model for Modern French Poetry Learning.

In educational contexts, these AI models offer personalized learning and timely feedback to students, enhancing their poetry style development. For scholars, they aid in

exploring poetry style evolution and genre relationships, and in recognizing contemporary trends. The paper discusses the importance of converting text into numerical data for machine learning, comparing TF-IDF and Doc2Vec techniques. Overall, AI-driven models present a substantial improvement in literary analysis, promising enhanced educational and research methodologies in poetry, with future AI advancements poised to offer even more sophisticated poetry analysis tools.

### *3.5. Pseudo Code for AI-Driven Poetry Classification Model*

This pseudocode outlines the algorithmic process for the AI-based modern French poetry classification in our study. It details the preprocessing of poetry data, including normalization, tokenization, and stopword removal, followed by feature extraction using TF-IDF to analyze word significance and meanings.

The subsequent classification phase employs machine learning models, using extracted features to categorize poetry. This clear, structured pseudocode, which we will refer to as Algorithm 1, not only describes our methodology but also makes the complex computational processes accessible to non-experts, bridging technical complexity and comprehensibility.

In summary, the pseudocode (Algorithm 1) is crucial for understanding our AI-based poetry classification approach, emphasizing transparency and methodological rigor in integrating AI with a literary analysis.

**Algorithm 1** AI-Based Poetry Classification Algorithm

**input:** poetry_dataset $X$, machine_learning_model $M$;
**output:** Classifications $C$;
**function** PREPROCESS_POETRY_DATA(dataset)
  **for** $i = 1$ to length(dataset) **do**
    Normalize text(dataset[i]);
    Tokenize text(dataset[i]);
    Remove stop words(dataset[i]);
    Apply TF-IDF(dataset[i]);
  **end for**
**end function**
**function** EXTRACT_FEATURES(processed_dataset)
  Initialize feature set $F$;
  **for** $i = 1$ to length(processed_dataset) **do**
    $F[i]$ =compute_TF_IDF(processed_dataset[i]);
  **end for**
**end function**
**function** CLASSIFY_POEMS(features, machine_learning_model)
  Initialize classifications $C$;
  **for** $i = 1$ to length(features) **do**
    $C[i]$ = machine_learning_model(features[i]);
  **end for**
**end function**
**function** MAIN(dataset, machine_learning_model)
  processed_dataset = PREPROCESS_POETRY_DATA(dataset);
  features = EXTRACT_FEATURES(processed_dataset);
  classifications = CLASSIFY_POEMS(features, machine_learning_model);
  **return** classifications;
**end function**

## 4. Experiment

Having delineated the intricate workings of our AI model methodology, which employs advanced natural language processing techniques, we now shift our focus to the practical application of this model. The transition from the theoretical underpinnings and technical robustness of the model, demonstrated through various algorithms and a bespoke training process to its real-world applicability in the realm of literary analysis, is pivotal.

This section contextualizes the AI-driven analysis within the broader literary ecosystem of Modernist French poetry. It emphasizes the importance of not only recognizing linguistic patterns and stylistic novelties but also understanding the historical, literary, and cultural contexts that shape these poems. This holistic approach, integrating both technical acumen and nuanced literary comprehension, ensures a more thorough and accurate analysis of the poetic corpus.

### *4.1. Source and Background of Dataset*

The study utilized a dataset from "Bonjour Poésie", a comprehensive source of classic French poetry at https://www.bonjourpoesie.fr/lesgrandsclassiques (accessed on 14 January 2024). Curated by experts, the site offers poems with annotations on the poet, the work, and its historical context. The dataset comprises 1646 poems, representing modern French poetry's three main streams: Symbolism (309 poems), Parnassism (390 poems), and Romanticism (453 poems). It is split into training (70%) and testing (30%) sets for effective AI model evaluation. The advantages of using "Bonjour Poésie" include: **Expert Curation:** Ensuring dataset quality and accuracy. **Contextual Annotations:** Providing insights into poems' backgrounds. **Comprehensive Representation:** Covering the genre's main streams. **Rigorous Dataset Split:** Facilitating thorough model training and evaluation. In summary, "Bonjour Poésie" offers a well-curated, context-rich, and genre-representative dataset, ideal for an AI-driven analysis of modern French poetry.

### *4.2. Training and Test Sets*

In this research, we classified 1646 poems from Modern French Poetry into Romanticism, Parnassian, and Symbolist categories, informed by literary expert opinions. The dataset was allocated into training (1152 poems) and test sets (494 poems), as outlined in Table 1.

**Table 1.** Distribution of Poetry Types in Training and Test Sets.

| Poetry Category | Training Sets | Test Sets |
|---|---|---|
| Romanticism | 453 | 183 |
| Parnasse | 390 | 171 |
| Symbolism | 309 | 140 |

We evaluated machine learning classifiers such as SVM, LR, RF, AdaBoost, GBDT, XGBoost, and LightGBM, using TF-IDF and Doc2Vec for text features. The data split was 70% for training and 30% for testing, with cross-validation to select the best models. SVM, with hyperparameters $C = 1.0$ and $\gamma = 0.1$, performed well, along with LR and GBDT, while RF and AdaBoost were less effective.

This study's thorough methods demonstrated the effectiveness of eight algorithms classification in French poetry classification and emphasized the role of hyperparameters, features, and ensemble techniques. These insights provide a foundation for future enhancements with advanced deep learning algorithms.

### *4.3. Training and Testing Design*

The stages of our experiment were as follows:

**Step 1.** We applied the TF-IDF algorithm to preprocess and transform poems into numerical feature vectors, which were then used for SVM training.

**Step 2.** Following machine learning norms, the TF-IDF processed data was divided into training (1152 samples) and test sets (494 samples), covering Parnasse, Romantisme, and Symbolisme.

**Step 3.** Utilizing grid search and cross-validation, we optimized SVM's parameters to $C = 1.0$ and $\gamma = 0.1$. The model was then trained on the training set and evaluated on the test set comprising various poetry categories.

**Step 4.** We thoroughly analyzed the results, focusing on the accuracy, precision, recall, and F1 score.

This experiment demonstrated the synergy between TF-IDF and SVM, highlighting their combined effectiveness in text classification. The findings offer valuable insights for future text classification research. Overall, SVM's robustness in classification, enhanced by its adjustable parameters, was pivotal in our study of French poetry classification using the TF-IDF feature extraction method.

### *4.4. Hyperparameter Optimization for Classifiers*

Python 3.8.15, pandas 1.5.2, gensim 4.3.0, and scikit-learn 1.2.0 were used for conducting experiments on the French poetry classification. Key steps included data pre-processing and vectorization, followed by training with scikit-learn. The models employed and their fine-tuned parameters were as follows:

- **SVM:** Adapted for high-dimensional poetry data, with rbf kernel. Parameters: $C = 1.0$, penalty = 'l2', max_iter = 1000.
- **LR:** Applied for probabilistic categorization. Parameters: penalty = 'l2', $C = 1.0$, max_iter = 1000, solver = 'lbfgs'.
- **Bagging:** To capture stylistic nuances. Parameters: n_estimators = 60, max_samples = 1.0.
- **RF:** Ensemble method suitable for diverse poetry styles. Parameters: n_estimators = 10, criterion = 'gini', max_features = 'sqrt'.
- **AdaBoost:** For iterative correction and detection of stylistic elements. Parameters: n_estimators = 50.
- **GBDT:** Iterative model capturing stylistic intricacies. Parameters: n_estimators = 100, max_depth = 3.
- **XGBoost:** Efficient for sparse stylistic data. Parameters: max_depth = 6, n_estimators = 100.
- **LightGBM:** Effective in capturing hierarchical poetic styles. Parameters: num_leaves = 31, n_estimators = 250.

Each classifier in our study was meticulously fine-tuned and evaluated using cross-validation on the training data. Following this, they were retrained on the complete dataset and tested on previously unobserved poetry. This methodological approach, which included careful parameter optimization, enabled us to accurately classify Modern French Poetry into distinct schools such as Romanticism, Parnasse, and Symbolism. The details of the classifier parameters and the dataset, comprising 1644 poems, are available at our GitHub repository: https://github.com/wkwg429/FrenchLiteratureTextClassification.git (accessed on 14 January 2024). Notably, the classifiers achieved an impressive accuracy rate of 90%.

In our research, the Support Vector Machine (SVM) emerged as a pivotal supervised machine learning method for classification tasks. SVM's efficacy hinges on its ability to find an optimized hyperplane, thus maximizing the margin between different categories. Figure 5 depicts how the optimal hyperplane is determined: it minimizes the distance between the nearest training data points, known as support vectors, and the hyperplane itself. This larger margin is crucial as it enhances the model's generalization capabilities when dealing with unknown data.

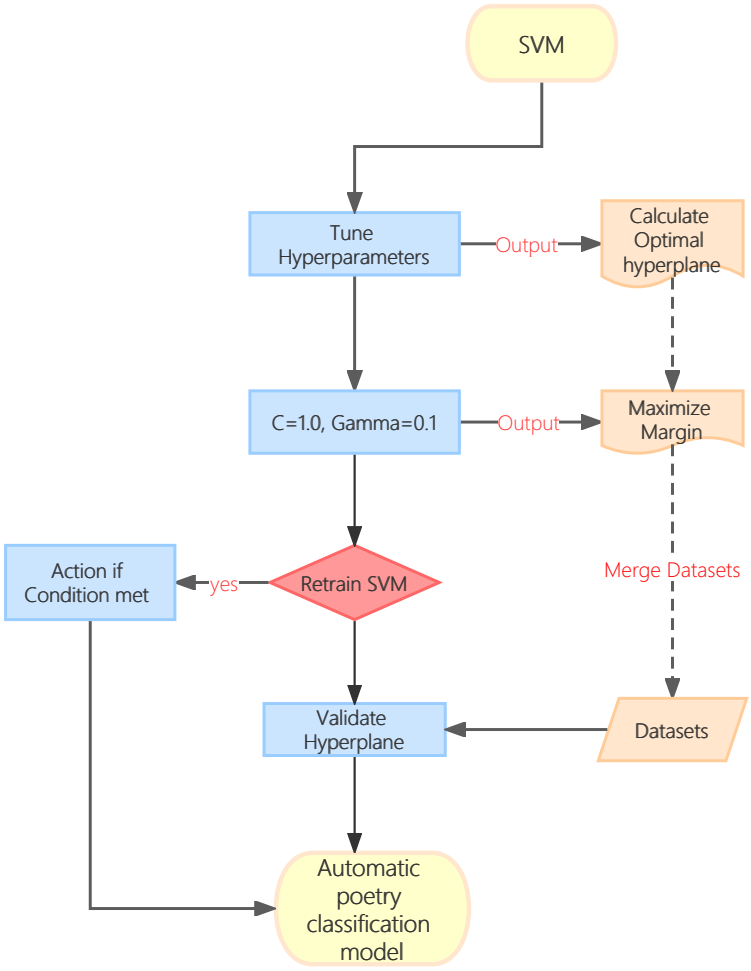


**Figure 5.** Optimal hyperplane in SVM maximizing the margin.

In the context of SVM, we utilized cross-validation to fine-tune its hyperparameters. The optimal model, as determined by this process, was subsequently retrained on the entire collection of training and testing poetry. A cornerstone of our study, the SVM models were adeptly implemented to classify the diverse styles of poetry. We integrated these models with traditional feature extraction methods, notably Term Frequency–Inverse Document Frequency (TF-IDF), to capture the nuanced differences among various poetic schools.

### *4.5. Analysis of the Results*

Due to the inherent nuances and complexity of poetry, the process of categorizing French poetry into separate stylistic genres, such as Romanticism, Parnasse, and Symbolism, can be a complex and nuanced endeavor. Our experimental approach consisted of using a variety of classifiers in conjunction with the TF-IDF feature extraction method so that we could efficiently distinguish between these stylistic variances. Results showed that SVM with TF-IDF features, which provided optimal results in the Table 2, achieved the best performance with an accuracy of 0.7571 and an F1 score of 0.7518, while other classifiers, despite a reasonable performance, were less effective.

**Table 2.** Performance of classifiers on French poetry classification with different features.

| Feature | Classification | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| tfidf | lr | 0.7045 | 0.7202 | 0.6872 | 0.6879 |
| tfidf | svm | 0.7571 | 0.7549 | 0.7504 | 0.7518 |
| tfidf | bagging | 0.7510 | 0.7533 | 0.7418 | 0.7442 |
| tfidf | rf | 0.5506 | 0.5469 | 0.5391 | 0.5387 |
| tfidf | adaboost | 0.6275 | 0.6952 | 0.6108 | 0.6054 |
| tfidf | gbdt | 0.6194 | 0.6172 | 0.6190 | 0.6177 |
| tfidf | xgb | 0.6377 | 0.6375 | 0.6410 | 0.6355 |
| tfidf | lgb | 0.6721 | 0.6688 | 0.6703 | 0.6685 |
| doc2vec_25 | lr | 0.6640 | 0.6570 | 0.6567 | 0.6566 |
| doc2vec_25 | svm | 0.6599 | 0.6542 | 0.6538 | 0.6531 |
| doc2vec_25 | bagging | 0.6640 | 0.6572 | 0.6571 | 0.6566 |
| doc2vec_25 | rf | 0.5891 | 0.5781 | 0.5752 | 0.5733 |
| doc2vec_25 | adaboost | 0.6538 | 0.6455 | 0.6457 | 0.6453 |
| doc2vec_25 | gbdt | 0.6842 | 0.6795 | 0.6798 | 0.6795 |
| doc2vec_25 | xgb | 0.6579 | 0.6515 | 0.6509 | 0.6508 |
| doc2vec_25 | lgb | 0.6761 | 0.6726 | 0.6694 | 0.6699 |
| ensemble_spar | lr | 0.7045 | 0.6983 | 0.6986 | 0.6982 |
| ensemble_spar | svm | 0.7429 | 0.7389 | 0.7367 | 0.7363 |
| ensemble_spar | bagging | 0.7368 | 0.7326 | 0.7315 | 0.7312 |
| ensemble_spar | rf | 0.5749 | 0.5742 | 0.5669 | 0.5683 |
| ensemble_spar | adaboost | 0.6579 | 0.6475 | 0.6477 | 0.6475 |
| ensemble_spar | gbdt | 0.7287 | 0.7266 | 0.7250 | 0.7242 |
| ensemble_spar | xgb | 0.7227 | 0.7252 | 0.7196 | 0.7172 |
| ensemble_spar | lgb | 0.7368 | 0.7394 | 0.7323 | 0.7322 |

The classifiers' performance in terms of accuracy is summarized in Table 3. The SVM classifier clearly outperforms the competition with an accuracy of 0.7571, demonstrating its ability to handle the stylistic variations present in the presented dataset with ease.

**Table 3.** Classifier Efficiency with TF-IDF Features.

| Classifier | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| LR | 0.7045 | 0.7202 | 0.6872 | 0.6879 |
| **SVM** | **0.7571** | **0.7549** | **0.7504** | **0.7518** |
| Bagging | 0.7510 | 0.7533 | 0.7418 | 0.7442 |
| RF | 0.5506 | 0.5469 | 0.5391 | 0.5387 |
| AdaBoost | 0.6275 | 0.6952 | 0.6108 | 0.6054 |
| GBDT | 0.6194 | 0.6172 | 0.6190 | 0.6177 |
| XGBoost | 0.6377 | 0.6375 | 0.6410 | 0.6355 |
| LightGBM | 0.6721 | 0.6688 | 0.6703 | 0.6685 |

This is in line with how the SVM is meant to function, which is to locate a hyperplane that can partition the dataset into classes in the most effective way possible. More specifically, this is consistent with the intended operation of the SVM that was developed. When it pertains to the subject of our research, the Support Vector Machine (SVM) is preferable to other classifiers in terms of its potential to recognize the change that takes place when shifting from one form of poetry to another. This is because the SVM is able to distinguish the shift more accurately.

Random Forest (RF) failed to perform well as a classifier, achieving an accuracy of just 0.5506. It is possible that this is because it struggled to handle the delicate stylistic components of the dataset. The AdaBoost classifier, on the other hand, performed exceptionally well in precision, earning a score of 0.6952, but its lower recall score shows that there is potential for development.

Notably, the Support Vector Machine, or SVM, won the competition for the greatest F1 Score. This score is a combination of accuracy and recall, and as such, it is the option that

is both the best balanced and an effective alternative for categorizing this specific dataset. Figure 6 is a graphic illustration of these essential indicators, which aims to promote a more intuitive understanding of the differences in performance that were observed.

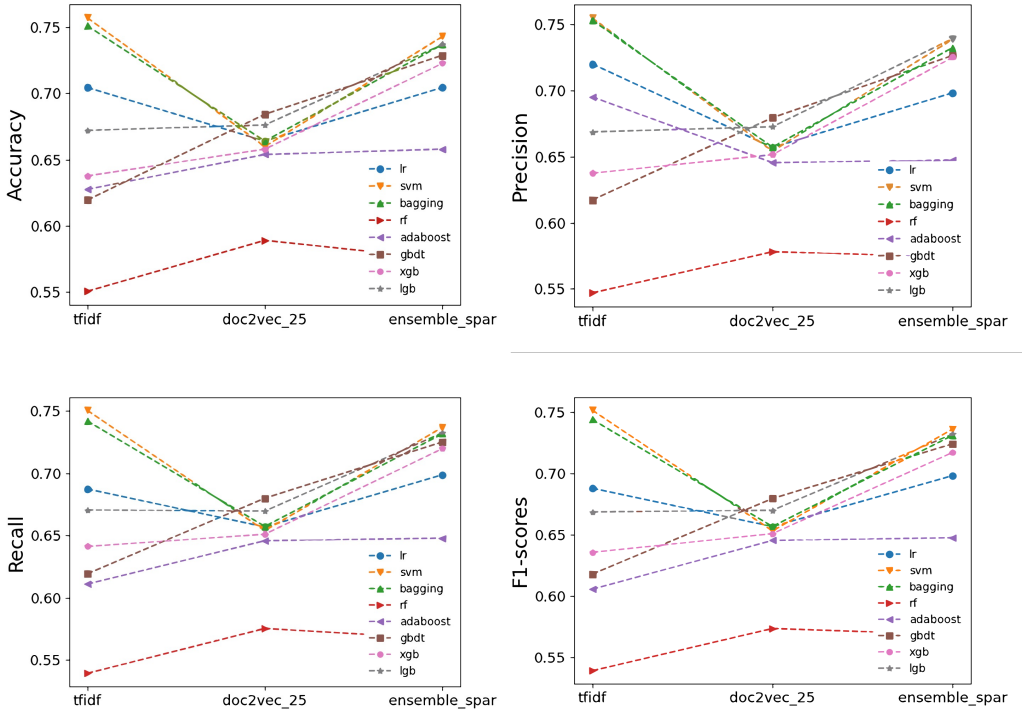


**Figure 6.** Four Classification Models Assessed Through Experiments With Real Data.

#### 4.5.1. Performance Using TF-IDF Features

Table 4 summarizes the results using TF-IDF features. SVM achieves the best performance, with an accuracy and F1 score of 0.757 and 0.752, respectively. The other classifiers have an inferior performance:

**Table 4.** Classifier performance using method TF-IDF.

| Classifier | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| **SVM** | **0.7571** | **0.755** | **0.750** | **0.7518** |
| LR | 0.704 | 0.720 | 0.687 | 0.688 |
| Bagging | 0.751 | 0.753 | 0.742 | 0.744 |
| RF | 0.551 | 0.547 | 0.539 | 0.539 |
| AdaBoost | 0.628 | 0.695 | 0.611 | 0.605 |
| GBDT | 0.619 | 0.617 | 0.619 | 0.618 |
| XGBoost | 0.638 | 0.637 | 0.641 | 0.636 |
| LightGBM | 0.672 | 0.669 | 0.670 | 0.668 |

#### 4.5.2. Performance Using Doc2Vec Features

Table 5 demonstrates that using Doc2Vec features leads to the highest levels of accuracy (0.684) and F1 score (0.679) for GBDT. The performance of SVM drops significantly to only 0.66 accuracy with Doc2Vec. LR, Bagging and LightGBM also perform reasonably well. But overall, results indicate that TF-IDF is better than Doc2Vec for this task.

**Table 5.** Classifier performance using Doc2Vec.

| Classifier | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| GBDT | 0.684 | 0.679 | 0.680 | 0.679 |
| LR | 0.664 | 0.657 | 0.657 | 0.657 |
| Bagging | 0.664 | 0.657 | 0.657 | 0.657 |
| LightGBM | 0.676 | 0.673 | 0.669 | 0.670 |
| SVM | 0.660 | 0.654 | 0.654 | 0.653 |
| XGBoost | 0.658 | 0.652 | 0.651 | 0.651 |
| AdaBoost | 0.654 | 0.645 | 0.646 | 0.645 |
| RF | 0.589 | 0.578 | 0.575 | 0.573 |

#### 4.5.3. Ensemble Technique

The ensemble method in Table 6 improves accuracy of most classifiers by 2–5% over individual features. Using ensemble features, SVM obtains a best-in-class accuracy of 0.743 and an F1 score of 0.736.

**Table 6.** Classifier performance using ensemble features.

| Classifier | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| **SVM** | **0.743** | **0.739** | **0.737** | **0.736** |
| LightGBM | 0.737 | 0.739 | 0.732 | 0.732 |
| GBDT | 0.729 | 0.727 | 0.725 | 0.724 |
| XGBoost | 0.723 | 0.725 | 0.720 | 0.717 |
| Bagging | 0.737 | 0.733 | 0.732 | 0.731 |
| LR | 0.704 | 0.698 | 0.699 | 0.698 |
| AdaBoost | 0.658 | 0.647 | 0.648 | 0.647 |
| RF | 0.575 | 0.574 | 0.567 | 0.568 |

In summary, the investigation shows that SVM is the best TF-IDF classifier for French poetry categorization. The ensemble technique improves performance. The findings suggest that the multi-class text categorization machine learning method's SVM classifier is the most successful in classifying French poetry into stylistic genres, but it is important to understand its strengths and weaknesses. Classifiers may vary depending on the needs, such as favoring precision above recall or vice versa. Further research into parameter tweaking and other feature extraction approaches may provide even better findings, revealing more about French literary genres. The enhanced SVM model has been developed on the entire set of training data with the help of ensemble features and the revised hyperparameters $C = 1.0$, as well as through the $\gamma = 0.1$ process of cross-validation. This model was tested on 494 unobserved poems. The model includes 0.743 test set accuracy, 0.73–0.74 precision, recall, and F1 score. This suggests a good generalization to new French poetry data.

## 5. Discussion

This study contributes to Modern French Poetry classification using machine learning, with our SVM-based model achieving an accuracy of 0.743. However, its performance, considering the limited dataset and basic feature set, suggests room for improvement. The model's generalization ability needs testing with a larger, more diverse dataset for robust evaluation. The feature representations used, TF-IDF and Doc2Vec, capture basic stylistic elements but may not fully grasp poetry's intricate linguistic and contextual nuances. Advanced features like word embeddings and latent semantic analysis could better differentiate poetic styles. Moreover, our focus on SVM, despite its effectiveness in the text classification, warrants an exploration of other algorithms like CNNs and RNNs, which might better capture poetry's semantic and contextual intricacies. The study also opens possibilities for integrating this style classification model into poetry education tools, potentially enhancing student engagement and understanding. However, further research is needed to ensure the effectiveness and personalization of these technologies

in educational settings. In summary, while our SVM model shows promise for classifying Modern French Poetry, exploring advanced features, diverse algorithms, and larger datasets could improve its performance. The model's integration into educational tools holds potential, but requires a careful evaluation of its impact on learning. Addressing these aspects could further advance computational poetry and educational technology.

### *5.1. In-Depth Comparative Analysis of AI-Driven Poetry Classification Model*

Building upon the foundations established in the preceding sections, this section presents a comparative analysis that incorporates the latest advancements in AI to augment our poetry classification model's effectiveness. The integration of these sophisticated methodologies is crucial for effectively navigating the unique complexities associated with Modernist French poetry. Furthermore, this integration lays the groundwork for future research directions, inspiring further exploration and innovation in the field.

Li et al. discussed the cutting-edge data processing techniques providing valuable insights into the refinement of our poetry classification approach [43]. By implementing similar data distillation methods, we can enhance the accuracy and efficiency of our model, ensuring that even the subtlest nuances of Modernist French poetry are captured and analyzed effectively. Further, Shingala explored the integration of NLP with text classification and underscores the potential of combined methodologies [44]. This approach is particularly beneficial in our model, as it improves the accuracy and applicability of the classification, enabling a deeper understanding of the intricate linguistic structures and themes inherent in the poetry. By evaluating and incorporating the strengths of various classification methods, our model can achieve a higher degree of precision in categorizing and interpreting poetic texts. Particularly noteworthy is the research by Wang et al. [48] on predicting the lithium-ion battery lifespan using the PF-BiGRU-TSAM model, which integrates Particle Filtering, Bidirectional Gated Recurrent Units, and a Time Series Attention Mechanism, offering a valuable parallel perspective to our study. Additionally, the method proposed by Bilski [52], "Lifetime Extension via Levenberg-Marquardt Neural Networks" demonstrates proficiency in handling complex, time-sensitive data. This proficiency is directly relevant to our poetry classification model, where understanding temporal and stylistic shifts in language is crucial. Finally, the recent advancements in AI, such as the research made by Wang et al. [40], which incorporates contrastive learning and adversarial training, offer novel approaches for tackling complex classification challenges.

In summary, this research synthesizes findings from both established and emerging areas of AI research to enhance our poetry classification model. Specifically, the methodologies adopted are tailored to tackle the unique challenges inherent in Modernist French poetry, including ambiguity, subjectivity, and unconventional syntax. These challenges necessitate a nuanced approach for a further comprehensive analysis in the future study. Consequently, our model not only addresses the specificities of Modernist French poetry but also aligns with the latest advancements in AI, resulting in a more robust and insightful analytical framework.

### *5.2. Summary of Findings and Future Directions*

Our study on Modern French Poetry classification using machine learning has led to significant insights:

- The SVM-based model effectively classified poetry with an accuracy of 0.743, demonstrating machine learning's potential in capturing poetic styles.
- Limitations in current feature representations (TF-IDF and Doc2Vec) were observed. Advanced methods like word embeddings could better capture poetic nuances.
- The dataset's size limits the model's ability to generalize. A larger dataset would offer a more robust evaluation of its performance.

The study acknowledges limitations, setting the course for future research:

- Enhancing generalization across diverse poetic styles requires a more extensive dataset.

- Investigating sophisticated feature representations like latent semantic analysis could improve style discrimination.
- Exploring other machine learning techniques, such as CNNs and RNNs, might capture poetic nuances more effectively.
- Integrating the model into educational platforms demands careful consideration of effectiveness and personalization. Further research should focus on ensuring these technologies augment educational outcomes and democratize learning.

In conclusion, while our SVM model shows promise, addressing these limitations and exploring new methodologies could significantly enhance its capabilities and educational impact. This research serves as a foundation for future studies in computational poetry analysis and its integration into educational settings.

**Author Contributions:** Conceptualization, L.Y.; methodology, L.Y.; software, G.W., L.Y.; validation, L.Y.; formal analysis, L.Y.; investigation, L.Y.; resources, L.Y.; data curation, G.W., L.Y.; writing—original draft preparation, L.Y.; writing—review and editing, H.W., L.Y.; visualization, L.Y.; supervision, H.W.; project administration, H.W.; funding acquisition, L.Y.. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the Central Basic Research Operation Fee Cultivation Project of Zhongnan University of Economics and Law for the year 2019 under Grant number Category B, No. 61. School of Foreign Studies, Zhongnan University of Economics and Law.

**Data Availability Statement:** The codes and datasets associated with the paper are shared at https://github.com/wkwg429/FrenchLiteratureTextClassification.git, accessed on 14 January 2024.

**Acknowledgments:** We extend our heartfelt thanks to Gang Wang for his significant contributions to software development and data curation, and to Hongjun Wang for his invaluable guidance and supervision. This article is the culmination of joint efforts from all contributors.

**Conflicts of Interest:** The authors declare no conflict of interest

## References

1. Middleton, P.; Marsh, N. *Teaching Modernist Poetry*; Cambridge University Press: Cambridge, UK, 2010. [CrossRef]
2. Camilleri, M.; Camilleri, A. The Sustainable Development Goal on Quality Education. In *The Future of the UN Sustainable Development Goals*; United Nations: New York, NY, USA, 2019. [CrossRef]
3. Gallagher, M. Ethics in the Absence of Reference: Levinas and the (Aesthetic) Value of Diversity. *Levinas Stud.* **2012**, *7*, 125–195. [CrossRef]
4. Röhnert, J. Vom Spleen de Paris zum spleenigen Paris. *Z. Für Lit. Und Linguist.* **2010**, *40*, 168–179. [CrossRef]
5. Sherry, V. *The Great War and the Language of Modernism*; Oxford University Press: Oxford, UK, 2003. [CrossRef]
6. Fetzer, G.W. Linguistique et analyse de la poésie: Apports actuels. *Contemp. Fr. Francoph. Stud.* **2016**, *20*, 470–477. [CrossRef]
7. Gardner, W.O. Advertising Tower: Japanese Modernism and Modernity in the 1920s. *Comp. Crit. Stud.* **2007**, *4*, 455–459. [CrossRef]
8. Tanasescu, C.; Paget, B.; Inkpen, D. Automatic Classification of Poetry by Meter and Rhyme. In Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference (FLAIRS), Key Largo, FL, USA, 16–18 May 2016; pp. 823–828.
9. Prieto, J.C.S.; Gamazo, A.; Cruz-Benito, J.; Therón, R.; García-Peñalvo, F. AI-Driven Assessment of Students: Current Uses and Research Trends. In Proceedings of the International Conference on Human-Computer Interaction, Virtual, 19–24 July 2020; Springer International Publishing: Cham, Germany, 2020; pp. 292–302. [CrossRef]
10. Huang, M.H.; Rust, R. Artificial Intelligence in Service. *J. Serv. Res.* **2018**, *21*, 155–172. [CrossRef]
11. Yang, S.J.H.; Ogata, H.; Matsui, T.; Chen, N. Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100008. [CrossRef]
12. Hosseini, M.; Nasrabadi, M.; Mollanoroozy, E.; Khani, F.; Mohammadi, Z.; Barzanoni, F.; Amini, A.; Gholami, A. Relationship of sleep duration and sleep quality with health-related quality of life in patients on hemodialysis in Neyshabur. *Sleep Med.* **2023**, *5*, 100064. [CrossRef] [PubMed]
13. Wei, H.; Geng, H. The Revival of Classical Chinese Poetry Composition: A Perspective from the New Liberal Arts. *Int. J. Comp. Lit. Transl. Stud.* **2022**, *10*, 18–22. [CrossRef]
14. Kempton, H.M.; School of Psychology. Holy be the Lay: A Way to Mindfulness Through Christian Poetry. *OBM Integr. Complement. Med.* **2021**, *7*, 011. [CrossRef]

15. Yusifov, T. Methods of Analysis of Lyrical Texts in Foreign Literature Classes in Higher Educational Institutions (Based on the Works of Nazim Hikmet). *Sci. Bull. Mukachevo State Univ. Ser. Pedagogy Psychol.* **2022**, *8*, 81–88. [CrossRef]
16. Alsharif, O.; Alshamaa, D.; Ghneim, N. Emotion Classification in Arabic Poetry using Machine Learning. *Int. J. Comput. Appl.* **2013**, *65*, 10–15.
17. Ahmed, M.A. A Classification of Al-hur Arabic Poetry and Classical Arabic Poetry by Using Support Vector Machine, Naïve Bayes, and Linear Support Vector Classification. *Iraqi J. Comput. Sci. Math.* **2022**, *3*, 128–137. [CrossRef]
18. Rostampour, S. Word Order of Noun and Verb Phrases in Contemporary Persian and English Poems. *J. Adv. Linguist.* **2017**, *8*, 1229–1235. [CrossRef]
19. Tangirova, D. Variety Of Forms And Genres In Modern Poetry. *Am. J. Interdiscip. Innov. Res.* **2020**, *2*, 32–41. [CrossRef]
20. Dadhich, A.; Thankachan, B. Opinion Classification of Product Reviews Using Naïve Bayes, Logistic Regression and Sentiwordnet: Challenges and Survey. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1099*, 012071. [CrossRef]
21. Puri, S.; Singh, S. Text recognition in bilingual machine printed image documents—Challenges and survey: A review on principal and crucial concerns of text extraction in bilingual printed images. In Proceedings of the 2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, 7–8 January 2016; pp. 1–8. [CrossRef]
22. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A Survey on Text Classification: From Shallow to Deep Learning. *arXiv* **2020**, arXiv:2008.00364. [CrossRef].
23. Kowsari, K.; Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.E.; Brown, D.E. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. [CrossRef]
24. Li, Z.; Qian, S.; Cao, J.; Fang, Q.; Xu, C. Adaptive Transformer-Based Conditioned Variational Autoencoder for Incomplete Social Event Classification. In Proceedings of the 30th ACM International Conference on Multimedia, New York, NY, USA, 10–14 October 2022. [CrossRef]
25. Wu, M.; Pan, S.; Zhu, X.; Zhou, C.; Pan, L. Domain-Adversarial Graph Neural Networks for Text Classification. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019. [CrossRef]
26. Rafeeque, P.C.; Sendhilkumar, S. A survey on Short text analysis in Web. In Proceedings of the 2011 Third International Conference on Advanced Computing, Chennai, India, 14–16 December 2011; pp. 365–371. [CrossRef]
27. Xia, C.; Yin, W.; Feng, Y.; Yu, P. Incremental Few-shot Text Classification with Multi-round New Classes: Formulation, Dataset and System. *arXiv* **2021**, arXiv:2104.11882; pp. 1351–1360. [CrossRef]
28. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent Convolutional Neural Networks for Text Classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2267–2273. [CrossRef]
29. Wang, P.; Hu, J.; Zeng, H.J.; Chen, Z. Using Wikipedia knowledge to improve text classification. *Knowl. Inf. Syst.* **2009**, *19*, 265–281. [CrossRef]
30. Abel, J.; Lantow, B. A Methodological Framework for Dictionary and Rule-based Text Classification. In Proceedings of the KDIR, Vienna, Austria, 16–18 September 2019; pp. 330–337.
31. Wang, H.; Hong, M. Supervised Hebb rule based feature selection for text classification. *Inf. Process. Manag.* **2019**, *56*, 167–191. [CrossRef]
32. Hadi, W.; Al-Radaideh, Q.A.; Alhawari, S. Integrating associative rule-based classification with Naïve Bayes for text classification. *Appl. Soft Comput.* **2018**, *69*, 344–356. [CrossRef]
33. Mishra, M.; Vishwakarma, S.K. Text Classification based on Association Rule Mining Technique. *Int. J. Comput. Appl.* **2017**, *169*, 46–50. [CrossRef]
34. Liu, J.; Bai, R.; Lu, Z.; Ge, P.; Aickelin, U.; Liu, D. Data-driven regular expressions evolution for medical text classification using genetic programming. In Proceedings of the 2020 IEEE Congress on Evolutionary Computation (CEC), Glasgow, UK, 19–24 July 2020; pp. 1–8.
35. Faguo, Z.; Fan, Z.; Bingru, Y.; Xingang, Y. Research on Short Text Classification Algorithm Based on Statistics and Rules. In Proceedings of the 2010 Third International Symposium on Electronic Commerce and Security, Nanchang, China, 29–31 July 2010; pp. 3–7. [CrossRef]
36. Choi, J.; Kilmer, D.; Mueller-Smith, M.; Taheri, S.A. Hierarchical Approaches to Text-based Offense Classification. *Sci. Adv.* **2023**, *9*, eabq8123. [CrossRef] [PubMed]
37. Elmahdy, A.; Inan, H.A.; Sim, R. Privacy Leakage in Text Classification: A Data Extraction Approach. *arXiv* **2022**, arXiv.2206.04591. [CrossRef].
38. Guo, Z.; Zhang, R.; Huan, H. Text Classification Method Based on PEGCN. Online Resource, 2024. Available online: https://www.authorea.com/doi/full/10.22541/au.168210369.91223406/v1 (accessed on 14 January 2024).
39. Wang, H.; Tian, K.; Wu, Z.; Wang, L. A Short Text Classification Method Based on Convolutional Neural Network and Semantic Extension. *Int. J. Comput. Intell. Syst.* **2020**, *14*, 367. [CrossRef]
40. Wang, X.; Zhang, J.; Zhao, L.; Wang, X. Research on Text Classification Technology Integrating Contrastive Learning and Adversarial Training. In Proceedings of the 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 26–28 May 2023; pp. 860–865.
41. Zhang, R.; Guo, Z.; Huan, H. Text Classification Based on an Improved Graph Neural Network. Online Resource, 2022. Available online: https://www.researchsquare.com/article/rs-2385115/v1 (accessed on 21 December 2022).

42. Kargupta, P.; Komarlu, T.; Yoon, S.; Wang, X.; Han, J. MEGClass: Extremely Weakly Supervised Text Classification via Mutually-Enhancing Text Granularities. *arXiv* **2023**, arXiv.2304.01969. [CrossRef].
43. Li, Y.; Li, W. Data Distillation for Text Classification. *arXiv* **2021**, arXiv.2104.08448. [CrossRef].
44. Shingala, A. Three Phase Synthesizing of Nlp and Text Classification for Query Generation. *Int. J. Res. Appl. Sci. Eng. Technol.* **2017**, *5*, 39–42. [CrossRef]
45. Abdulla, H.H.H.A.; Awad, W.S. Text Classification of English News Articles using Graph Mining Techniques. In Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022), Virtual, 3–5 February 2022; pp. 926–937.
46. Zhang, S.; Li, H.; Zhang, S. Job opportunity finding by text classification. *Procedia Eng.* **2012**, *29*, 1528–1532. [CrossRef]
47. Liu, X.; Mou, L.; Cui, H.; Lu, Z.; Song, S. Finding decision jumps in text classification. *Neurocomputing* **2020**, *371*, 177–187. [CrossRef]
48. Wang, Y.; Hei, C.; Liu, H.; Zhang, S.; Wang, J. Prognostics of Remaining Useful Life for Lithium-Ion Batteries Based on Hybrid Approach of Linear Pattern Extraction and Nonlinear Relationship Mining. *IEEE Trans. Power Electron.* **2023**, *38*, 1054–1063. [CrossRef]
49. Forman, G. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* **2003**, *3*, 1289–1305.
50. Ameer, I.; Sidorov, G.; Gómez-Adorno, H.; Nawab, R.M.A. Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods. *IEEE Access* **2022**, *10*, 8779–8789. [CrossRef]
51. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781. [CrossRef].
52. Bilski, J.; Smoląg, J.; Kowalczyk, B.; Grzanek, K.; Izonin, I. Fast Computational Approach to the Levenberg-Marquardt Algorithm for Training Feedforward Neural Networks. *J. Artif. Intell. Soft Comput. Res.* **2023**, *13*, 45–61. [CrossRef]