mathematics

MDPI

Article

Multi-Granularity Invariant Structure Learning for Text Classification in Entrepreneurship Policy

Xinyu Sun [1,2,*] and Meifang Yao [1]

[1] School of Business and Management, Jilin University, Changchun 130012, China; yaomf@jlu.edu.cn
[2] Scientific Research Institute, Changchun Finance College, Changchun 130022, China
* Correspondence: sunxy17@mails.jlu.edu.cn

Abstract

Data-driven text classification technology is crucial for understanding and managing a large number of entrepreneurial policy-related texts, yet it is hindered by two primary challenges. First, the intricate, multi-faceted nature of policy documents often leads to insufficient information extraction, as existing models struggle to synergistically leverage diverse information types, such as statistical regularities, linguistic structures, and external factual knowledge, resulting in semantic sparsity. Second, the performance of state-of-the-art deep learning models is heavily reliant on large-scale annotated data, a resource that is scarce and costly to acquire in entrepreneurial policy domains, rendering models susceptible to overfitting and poor generalization. To address these challenges, this paper proposes a Multi-granularity Invariant Structure Learning (MISL) model. Specifically, MISL first employs a multi-view feature engineering module that constructs and fuses distinct statistical, linguistic, and knowledge graphs to generate a comprehensive and rich semantic representation, thereby alleviating semantic sparsity. Furthermore, to enhance robustness and generalization from limited data, we introduce a dual invariant structure learning framework. This framework operates at two levels: (1) sample-invariant representation learning uses data augmentation and mutual information maximization to learn the essential semantic core of a text, invariant to superficial perturbations; (2) neighborhood-invariant semantic learning applies a contrastive objective on a nearest-neighbor graph to enforce intra-class compactness and inter-class separability in the feature space. Extensive experiments demonstrate that our proposed MISL model significantly outperforms state-of-the-art baselines, proving its effectiveness and robustness for classifying complex texts in entrepreneurial policy domains.

Keywords: text classification; multi-granularity invariant learning; multi-view feature engineering

MSC: 68U15; 68T01

check for updates

Academic Editor: Jonathan Blackledge

Received: 14 October 2025
Revised: 4 November 2025
Accepted: 10 November 2025
Published: 14 November 2025

Citation: Sun, X.; Yao, M. Multi-Granularity Invariant Structure Learning for Text Classification in Entrepreneurship Policy. *Mathematics* **2025**, *13*, 3648. https://doi.org/10.3390/math13223648

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In an increasingly innovation-driven global economy, governments worldwide have enacted a vast number of policies to foster entrepreneurial ecosystems, aiming to promote economic growth and employment [1–5]. These policies, ranging from macro-level regulations to specific financial subsidies and tax incentives, have generated a massive volume of complex and fragmented policy texts. This information explosion presents significant

challenges for policymakers, researchers, and entrepreneurs, including information overload, difficulties in policy evaluation and monitoring, and a lack of transparency [6,7]. To address these challenges, the application of Artificial Intelligence (AI) and Natural Language Processing (NLP) for entrepreneurship policy text classification has emerged [8,9]. This technology automates the process of categorizing immense volumes of unstructured texts, such as policy documents, evaluation reports, press releases, and public feedback, into predefined categories, for instance, by policy instrument, target group, or policy objective. This greatly enhances the efficiency and accuracy of policy analysis. It not only provides data-driven support for evidence-based policy evaluation and optimization but also increases policy transparency. By enabling entrepreneurs and research institutions to more easily access and understand relevant information, it offers crucial technical support for building a healthy and efficient entrepreneurial ecosystem [10,11].

Current data-driven text classification methods are primarily divided into two main categories [12–15]. The first is based on traditional machine learning, which represents a classic paradigm in the field [16]. Its core process involves transforming unstructured text into machine-readable numerical vectors through handcrafted feature engineering [17]. This approach typically represents text using either the Bag-of-Words model, which treats a document as an unordered collection of word frequencies often weighted by the TF-IDF algorithm, or the N-gram model, which preserves local word order by capturing sequences of consecutive words. Once the text is vectorized, these feature vectors are fed into supervised learning classifiers such as Naive Bayes, Support Vector Machine, or Logistic Regression for training, enabling the model to learn the mapping between text and its category [18,19]. The effectiveness of this class of methods is heavily dependent on the quality of the feature engineering [20].

The second category is composed of methods based on deep learning, which represent the state of the art in text classification technology [21,22]. Their primary advantage lies in the ability to automatically extract high-level semantic features from text through an end-to-end learning process, thus circumventing the need for cumbersome and experience-dependent handcrafted feature engineering [23,24]. These methods begin by mapping words into low-dimensional, dense vectors using word embedding techniques. Subsequently, different neural network architectures are employed to capture textual features: Convolutional neural networks efficiently extract key local phrase information through convolution and pooling operations [25]; long short-term memory networks, a variant of recurrent neural networks, excel at processing text sequences to capture long-range contextual dependencies via their gating mechanisms [26]; and architectures represented by the Transformer model leverage a core self-attention mechanism to process text in parallel, directly calculating the relational strength between words to build exceptionally powerful contextual representations, achieving breakthrough results in virtually all text classification tasks [27–30].

Despite the superior performance of text classification methods, two critical issues remain, limiting their practical application in the entrepreneurship policy domain. First, the intricate and multi-faceted nature of policy texts often leads to insufficient information extraction. Existing models frequently struggle to synergistically leverage the diverse types of information embedded in the text—such as statistical regularities, linguistic structures, and connections to external factual knowledge—resulting in a superficial semantic understanding and failing to overcome the semantic sparsity prevalent in this field. Second, the performance of state-of-the-art deep learning models is heavily reliant on large-scale annotated datasets for training. In specialized domains like entrepreneurship policy, acquiring such labeled data is a major bottleneck, as it is both labor-intensive and requires domain expertise. This dependency on limited labeled data renders models highly susceptible to

overfitting, which severely impairs their ability to generalize to new and unseen policy documents in real-world applications.

To this end, a Multi-granularity Invariant Structure Learning (MISL) model is proposed for text classification in entrepreneurship policy. Specifically, our approach comprises three core components designed to systematically address the aforementioned challenges. First, we introduce a multi-view feature engineering module to tackle the issue of insufficient information extraction. This module constructs three distinct graphs to capture semantics from different perspectives: a statistical graph to model word co-occurrence patterns, a linguistic graph to capture syntactic roles and resolve ambiguity, and a knowledge graph to integrate external factual information. By fusing these heterogeneous views, the model generates a comprehensive and rich representation of the policy text, effectively alleviating semantic sparsity. Second, to mitigate the problem of overfitting on limited labeled data, we employ a sample-invariant representation learning strategy. This is achieved through a data augmentation process where we generate varied instances of the original text. The model is then trained to maximize the mutual information between the representations of the original text and its augmented versions. This forces the model to learn the essential and stable semantic core of the text, thereby enhancing its robustness and generalization capabilities. Finally, the model further refines these representations through neighborhood-invariant semantic learning. By constructing a nearest-neighbor graph and applying a contrastive objective, this component encourages intra-class compactness and inter-class separability in the feature space. It pulls semantically similar policy texts closer together while pushing dissimilar ones apart, which significantly improves the discriminative power of the final representations for more accurate classification.

The main contributions are threefold:

- We propose a novel multi-view feature engineering module that integrates statistical, linguistic, and knowledge-based representations through distinct graph structures. This approach effectively alleviates semantic sparsity in policy texts by capturing comprehensive contextual information from multiple perspectives.
- We introduce a sample-invariant representation learning strategy that leverages data augmentation and mutual information maximization. This method enhances the model's robustness and generalization capabilities by focusing on the essential and stable semantic core of the text, even with limited labeled data.
- We develop a neighborhood-invariant semantic learning component that constructs a nearest-neighbor graph and applies a contrastive objective. This technique improves the discriminative power of the final representations by promoting intra-class compactness and inter-class separability, leading to more accurate text classification.

2. The Proposed Method

Given a text document D composed of a sequence of L tokens, where the document belongs to one of C predefined categories, the goal of the text classification task aims to learn a mapping function $f(\cdot)$ to predict the corresponding category label $y \in \{1, 2, \ldots, C\}$, i.e., $f : D \rightarrow y$. To achieve the above objective, this paper proposes a multi-granularity invariant structure learning method (MISL) for mining short text patterns, which contains the multi-view feature engineering, the sample-invariant representation learning, and the neighborhood-invariant semantic learning, as shown in Figure 1.

2.1. Multi-View Feature Engineering

Multi-view feature engineering aims to synergistically leverage the intrinsic statistical regularities and linguistic structures of the text as well as the factual knowledge provided by external knowledge bases, to alleviate semantic sparsity prevalent in texts. To achieve

this objective, we design and construct three distinct graph structures to capture and model each of these information types, thereby generating a richer contextual representation for short texts.

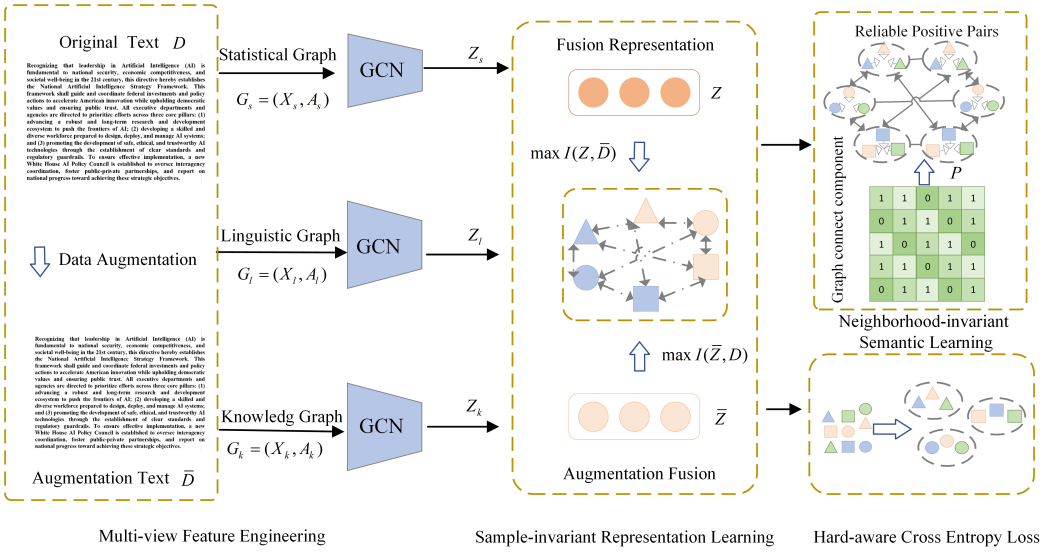


Figure 1. The illustration of the proposed method.

Statistical representation. The statistical regularities of a text reflect the distributional properties and co-occurrence patterns of its vocabulary. To this end, we construct a statistical graph $G_s = \{V_s, X_s, A_s\}$ to explicitly model this type of information. In this graph, V_s represents the set of word nodes. The connection strength between nodes is defined by an adjacency matrix $A_s \in \mathbb{R}^{|V_s| \times |V_s|}$, where weights are calculated using Pointwise Mutual Information (PMI) to quantify the co-occurrence probability of word pairs:

$$A_{s,ij} = \max(\text{PMI}(v_i, v_j), 0), \quad v_i, v_j \in V_s \tag{1}$$

For node features X_s, we use GloVe word vectors for initialization. Then, we input the statistical graph in a two-layer graph convolutional networks $GCN(\cdot)$ to learn statistical representations Z_s:

$$Z_s = GCN(GCN(X_s, A_s), A_s) \tag{2}$$

Linguistic representation. Linguistic structure, particularly syntactic information, is crucial for accurately interpreting the meaning of a text. We focus on utilizing Part-of-Speech (POS) tagging to parse the syntactic roles of words (e.g., distinguishing adjectives from adverbs), thereby resolving lexical ambiguity. To this end, we construct a linguistic graph $G_l = \{V_l, X_l, A_l\}$, where V_l is the set of nodes for all unique POS tags. The adjacency matrix A_l is also computed based on PMI to model the co-occurrence patterns of different POS tags:

$$A_{p,ij} = \max(\text{PMI}(v_i, v_j), 0), \quad v_i, v_j \in V_l \tag{3}$$

The initial feature for each POS node, X_l, is set as a one-hot vector. Similar to the statistical representation, we apply a two-layer graph convolutional networks to generate linguistic representation Z_l.

$$Z_l = GCN(GCN(X_l, A_l), A_l) \tag{4}$$

Knowledge representation. External factual knowledge can supplement short texts with necessary background information, significantly enhancing the performance of classification models. To incorporate this knowledge, we first identify and link entities from

the text to a knowledge graph, and then construct a knowledge graph $G_k = \{V_k, X_k, A_k\}$. In the implementation, we use the TAGME toolkit to perform entity linking against the NELL knowledge base, where V_k is the resulting set of entity nodes. The initial embeddings for these entities, X_k, are generated using the TransE model, which is adept at learning entity relationships in a knowledge graph. The adjacency matrix A_k is derived from the cosine similarity between entity embedding pairs to capture semantic proximity:

$$A_{k,ij} = \max(\cos(X_{k,i}, X_{k,j}), 0)$$

Likewise, we apply a two-layer graph convolutional networks to generate linguistic representation Z_k.

$$Z_k = GCN(GCN(X_k, A_k), A_k) \tag{5}$$

Based on the above three representations, a multi-view fusion function $Fusion(\cdot)$ is designed to aggregate complementary information:

$$Z = Fusion(Z_s, Z_l, Z_k) \tag{6}$$

where Z denotes the fusion representation of the text. In the experiments, $Fusion(\cdot)$ is implemented as the concatenation operations of three representations.

2.2. Sample-Invariant Representation Learning

Sample-invariant representation learning aims to utilize invariant structure hidden in each text to boost the discriminative ability of representations. To accomplish this, we introduce the data augmentation strategy to generate augmentation text and then maximize mutual information between the augmentation text and original text to endow representations with sample-invariant structures.

Specifically, we define a text data augmentation strategy (Aug) to generate augmentation text:

$$\begin{aligned} Aug^a(D, \alpha), Aug^b(D, \beta), Aug^c(D, \gamma) \sim Aug \\ \bar{D} = Aug^j(D), j \in \{a, b, c\} \end{aligned} \tag{7}$$

where Aug^a, Aug^b, and Aug^c denote three augmentation methods, i.e., back translation, random deletion, and synonym replacement. α, β, and γ correspond to the back translation rate, random deletion rate, and synonym replacement rate, respectively. $\bar{D}$ denotes the text augmentation of the text D. Then, through the multi-view feature engineering, the augmentation fusion representation $\bar{Z}$ is obtained.

In general, data augmentation should not alter the semantic information of the text. That is, the original text and augmentation text should have consistent semantics. Thus, the mutual information between original fusion representations and augmentation text is maximized to achieve the above objective:

$$max I(Z, \bar{D}) \tag{8}$$

where $I(,)$ denotes the mutual information function. Based on the definition of mutual information, the following holds:

$$I(Z, \bar{D}) = \iint p(Z \mid \bar{D}) p(\bar{D}) \log \frac{p(Z \mid \bar{D})}{p(Z)} \, d\bar{D} \, dZ = \mathrm{KL}(p(Z \mid \bar{D}) p(\bar{D}) \parallel p(Z) p(\bar{D})) \tag{9}$$

where $p(\bar{D})$ denotes the marginal distribution of the augmented text, $p(Z)$ is the marginal distribution of the representations, and $p(Z \mid \bar{D})$ is the conditional distribution of Z given $\bar{D}$. $\mathrm{KL}(\cdot)$ denotes the Kullback–Leibler divergence.

However, the direct optimization of a KL divergence-based objective presents a significant challenge in the context of deep learning. The unbounded nature of the KL divergence can lead to unstable training dynamics, where the loss can grow indefinitely, causing exploding gradients. To mitigate this, we seek a more robust and well-behaved alternative. The Jensen–Shannon (JS) divergence is an ideal candidate because it is both symmetric and, crucially, bounded within the range $[0, \log 2]$. By substituting JS divergence for KL divergence, we reformulate our objective as follows:

$$maxI(Z, \bar{D}) = \mathrm{JS}(p(Z \mid \bar{D})p(\bar{D}) \parallel p(Z)p(\bar{D})) \tag{10}$$

While theoretically appealing, the direct computation of JS divergence is often intractable, as it requires knowledge of the distribution $p(Z)$, which is typically unknown. To overcome this, we employ a variational estimation approach. The variational lower bound of the JS divergence between two distributions, $p(X)$ and $q(X)$, is defined as:

$$\mathrm{JS}(p(X) \parallel q(X)) \geq \sup_{T}(\mathbb{E}_{x\sim p(X)}[\log \sigma(T(x))] + \mathbb{E}_{x\sim q(X)}[\log(1 - \sigma(T(x)))]) \tag{11}$$

where $T(\cdot)$ is a function, typically a neural network, that acts as a discriminator, and $\sigma(\cdot)$ is the sigmoid function. $\mathbb{E}$ denotes the expectation operator. This framework effectively reframes the problem of density ratio estimation as a more tractable binary classification task. Then, we replace $p(X)$ with $p(Z \mid \bar{D})p(\bar{D})$ and $q(X)$ with $p(Z)p(\bar{D})$, and we obtain:

$$\begin{aligned} maxI(Z, \bar{D}) = {} & \mathbb{E}_{(\bar{D},Z)\sim p(Z|\bar{D})p(\bar{D})}[\log \sigma(T(\bar{D}, Z))] \\ & + \mathbb{E}_{(\bar{D},Z)\sim p(Z)p(\bar{D})}[\log(1 - \sigma(T(\bar{D}, Z)))]) \end{aligned} \tag{12}$$

In practice, the expectation over the product of marginals is approximated using negative sampling. We recast an augmentation text $\bar{D}$ and fusion representations Z of the corresponding original text D as a positive pair. We recast the same augmentation text $\bar{D}$ and fusion representations Z of the other original text. By training the discriminator to distinguish these pairs, the representation learning model is implicitly forced to maximize the mutual information.

Finally, the loss L_{spl} of sample-invariant representation learning is defined as:

$$L_{spl} = -(maxI(Z, \bar{D}) + I(\bar{Z}, D)) \tag{13}$$

2.3. Neighborhood-Invariant Semantic Learning

Neighborhood-invariant semantic learning aims to bring similar samples with high confidence closer together and keep dissimilar samples with high confidence away to capture text patterns with intra-class compactness and inter-class separability.

Specifically, given fusion representations, we introduce the nearest-neighbor graph G_n via utilizing k-nearest neighbors to select the top k most similar neighbors for each text sample D:

$$G_{n,ij} = \begin{cases} 1 & (D_i \in \varphi(D_j) \text{ or } D_j \in \varphi(D_i)) \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where $\varphi(D_j)$ denotes the nearest set of the j-th text sample D_j. $G_{n,ij}$ denotes the element of the nearest-neighbor graph G_n.

Following the initial identification of nearest neighbors for every sample, a naive method might consider this entire neighbor set as the basis for forming positive pairs. Such a strategy, however, risks contaminating the learning signal with semantically inconsistent pairings, as some neighbors may be incidental rather than truly related. To address this potential for noise, we implement a robust filtering mechanism that consults the pre-computed, stable class assignments. Our model mandates that for a pair of samples to be considered positive, they must simultaneously fulfill two requirements: a local relationship (being nearest neighbors) and a global one (residing in the same class). In contrast, pairs are defined as negative when they exhibit neither a local nor a global semantic connection. By imposing this strict, two-tiered validation, we effectively purify the set of positive and negative pairs, ensuring a more accurate and stable optimization process.

Based on the above analysis, for the nearest-neighbor graph G_n, we identify its disjoint subgraphs as different classes using the connected component labeling algorithm. This algorithm efficiently partitions the set of nodes into components $\{C_1, C_2, \ldots, C_K\}$, where any two nodes within a component C_k are connected by a path. We treat each component as a pseudo-label, assuming that all texts within it are semantically similar. Let $\mathcal{C}(D_i)$ be the function that returns the component ID for text D_i. From this, we derive a binary pseudo-label matrix $P \in \{0,1\}^{N\times N}$ between N text samples:

$$P_{ij} = \begin{cases} 1 & \text{if } \mathcal{C}(D_i) = \mathcal{C}(D_j) \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Here, $P_{ij} = 1$ signifies that texts D_i and D_j belong to the same semantic component. Based on neighborhood structure information and class semantic information, we construct the reliable positive pairs set R_i^p and negative pairs set R_i^n as follows:

$$R_i^p = \{j \mid G_{n,ij} = 1, p_{ij} = 1, \quad \forall j \in [1, N]\}, \tag{16}$$

$$R_i^n = \{j \mid G_{n,ij} = 0, p_{ij} = 0, \quad \forall j \in [1, N]\}, \tag{17}$$

Then, the loss L_{nsl} of the neighborhood-invariant semantic learning is defined as:

$$L_{nsl} = -\log \frac{\sum_{j\in R_i^p} \exp(sim(z_i, \bar{z}_j))}{\sum_{j\in R_i^n} \exp(sim(z_i, \bar{z}_j)) + \sum_{j\in R_i^p} \exp(sim(z_i, z_j))} \tag{18}$$

where $\exp(\cdot)$ and $sim(\cdot)$ are the exponential function and the cosine function, respectively.

2.4. The Overall Loss Function

The model is trained end-to-end by optimizing a composite objective that linearly combines the neighborhood-invariant semantic learning loss, the sample-invariant representation learning loss, and the hard-aware cross entropy loss, as follows:

$$L = L_{hce} + \lambda L_{nsl} + \delta L_{srl} \tag{19}$$

where λ and δ are trade-off parameters. L_{hce} is the hard-aware cross entropy loss:

$$L_{hce} = -\frac{1}{N}\sum_{i=1}^{N}(1 - q_{i,y_i})^{\tau} \log(q_{i,y_i}) \tag{20}$$

where q_{i,y_i} is the predicted probability for the true class y_i of the i-th text sample. The core idea is to dynamically scale the cross-entropy loss with a weighting factor τ that diminishes the contribution of easy samples and, in turn, magnifies the focus on hard-to-classify

samples. q_i can be obtained via inputting the fusion representation Z_i into a linear classifier. The detailed training process is shown in Algorithm 1.

Algorithm 1 Algorithm for Multi-granularity Invariant Structure Learning (MISL)

1: /* Training Phase */
2: **Input:** Training dataset $D = \{(D_i, y_i)\}_{i=1}^{N}$; the category number C; loss weights λ, δ.
3: **Initialize:** Initialize the parameters θ of MISL.
4: **while** not converged **do**
5: Sample a batch of text documents $\{(D_i, y_i)\}$ from D.
6: **for** each document D_i in the batch **do**
7: Construct statistical graph G_s, linguistic graph G_l, and knowledge graph G_k.
8: Obtain view-specific representations Z_s, Z_l, Z_k.
9: Fuse representations to get the final representation $Z_i = \text{Fusion}(Z_s, Z_l, Z_k)$.
10: Generate an augmented view $\bar{D}_i = \text{Aug}(D_i)$.
11: Obtain the augmentation representation $\bar{Z}_i$ using the same process as lines 7–9.
12: **end for**
13: Calculate the hard-aware cross-entropy loss L_{hce}.
14: Calculate the sample-invariant representation learning loss L_{spl}.
15: calculate the neighborhood-invariant semantic learning L_{nsl}.
16: Compute the total loss: $L = L_{hce} + \lambda L_{nsl} + \delta L_{spl}$.
17: Update model parameters θ using gradient descent.
18: **end while**
19: **Output:** The model parameter θ.
20: /* Inference Phase */
21: **Input:** A test document D_{test}.
22: Apply the trained model with parameter θ on D_{test} to obtain the corresponding label predictions.
23: **Output:** Label predictions of D_{test}.

3. Experimental Evaluation

3.1. Setup

Datasets. Three text datasets are leveraged to test the performance of MISL, i.e., Ohsumed, TagMyNews, and Snippets. Following current works, a small pool of 40 labeled documents is randomly selected for each class, and then this pool is evenly partitioned, yielding 20 documents per class for the training set and another 20 for the validation set. The majority of the remaining data is designated as the test set. Specifically, the Ohsumed dataset includes 7400 documents, with 460 for training (6.22%), 11,764 words, 4507 entities, 38 tags, 23 classes. The TagMyNews dataset is the most extensive, with 32,549 documents, 140 for training (0.43%), 38,629 words, 14,734 entities, 42 tags, and 7 classes. The Snippets dataset is larger, with 12,340 documents, 160 for training (1.30%), 29,040 words, 9737 entities, 34 tags, and it covers 8 classes.

Evaluation Metrics: Following [31–33], to evaluate the performance of MISL, Accuracy (ACC), Recall (Rec), Precision (Pre), and F1 are leveraged in the experiments. The larger the values of the four metrics, the better the performance. The experiment results are the average of five times.

Comparison methods. Eleven text classification methods are selected as baselines, including QSIM [7], EMGAN [10], GTC [13], PTE [17], LSTMNN [19], NFS [20], Hy-TC [22], IEG-GAT [24], IGCL [30], BERT-avg [34] and BERT-cls [34].

Implementation details. In the experiments, the Adam optimizer is used to optimize the overall network with a learning rate of 0.0003 and an epoch number of 500. λ and δ are determined by conducting grid search experiments on three datasets. An early stopping mechanism was employed to prevent overfitting. The training process was terminated if the loss on the validation set did not decrease for 10 consecutive epochs, and the model weights from the best-performing epoch were restored for the final evaluation. For Ohsumed, TagMyNews, and Snippets, λ is set to 1, 1, and 0.1, respectively. δ is set to 1, 1, and 1, respectively.

3.2. Comparison Evaluation

The results of the comparison evaluation, as detailed in Tables 1–3, demonstrate the superior performance of the proposed MISL model across all three datasets: Ohsumed, TagMyNews, and Snippets. In all evaluated metrics, our model consistently outperforms eleven baseline methods, often by a significant margin. These consistent and substantial gains across diverse datasets underscore the robustness and effectiveness of the proposed approach.

The remarkable performance of our proposed model can be attributed to several key methodological innovations designed to address the specific challenges of text classification in entrepreneurship policy. Firstly, the novel multi-view feature engineering module effectively mitigates semantic sparsity by integrating statistical, linguistic, and knowledge-based representations. This holistic approach captures a more comprehensive contextual understanding of the policy texts. Secondly, the sample-invariant representation learning strategy, which leverages data augmentation and mutual information maximization, enhances the model's robustness and generalization capabilities, particularly in scenarios with limited labeled data. By forcing the model to learn the essential semantic core of the text, it becomes less susceptible to superficial variations. Lastly, the neighborhood-invariant semantic learning component improves the discriminative power of the final representations by promoting intra-class compactness and inter-class separability, leading to more accurate classification.

Meanwhile, to validate the statistical significance of our results, we performed paired t-tests between our proposed MISL model and each baseline method for every evaluation metric across all datasets. The tests were conducted using the results from five independent runs with different random seeds. Significance levels are indicated in the tables: ** denotes $p < 0.01$ when comparing baseline methods against our proposed model, confirming that all observed improvements are statistically meaningful.

Table 1. Results (%) of four metrics on the Ohsumed dataset. The average results with standard deviation over five random runs are reported. Bold indicates the best result. The symbols ** indicate that the performance is significantly worse than the proposed model (Ours) at $p < 0.01$ levels based on paired *t*-tests.

Method	Accuracy	Recall	Precision	F1
QSIM	44.58 ± 0.26 **	27.75 ± 0.42 **	40.36 ± 0.42 **	28.67 ± 0.33 **
EMGAN	40.12 ± 3.12 **	23.31 ± 3.36 **	37.74 ± 3.02 **	25.14 ± 2.64 **
GTC	36.95 ± 2.38 **	20.78 ± 1.17 **	35.54 ± 1.85 **	23.57 ± 1.82 **
PTE	37.58 ± 3.87 **	18.74 ± 2.72 **	32.00 ± 2.97 **	16.40 ± 2.63 **
LSTMNN	38.74 ± 2.71 **	20.12 ± 1.11 **	34.44 ± 2.04 **	18.07 ± 2.47 **
NFS	39.95 ± 1.22 **	21.45 ± 1.03 **	35.99 ± 1.77 **	20.47 ± 0.99 **
BERT-avg	42.68 ± 1.17 **	24.82 ± 1.30 **	37.26 ± 1.41 **	25.89 ± 1.38 **
BERT-cls	43.24 ± 0.87 **	25.69 ± 0.92 **	37.75 ± 1.03 **	26.46 ± 0.97 **
Hy-TC	40.07 ± 0.77 **	23.11 ± 0.65 **	35.21 ± 0.73 **	26.89 ± 0.46 **
IEG-GAT	43.24 ± 0.54 **	25.86 ± 0.47 **	37.77 ± 0.68 **	27.34 ± 0.52 **
IGCL	42.92 ± 1.25 **	25.64 ± 1.12 **	38.10 ± 1.78 **	27.69 ± 0.73 **
Ours	**47.23 ± 0.57**	**29.23 ± 0.72**	**43.54 ± 0.69**	**30.72 ± 0.38**

Table 2. Results (%) of four metrics on the TagMyNews dataset. The average results with standard deviation over five random runs are reported. Bold indicates the best result. The symbols ** indicate that the performance is significantly worse than the proposed model (Ours) at $p < 0.01$ levels based on paired *t*-tests.

Method	Accuracy	Recall	Precision	F1
QSIM	56.71 ± 0.31 **	47.77 ± 0.39 **	54.31 ± 0.53 **	50.71 ± 0.25 **
EMGAN	50.24 ± 2.74 **	45.36 ± 2.57 **	50.39 ± 2.68 **	46.77 ± 1.90 **
GTC	43.75 ± 2.21 **	25.41 ± 1.33 **	40.12 ± 2.05 **	26.11 ± 1.27 **
PTE	46.12 ± 2.37 **	27.67 ± 2.42 **	43.42 ± 2.28 **	29.12 ± 2.34 **
LSTMNN	46.33 ± 2.22 **	28.01 ± 1.04 **	44.07 ± 1.90 **	30.32 ± 1.72 **
NFS	50.12 ± 0.94 **	33.74 ± 0.87 **	49.04 ± 1.41 **	33.74 ± 0.81 **
BERT-avg	54.84 ± 0.89 **	44.78 ± 1.23 **	52.23 ± 0.77 **	48.17 ± 1.54 **
BERT-cls	52.11 ± 1.65 **	43.75 ± 1.58 **	50.92 ± 1.66 **	46.36 ± 1.86 **
Hy-TC	53.00 ± 0.54 **	44.22 ± 0.44 **	51.06 ± 0.44 **	47.99 ± 0.37 **
IEG-GAT	56.84 ± 0.70 **	48.03 ± 0.68 **	53.49 ± 0.93 **	51.03 ± 0.60 **
IGCL	55.27 ± 0.96 **	46.75 ± 0.85 **	52.33 ± 1.02 **	50.70 ± 0.57 **
Ours	**60.74 ± 0.44**	**51.47 ± 0.37**	**58.74 ± 0.69**	**55.88 ± 0.62**

Table 3. Results (%) of four metrics on the Snippets dataset. Bold indicates the best result. The average results with standard deviation over five random runs are reported. The symbols ** indicate that the performance is significantly worse than the proposed model (Ours) at $p < 0.01$ levels based on paired *t*-tests.

Method	Accuracy	Recall	Precision	F1
QSIM	75.12 ± 0.54 **	76.09 ± 0.67 **	75.86 ± 0.60 **	75.32 ± 0.57 **
EMGAN	69.69 ± 1.74 **	68.57 ± 2.02 **	69.22 ± 2.14 **	68.25 ± 1.87 **
GTC	60.12 ± 2.44 **	62.12 ± 2.24 **	59.96 ± 1.69 **	60.57 ± 1.93 **
PTE	56.62 ± 2.20 **	55.22 ± 2.75 **	56.12 ± 2.41 **	56.37 ± 2.23 **
LSTMNN	58.74 ± 2.36 **	58.75 ± 2.44 **	54.94 ± 2.35 **	58.07 ± 2.23 **
NFS	67.39 ± 1.47 **	67.21 ± 1.45 **	67.35 ± 1.99 **	67.20 ± 1.47 **
BERT-avg	70.74 ± 0.88 **	71.04 ± 0.67 **	70.42 ± 1.25 **	71.07 ± 0.93 **
BERT-cls	69.25 ± 0.85 **	70.07 ± 0.83 **	69.37 ± 1.34 **	70.28 ± 0.76 **
Hy-TC	68.55 ± 0.92 **	69.97 ± 0.74 **	68.82 ± 0.89 **	69.89 ± 0.86 **
IEG-GAT	73.24 ± 0.64 **	74.86 ± 0.89 **	75.49 ± 0.84 **	75.27 ± 0.93 **
IGCL	72.20 ± 1.75 **	72.95 ± 0.68 **	71.19 ± 1.44 **	72.77 ± 1.36 **
Ours	**79.75 ± 0.54**	**78.35 ± 0.88**	**78.62 ± 0.76**	**78.38 ± 0.75**

3.3. Ablation Analysis

An ablation study is conducted to validate the contribution of each representation within our multi-view feature engineering module: statistical (Z_s), linguistic (Z_l), and knowledge-based (Z_k). The results in Table 4 consistently demonstrate the overall effectiveness of the multi-view approach, as the complete model incorporating all three representations achieved the best performance across all datasets, highlighting a synergistic effect where each view provides complementary and indispensable information. Notably, the removal of the knowledge-based representation resulted in the most substantial decline in performance, underscoring its critical role in resolving semantic sparsity by enriching the text with crucial background and contextual information from external knowledge bases. Eliminating the statistical representation also led to a significant performance degradation, confirming that information derived from word co-occurrence patterns serves as the fundamental building block for capturing the core topics of the text. Finally, while the removal of the linguistic representation had a comparatively lesser impact, it still caused a noticeable decrease in performance, suggesting its value as an auxiliary source of fine-grained semantic clues. In summary, the analysis reveals a clear hierarchy of importance for the representations and confirms that their effective integration is key to the model's superior performance.

Meanwhile, to assess the contribution of each component within our composite objective, we performed an ablation study on the loss functions, with results presented in Table 5. The results unequivocally establish the hard-aware cross-entropy loss (L_{hce}) as the foundational component; its removal led to a catastrophic collapse in performance, as it

provides the primary supervised signal essential for the classification task. Furthermore, both the neighborhood-invariant (L_{nsl}) and sample-invariant (L_{srl}) learning losses proved to be critically important. Removing the neighborhood-invariant loss significantly weakened the discriminative power of the feature space, while removing the sample-invariant loss diminished the model's robustness and ability to generalize by learning the core semantics of the text. The superior performance of the complete model over any of its ablated variants strongly demonstrates a powerful synergy, confirming that the integration of a core classification objective with these two distinct representation learning objectives is essential for achieving a highly accurate and robust final model.

Table 4. View representation ablation on three datasets in terms of Accuracy and F1 (%). The average results with standard deviation over five random runs are reported. Bold indicates the best result.

Dataset	**Ohsumed**		**TagMyNews**		**Snippets**	
Metric	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
Z w/o Z_s	45.24 ± 0.74	29.57 ± 0.87	58.57 ± 1.11	54.64 ± 1.44	77.46 ± 2.45	76.56 ± 1.92
Z w/o Z_l	46.37 ± 0.85	29.75 ± 0.69	58.28 ± 0.82	55.55 ± 0.73	78.08 ± 0.76	77.89 ± 0.89
Z w/o Z_k	45.66 ± 0.43	28.89 ± 0.35	58.36 ± 0.38	53.92 ± 0.42	77.67 ± 0.59	76.43 ± 0.43
Ours	$\mathbf{47.23 \pm 0.57}$	$\mathbf{30.72 \pm 0.38}$	$\mathbf{60.74 \pm 0.44}$	$\mathbf{55.88 \pm 0.62}$	$\mathbf{79.75 \pm 0.54}$	$\mathbf{78.38 \pm 0.75}$

Table 5. Loss ablation on three datasets in terms of Accuracy and F1 (%). The average results with standard deviation over five random runs are reported. Bold indicates the best result.

Dataset	**Ohsumed**		**TagMyNews**		**Snippets**	
Metric	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
L w/o L_{hce}	34.12 ± 1.12	17.02 ± 1.46	42.39 ± 1.87	35.18 ± 2.01	57.57 ± 1.78	56.25 ± 1.53
L w/o L_{nsl}	44.96 ± 0.73	27.49 ± 0.66	57.82 ± 0.83	53.74 ± 0.97	76.55 ± 0.65	75.29 ± 0.72
L w/o L_{srl}	43.66 ± 0.84	27.07 ± 0.76	56.57 ± 0.59	52.62 ± 0.55	75.33 ± 1.22	75.03 ± 1.36
Ours	$\mathbf{47.23 \pm 0.57}$	$\mathbf{30.72 \pm 0.38}$	$\mathbf{60.74 \pm 0.44}$	$\mathbf{55.88 \pm 0.62}$	$\mathbf{79.75 \pm 0.54}$	$\mathbf{78.38 \pm 0.75}$

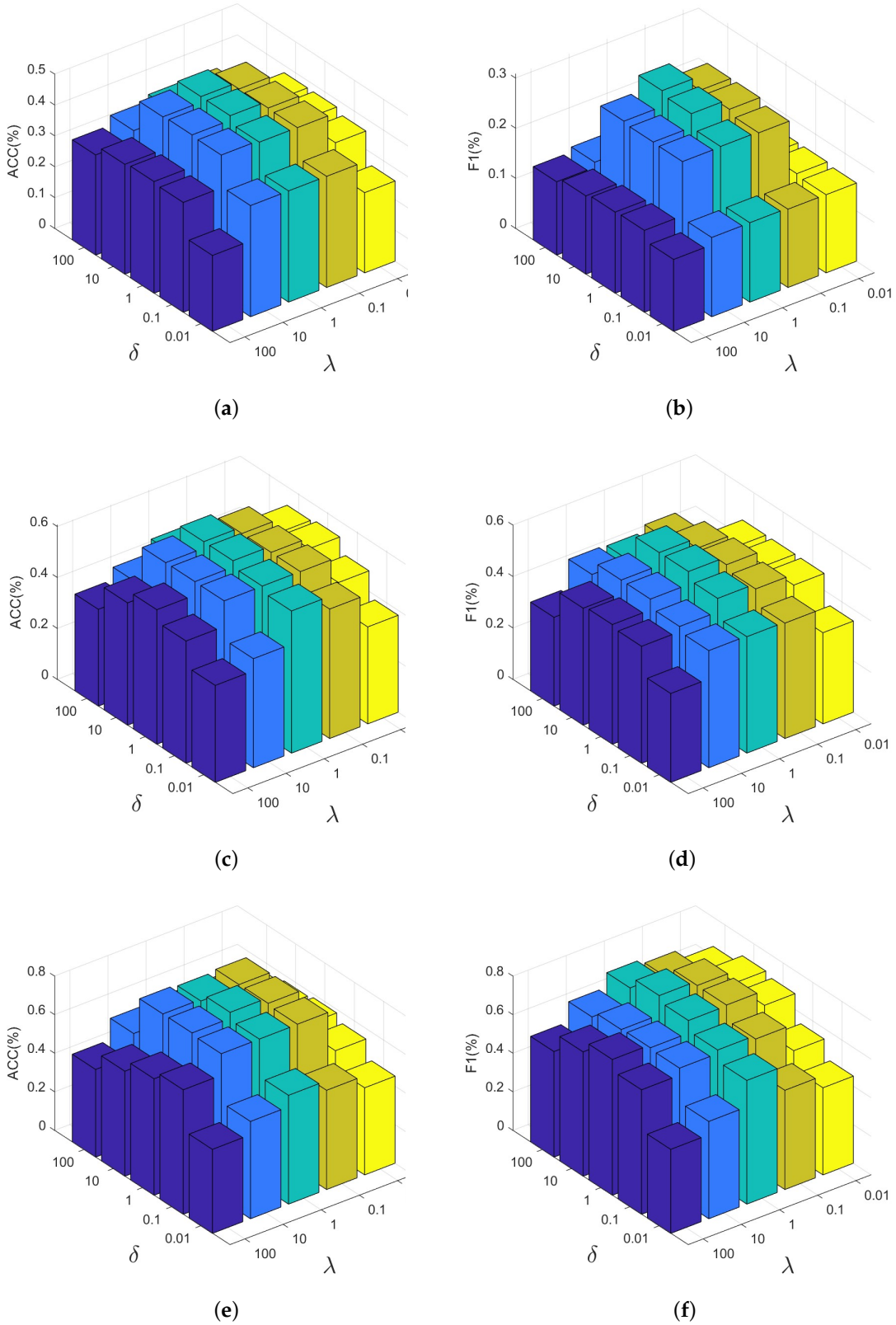
3.4. Parameter Analysis

λ and δ act as weighting factors, controlling the influence of the neighborhood-invariant and sample-invariant losses, respectively. To determine the optimal values for these hyperparameters, a comprehensive grid search is conducted, which explores a range of values for both λ and δ, specifically from the set $\{100, 10, 1, 0.1, 0.01\}$. This process allows for a thorough evaluation of how different balances between the loss components affect the model's ability to classify text accurately. The results in Figure 2 demonstrate a consistent and telling pattern across all three datasets: Ohsumed, TagMyNews, and Snippets. The model's performance, measured by ACC and F1, consistently peak when the values for λ and δ were set within the range of $[0.1, 1]$. This phenomenon indicates that the MISL model is most effective when the contributions of the neighborhood-invariant and sample-invariant learning components are carefully balanced with the primary cross-entropy loss. When the weights are too high (e.g., 10 or 100), these specialized loss components may dominate the learning process, potentially overshadowing the fundamental classification task. Conversely, when the weights are too low (e.g., 0.01), their regularizing and structure-enforcing benefits are diminished, leading to suboptimal performance.

3.5. Fusion Analysis

To determine the most effective method for integrating the multi-view representations, we evaluate four distinct fusion operations: concatenation, average, weighting, and sum. The ACC of each method was tested on the Ohsumed, TagMyNews, and Snippets datasets. The results in Figure 3 consistently show that the concatenate and weighting operations are highly competitive, both significantly outperforming the average and sum methods across all three datasets. While the weighting mechanism yields strong results, it introduces

additional learnable parameters and computational overhead. Given that the concatenate operation achieves comparable or superior performance without this extra burden, it was selected as the fusion strategy for our model due to its optimal balance of effectiveness and efficiency.



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 2. The parameter analysis on three datasets. (**a**) Ohsumed (ACC). (**b**) Ohsumed (F1). (**c**) TagMyNews (ACC). (**d**) TagMyNews (F1). (**e**) Snippets (ACC). (**f**) Snippets (F1).

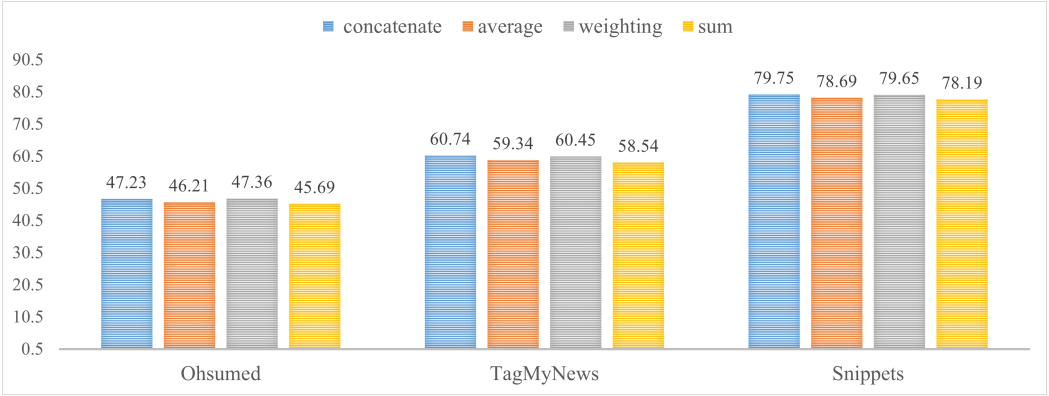


Figure 3. Four fusion operation analysis on three datasets.

3.6. Model Analysis

Based on the comprehensive model analysis in Table 6, the proposed MISL framework demonstrates significant advantages in computational efficiency and practical deployment potential. The model achieves exceptional parameter efficiency, requiring substantially fewer parameters while maintaining competitive performance. The results conclusively show that our method successfully navigates the trade-off between model complexity and operational efficiency, offering a practical and effective approach for text classification tasks.

Table 6. Model analysis in terms of parameter, inference time and FLOPS on the Ohsumed dataset.

Metric	QSIM	LSTMNN	BERT-avg	Ours
Parameter	81.65 M	564 K	110 M	779 K
Inference time	0.85 s	0.15 s	1.19 s	0.18 s
FLOPS	15.39	3.41	11.4	5.39

4. Conclusions

This paper addresses two core challenges in the text classification of entrepreneurship policy: the issue of semantic sparsity arising from the intrinsic complexity of policy texts, and the problems of model overfitting and poor generalization caused by the scarcity of labeled data in this specialized domain. To overcome these challenges, we propose an innovative multi-granularity invariant structure learning model. To combat semantic sparsity, the model first introduces a multi-view feature engineering module. This module constructs and fuses three distinct graph structures—statistical, linguistic, and knowledge-based—to capture textual information from multiple dimensions, thereby generating a comprehensive and rich semantic representation. To resolve the issue of model robustness in the context of data scarcity, we introduce a dual invariant structure learning framework. This framework operates on two levels: first, a sample-invariant representation learning that uses data augmentation and mutual information maximization to help the model learn the essential and stable semantic core of a text, making it insensitive to superficial perturbations. Second, a neighborhood-invariant semantic learning that applies a contrastive objective on a nearest-neighbor graph to enhance intra-class compactness and inter-class separability in the feature space. In summary, through its meticulously designed multi-view feature fusion and dual invariant learning mechanisms, the MISL model systematically solves the key difficulties in entrepreneurship policy text classification. It provides a powerful solution for achieving efficient, robust, and accurate classification of complex texts, even with limited data. While MISL demonstrates strong performance, we acknowledge several limitations that present opportunities for future work. First, the model's knowledge representation module relies on external knowledge bases. Consequently, its effectiveness can be constrained by the coverage and quality of these bases, potentially leading to information

loss for entities or concepts not well-represented in them. Second, the construction of multiple graph views and the subsequent graph convolutional operations introduce higher computational complexity compared to simpler text classification models, which may limit its application in scenarios requiring strict real-time processing.

Author Contributions: Conceptualization, X.S.; methodology, X.S. and M.Y.; formal analysis, X.S. and M.Y.; writing—original draft preparation, X.S.; writing—review and editing, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by 1. National Natural Science Foundation of China (Grant No. 72091313); 2. Changchun Philosophy and Social Sciences Planning Project (Grant No. CSKT2024ZX—107).

Data Availability Statement: The data presented in this study are openly available on Github at https://github.com/Machinelearning20/MISL, accessed on 1 November 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Allam, H.; Makubvure, L.; Gyamfi, B.; Graham, K.N.; Akinwolere, K. Text classification: How machine learning is revolutionizing text categorization. *Information* **2025**, *16*, 130. [CrossRef]
2. Zhang, D.W.; Mi, R.X.; Zhou, P.Y.; Jin, D.; Zhang, M.; Song, T. Optimization Method for Imbalanced Text Classification Based on Large Models. *Softw. Eng.* **2025**, *28*, 47–50+78. [CrossRef]
3. Wu, Y.; Wan, J. A survey of text classification based on pre-trained language model. *Neurocomputing* **2025**, *616*, 128921. [CrossRef]
4. Gao, N.; Wang, Y.; Chen, P.; Zheng, X. Graph-Guided Multi-view Text Classification: Advanced Solutions for Fast Inference. In *International Conference on Artificial Neural Networks*; Springer: Cham, Switzerland, 2024; pp. 126–142.
5. Gao, J.; Liu, M.; Li, P.; Zhang, J.; Chen, Z. Deep Multiview Adaptive Clustering with Semantic Invariance. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 12965–12978. [CrossRef] [PubMed]
6. Gao, J.; Liu, M.; Li, P.; Laghari, A.A.; Javed, A.R.; Victor, N.; Gadekallu, T.R. Deep Incomplete Multiview Clustering via Information Bottleneck for Pattern Mining of Data in Extreme-environment IoT. *IEEE Internet Things J.* **2023**, *11*, 26700–26712. [CrossRef]
7. Gao, H.; Zhang, P.; Zhang, J.; Yang, C. Qsim: A quantum-inspired hierarchical semantic interaction model for text classification. *Neurocomputing* **2025**, *611*, 128658. [CrossRef]
8. Reusens, M.; Stevens, A.; Tonglet, J.; De Smedt, J.; Verbeke, W.; Vanden Broucke, S.; Baesens, B. Evaluating text classification: A benchmark study. *Expert Syst. Appl.* **2024**, *254*, 124302. [CrossRef]
9. Jamshidi, S.; Mohammadi, M.; Bagheri, S.; Najafabadi, H.E.; Rezvanian, A.; Gheisari, M.; Ghaderzadeh, M.; Shahabi, A.S.; Wu, Z. Effective text classification using BERT, MTM LSTM, and DT. *Data Knowl. Eng.* **2024**, *151*, 102306. [CrossRef]
10. Ai, W.; Wei, Y.; Shao, H.; Shou, Y.; Meng, T.; Li, K. Edge-enhanced minimum-margin graph attention network for short text classification. *Expert Syst. Appl.* **2024**, *251*, 124069. [CrossRef]
11. Sun, G.; Cheng, Y.; Zhang, Z.; Tong, X.; Chai, T. Text classification with improved word embedding and adaptive segmentation. *Expert Syst. Appl.* **2024**, *238*, 121852. [CrossRef]
12. Liu, Y.; Li, M.; Pang, W.; Giunchiglia, F.; Huang, L.; Feng, X.; Guan, R. Boosting Short Text Classification with Multi-Source Information Exploration and Dual-Level Contrastive Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 24696–24704.
13. Li, X.; Wang, B.; Wang, Y.; Wang, M. Graph-based text classification by contrastive learning with text-level graph augmentation. *ACM Trans. Knowl. Discov. Data* **2024**, *18*, 1–21. [CrossRef]
14. Yang, Z. Emmert-Streib F. Optimal performance of Binary Relevance CNN in targeted multi-label text classification. *Knowl. Based Syst.* **2024**, *284*, 111286. [CrossRef]
15. Wei, Y.; Wang, Z.; Li, J.; Li, T. The unsupervised short text classification method based on GCN encoder–decoder and local enhancement. *Expert Syst. Appl.* **2025**, *282*, 127678. [CrossRef]
16. Zhu, Y.; Wang, Y.; Mu, J.; Li, Y.; Qiang, J.; Yuan, Y.; Wu, X. Short text classification with soft knowledgeable prompt-tuning. *Expert Syst. Appl.* **2024**, *246*, 123248. [CrossRef]
17. Tang, J.; Qu, M.; Mei, Q. Pte: Predictive text embedding through large-scale heterogeneous text networks. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 1165–1174.

18. Soni, S.; Chouhan, S.S.; Rathore, S.S. TextConvoNet: A convolutional neural network based architecture for text classification. *Appl. Intell.* **2023**, *53*, 14249–14268. [CrossRef] [PubMed]
19. Pozzi, A.; Incremona, A.; Tessera, D.; Toti, D. Mitigating exposure bias in large language model distillation: An imitation learning approach. *Neural Comput. Appl.* **2025**, *37*, 12013–12029. [CrossRef]
20. Okkalioglu, M. A novel redistribution-based feature selection for text classification. *Expert Syst. Appl.* **2024**, *246*, 123119. [CrossRef]
21. Behzadidoost, R.; Mahan, F.; Izadkhah, H. Granular computing-based deep learning for text classification. *Inf. Sci.* **2024**, *652*, 119746. [CrossRef]
22. Maragheh, H.K.; Gharehchopogh, F.S.; Majidzadeh, K.; Sangar, A.B. A hybrid model based on convolutional neural network and long short-term memory for multi-label text classification. *Neural Process. Lett.* **2024**, *56*, 42. [CrossRef]
23. Cheng, Q.; Shi, W. Hierarchical multi-label text classification of tourism resources using a label-aware dual graph attention network. *Inf. Process. Manag.* **2025**, *62*, 103952. [CrossRef]
24. Liu, H.; Huang, X.; Liu, X. Improve label embedding quality through global sensitive GAT for hierarchical text classification. *Expert Syst. Appl.* **2024**, *238*, 122267. [CrossRef]
25. Yuan, B.; Chen, Y.; Tan, Z.; Jinyan, W.; Liu, H.; Zhang, Y. Label distribution learning-enhanced dual-knn for text classification. In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM), SIAM, Houston, TX, USA, 18–20 April 2024; pp. 400–408.
26. Liang, Z.; Guo, J.; Qiu, W.; Huang, Z.; Li, S. When graph convolution meets double attention: Online privacy disclosure detection with multi-label text classification. *Data Min. Knowl. Discov.* **2024**, *38*, 1171–1192. [CrossRef]
27. Wang, Z.; Zheng, X.; Zhang, J.; Zhang, M. Three-branch BERT-based text classification network for gastroscopy diagnosis text. *Int. J. Crowd Sci.* **2024**, *8*, 56–63. [CrossRef]
28. Wen, Z.; Fang, Y. Prompt tuning on graph-augmented low-resource text classification. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 9080–9095. [CrossRef]
29. Zhang, Y.; Yang, R.; Xu, X.; Li, R.; Xiao, J.; Shen, J.; Han, J. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 2032–2042.
30. Liu, Y.; Huang, L.; Giunchiglia, F.; Feng, X.; Guan, R. Improved graph contrastive learning for short text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024; Volume 38, pp. 18716–18724.
31. Xu, J.F.; Wu, Y.C. A news text classification study based on BERT-BiLSTM-CNN model. *Softw. Eng.* **2023**, *26*, 11–15. [CrossRef]
32. Dou, Z.; Bai, G.; Han, Z.; Li, W.; Li, Y. PFGL-Net: A Personalized Federated Graph Learning Framework for Privacy-Preserving Disease Prediction. *J. Artif. Intell. Res.* **2025**, *2*, 12–23. [CrossRef]
33. Gao, I.J.; Liu, G.; Zhu, B.; Zhou, S.; Zheng, H.; Liao, X. Multi-level attention and contrastive learning for enhanced text classification with an optimized transformer. In Proceedings of the 2025 5th International Conference on Consumer Electronics and Computer Engineering (ICCECE), IEEE, Dongguan, China, 28 February–2 March 2025; pp. 499–503.
34. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.