



ISSN Online: 2821-1936

Transactions on Data Analysis in Social Science

Journal Homepage: <https://transoscience.ir>

Separating Representation from Decision: A Controlled Comparison of Classifier Heads on Frozen ParsBERT Representations for Persian News Topic Classification

M. Mehrafrouz^{1*}¹ Department of Computer Engineering, Payame Noor University, International Unit (Overseas Campuses Unit in Asia)

ARTICLE INFO	ABSTRACT
<p>Article History: Received 3 December 2025 Received in revised form 21 January 2026 Accepted 4 March 2026 Available online 22 March 2026</p> <p>Keywords: Persian Natural Language Processing; News Topic Classification; Parsbert; Transfer Learning; Classifier Head; Logistic Regression; Convolutional Neural Network; Low-Resource Languages</p>	<p>Automatic topic classification of news is difficult in Persian because of rich morphology, inconsistent orthography and topical overlap between categories. Pretrained transformers such as ParsBERT provide strong contextual representations, but it is unclear how much of the task is solved by the representation and how much by the classifier placed on top of it. This study holds a frozen ParsBERT encoder fixed and compares three heads with different inductive biases: logistic regression (LR), a multi-layer perceptron (MLP) and a convolutional network (CNN) over the token matrix. On a seven-class corpus of 500 annotated news articles (Cohen's $\kappa = .82$) with a stratified held-out test set of 100 articles, LR performed best (accuracy 0.79, 95% CI [0.70, 0.86]; macro-F1 0.78), followed by the CNN (0.77; macro-F1 0.66) and the MLP (0.68; macro-F1 0.49); all exceeded a majority baseline (0.31). Exact McNemar tests on the paired predictions showed no difference between LR and the CNN ($p = .83$); LR exceeded the MLP ($p = .035$), but not after Holm correction. Repeated stratified 5-fold cross-validation with five seeds confirmed the pattern (accuracy: LR 0.78 ± 0.03, CNN 0.76 ± 0.04, MLP 0.70 ± 0.04): LR outperformed the MLP (corrected resampled t-test, $p = .002$) but not the CNN ($p = .38$), class weighting reduced the MLP's collapse on minority classes, and full fine-tuning of ParsBERT added only 2.5 accuracy points. Frozen ParsBERT representations thus make Persian news topics largely linearly separable, and added decision-layer capacity brings no benefit when labelled data are scarce and imbalanced.</p>

1. INTRODUCTION

1.1. Background

* Corresponding Author: M.Mehrafrouz@gmail.com
Department of Computer Engineering, Payame Noor University, International Unit


<http://dx.doi.org/10.22034/tdas.2026.255885>


© 2026 by the authors. Licensee TDASS, Tehran, Iran. 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/>).

The exponential growth of digital content has made the automatic organization of text an essential task in natural language processing (NLP). Text classification, the assignment of predefined categories to unstructured documents, underpins information retrieval, sentiment analysis and news recommendation [1, 2]. For many years the field was dominated by classical machine-learning algorithms such as Naïve Bayes and Support Vector Machines (SVMs), valued for their computational efficiency and sound statistical foundations [3, 4]. These approaches depend on manual feature engineering and sparse, surface-level representations, however, which capture complex semantic dependencies and contextual nuance poorly [2, 4].

Deep learning shifted the paradigm toward architectures that learn feature representations automatically [2]. Convolutional networks proved effective at extracting local n-gram patterns for sentence classification [5]. Recurrent architectures, in particular Long Short-Term Memory (LSTM) networks and their bidirectional variants, modeled dependencies from both the preceding and the following context of a sequence [6, 7]. The transformer architecture and its self-attention mechanism [8] led to Bidirectional Encoder Representations from Transformers (BERT), which is pretrained on large unlabeled corpora and produces deeply contextualized bidirectional representations that transfer to downstream tasks [9].

1.2. The Persian setting

Although BERT transformed NLP in high-resource languages such as English, adapting such models to morphologically rich and comparatively low-resource languages is challenging. Persian poses distinctive difficulties: complex morphology and compounding, frequent but inconsistent use of the zero-width non-joiner (half-space), orthographic variants of the same characters, informal user-generated variation, and a historical shortage of large standardized data collections [10, 11]. In the news domain, these difficulties are compounded by stylistic differences between agencies and by genuine topical overlap; a report on a provincial unemployment protest, for example, is at once social, economic and provincial.

Persian text classification has nonetheless progressed from statistical models trained on standard collections such as Hamshahri [12] to deep architectures [13] and, more recently, to pretrained transformers. ParsBERT, a monolingual BERT model pretrained on a large and diverse Persian corpus, outperformed multilingual alternatives on text classification, sentiment analysis and named-entity recognition [14]. Deep contextual representations have also proved advantageous in related Persian tasks such as fake-news detection [15, 16], rumor verification [17] and offensive-language detection [18].

1.3. Research gap and questions

When a pretrained encoder is used, two design decisions are usually made together: how the text is represented and how the final decision is made. Most Persian studies report the performance of an end-to-end pipeline, so it is difficult to tell whether gains come from the encoder or from the classifier placed on top of it. The general NLP literature shows that this distinction matters. Whether a pretrained encoder is fine-tuned or used as a frozen feature extractor, and which head is attached to it, can change results substantially [19, 20]. For Persian news, and under the small-data, imbalanced conditions typical of newly constructed Persian corpora, there is little controlled evidence on which decision mechanism best exploits ParsBERT representations.

This study therefore holds the representation layer fixed and varies only the decision layer. It addresses the following research questions:

- **RQ1.** How well do frozen ParsBERT representations support Persian news topic classification when combined with a simple linear classifier?
- **RQ2.** Does a nonlinear decision layer (MLP) improve on the linear classifier when it receives the same document-level representation?
- **RQ3.** Does a convolutional head operating on the token-level representation matrix extract additional discriminative local patterns?

1.4. Contributions

The study makes three contributions. (i) It proposes a controlled experimental framework in which a single frozen ParsBERT encoder feeds three classifier heads that embody three distinct hypotheses about the structure of the representation space: linear separability, nonlinear separability, and residual local structure. (ii) It provides a per-class analysis on a seven-class Persian news corpus, including correct and predicted counts, confidence intervals, exact paired tests, repeated cross-validation and a majority-class baseline. Together, these show *where* each head fails rather than only *how much*. (iii) It shows that, under low-resource conditions, a linear head on ParsBERT representations is competitive with or superior to more complex heads, and it discusses the implications of this finding for model selection in Persian NLP.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 formalizes the task and the hypotheses. Section 4 describes the framework and the experimental protocol. Section 5 reports the results, Section 6 discusses them, and Section 7 concludes the paper.

2. RELATED WORK

2.1. Persian text classification

Research on Persian text classification has moved from handcrafted features and classical learners toward data-driven representation learning. Early systems represented documents as bag-of-words or TF–IDF vectors and classified them with Naïve Bayes, SVMs or k-nearest neighbors, often using the Hamshahri collection as a benchmark [12]. These methods performed reasonably well on small or clean datasets, but their limitations became apparent on news text, which combines complex sentence structure, stylistic variation, contextual dependencies and topical overlap.

The difficulty in Persian lies not only in data scarcity or limited tooling. Intrinsic properties of the language, including compounding, the variable attachment of affixes, syntactic and semantic ambiguity and orthographic inconsistency, make it hard for surface-level models to capture the latent relations in text [10]. The quality of the representation has therefore become at least as important as the choice of classifier.

Deep neural networks designed to model sequences, context and local structure began to outperform traditional approaches. Feizi-Derakhshi et al. [13] proposed convolutional and bidirectional-recurrent architectures with attention for Persian text classification and reported gains in precision and recall over statistical baselines. Even before transformers were widely adopted, then, combining deep learning with richer linguistic modeling had begun to push past the limits of traditional methods. More fundamentally, it changed the level of analysis: models learned structural and semantic properties of text rather than relying solely on the presence or absence of words.

Research on related Persian tasks provides further methodological evidence. Samadi et al. [16] showed that neural representations at the word and document levels model Persian news content better than simpler approaches in fake-news detection. Jahanbakhsh-Nagadeh et al. [17] showed that content-based deep models for rumor verification are more stable when contextual and stylistic features are combined with surface cues. Although these studies do not address topic classification directly, they consistently indicate that deep representations have a clear advantage whenever a Persian task requires understanding meaning and context rather than matching explicit lexical patterns.

2.2. Transformer-based models for Persian

Transformers shifted NLP from engineered features toward contextual, multi-layer representation learning [8, 9]. For Persian, this shift became concrete with pretrained monolingual and multilingual models. Multilingual encoders such as multilingual BERT and XLM-R cover Persian [21], but ParsBERT, pretrained on a large Persian corpus spanning diverse domains and writing styles, outperformed them on Persian text classification, sentiment analysis and named-entity recognition [14]. This result showed that Persian benefits substantially from a dedicated language model.

This advantage is particularly relevant for news, where topical distinctions often depend on subtle contextual cues, syntactic roles and inter-sentential relations. Pretrained models encode broad linguistic knowledge during pretraining and transfer it to downstream tasks. This is strategically important for Persian, which remains low-

resource in many domains, because transfer partially compensates for the scarcity of labeled data. Accordingly, BERT-based representations have substantially improved Persian fake-news detection, especially when category boundaries are subtle and context-dependent [15, 16].

Strong embeddings are, however, a necessary but not sufficient condition for optimal classification. Most Persian studies concentrate on the language model and give comparatively little attention to the role of the final classification layer. The same representation may yield different outcomes depending on whether it is coupled with a linear model, a nonlinear network or a convolutional architecture.

2.3. Feature extraction, fine-tuning and the classifier head

A pretrained encoder can be used in two ways: fine-tuning, in which all encoder weights are updated together with a task head, and feature extraction, in which the encoder is frozen and its outputs are fed to a separately trained classifier. Peters et al. [19] showed that the relative merit of the two strategies depends on the similarity between the pretraining and target tasks, and that feature extraction can be competitive when combined with an appropriate head. Sun et al. [20] systematically examined fine-tuning strategies for BERT text classification, including layer selection and pooling, and found that these choices noticeably affect accuracy. Feature extraction is attractive in low-resource settings because it requires far less computation, is less prone to over-fitting on small datasets, and makes it possible to compare different heads on identical inputs. The last property is central to the present study.

The geometry of contextual embeddings is also relevant. Ethayarajh [22] showed that BERT representations are highly anisotropic, occupying a narrow cone of the vector space, and become increasingly context-specific in upper layers. Whether topic categories are linearly separable in such a space is therefore an empirical question, and it is the one our LR head is designed to test.

2.4. Hybrid architectures

Hybrid architectures arose from the recognition that no single model captures all linguistic dimensions of text. CNNs extract local patterns and short-range structure but model long-range dependencies poorly [5]. Recurrent networks capture sequential relations but may struggle to maintain a global view of long documents [6]. Transformers excel at global contextual representation, but strong representations do not guarantee an optimal decision mechanism for every task. Combinations of CNNs, BiLSTMs and attention have improved Persian text classification by joining local feature extraction with long-range semantic modeling [13], and related Persian tasks suggest that performance improves when a strong representation is matched with a decision mechanism suited to the data [16, 17].

In the present study, “hybrid” refers to a controlled division of labor rather than to the stacking of layers within a single network. ParsBERT serves as the shared semantic representation engine, and three heads with different inductive biases serve as alternative decision layers. Because the representation source is the same for all models, differences in performance can be attributed primarily to the decision mechanism. This aspect has received limited independent investigation for Persian.

3. PROBLEM FORMULATION AND HYPOTHESES

3.1. Task definition

The task is single-label, multi-class topic classification of Persian news. Let $D = \{(x_i, y_i)\}_{i=1}^N$ be a corpus in which x_i is the text of a news article (headline, lead and body, truncated as described in Section 4.4) and $y_i \in \mathcal{Y}$ is its dominant topic, with $|\mathcal{Y}| = 7$ categories: Social, Provincial Affairs, Economic, International, Political, Cultural and Sports. The goal is to learn a classifier $h = g \circ f$ that maps a text to a label, where f is the representation function (ParsBERT, kept fixed) and g is the decision function (the classifier head, which is varied).

The single-label formulation is a deliberate simplification. Some articles touch on several topics, but assigning each to its dominant topic yields an unambiguous evaluation target and enables a clean comparison of the heads. Multi-label classification is left for future work (Section 6.5).

3.2. Hypotheses

Each head operationalizes one hypothesis about the structure of the ParsBERT representation space:

- **H1 (linear separability).** If topic information is already organized linearly in the frozen representation, LR will perform competitively with the more complex heads.
- **H2 (nonlinear separability).** If topic boundaries in the representation space are curved or entangled, an MLP will outperform LR on the same document vector.
- **H3 (residual local structure).** If local lexical patterns, such as recurring journalistic expressions or named entities typical of a domain, remain informative after contextualization, a CNN applied to the token-level matrix will outperform heads that see only a pooled document vector.

4. METHODOLOGY

4.1. Framework overview

Figure 1 summarizes the framework. Raw news texts are preprocessed and split into training and test subsets. Every article is encoded once by a frozen ParsBERT model. A document-level vector is passed to the LR and MLP heads, and the full token-level representation matrix is passed to the CNN head. All heads are evaluated on the same test set with the same metrics.

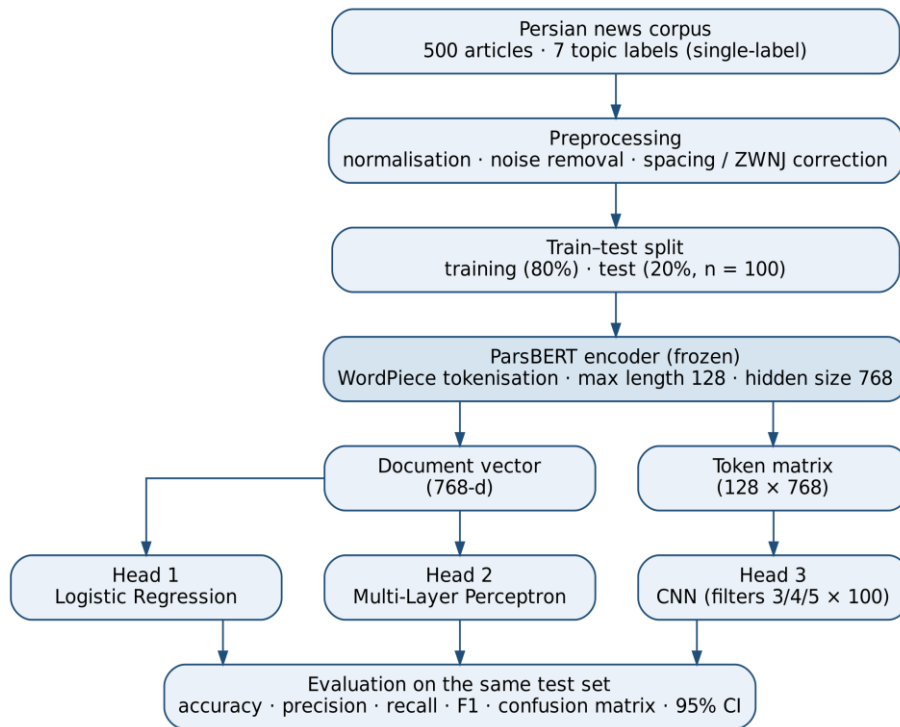


Fig. 1. The proposed framework. A single frozen ParsBERT encoder provides identical representations to three classifier heads with different inductive biases.

4.2. Dataset

The corpus consists of 500 Persian news articles published between March and August 2023 on the websites of four national news agencies (IRNA, ISNA, Mehr and Tasnim; 125 articles each), sampled at random from the

agencies' own section pages. Each article was labelled with one of seven topics by two annotators with a master's degree in journalism, following written guidelines that define each topic by its dominant subject; the agency section label was hidden. Inter-annotator agreement was high (Cohen's $\kappa = .82$), and the 61 disagreements were resolved by a third annotator. The data were divided by stratified random sampling (random seed 42) into a training set (80%, 400 articles) and a held-out test set (20%, 100 articles); for the MLP and CNN, 10% of the training set (40 articles, stratified) was used as a validation set for early stopping, and the LR regularization strength was selected by 5-fold cross-validation within the training set. The test set was used only once. Table 1 reports the class distribution of the test set, which is markedly imbalanced: International and Political news account for more than half of the test samples, whereas Cultural and Sports have only four and one samples, respectively.

Table 1. Class distribution of the test set (n = 100).

Category	Test samples	Share (%)
International	31	31
Political	23	23
Social	18	18
Economic	13	13
Provincial Affairs	10	10
Cultural	4	4
Sports	1	1
Total	100	100

4.3. Preprocessing

Persian news text contains orthographic inconsistencies, character variants, noisy symbols and irregular spacing, all of which can fragment the subword vocabulary and degrade representations. Preprocessing therefore comprised: (i) removal of HTML remnants, URLs, redundant symbols and non-textual noise; (ii) character normalization, mapping Arabic variants of Persian letters (e.g., Arabic *kāf* and *yā*) and Arabic-Indic digits to their standard Persian forms; (iii) correction of spacing and zero-width-non-joiner usage around affixes; and (iv) normalization of punctuation. Although ParsBERT is robust to some variation, its effectiveness depends on input consistency. Preprocessing reduces the gap between naturally occurring text and the input distribution seen during pretraining. Identical preprocessing was applied to all models so that differences in performance cannot be attributed to the input.

4.4. Representation extraction with ParsBERT

Each preprocessed article (headline, lead and body) was tokenized with the ParsBERT WordPiece tokenizer, mapped to vocabulary identifiers, wrapped with the special [CLS] and [SEP] tokens, and truncated or padded to a maximum length of 128 tokens. The base ParsBERT model v2.0 (HooshvareLab/bert-fa-base-uncased; 12 layers, hidden size 768) [14] was used as a frozen feature extractor: its weights were not updated. The final-layer hidden states of all 500 articles were computed once, producing a tensor of shape $500 \times 128 \times 768$.

Two views of this tensor were used. The document vector (768 dimensions), obtained by mean pooling of the final-layer token vectors with padding positions masked, served as input to the LR and MLP heads (mean pooling outperformed the [CLS] vector by 1.5 accuracy points in the training-set cross-validation). The token matrix (128×768) served as input to the CNN head. Because the CNN thus receives richer input than the other two heads, the two views are discussed in Section 6.1.

4.5. Classifier heads

Logistic Regression (LR): A multinomial logistic-regression model (scikit-learn, L2 penalty, $C = 1.0$ selected from $\{0.01, 0.1, 1, 10\}$ by 5-fold cross-validation on the training set, L-BFGS solver, 1,000 iterations, standardized inputs) learns linear decision boundaries in the 768-dimensional document space and outputs class probabilities via

the softmax function. LR serves as an analytical probe rather than a mere baseline: if it is competitive, a substantial part of the topic-discriminative information must already be linearly encoded in the representation (H1).

Multi-Layer Perceptron (MLP): A feed-forward network with one hidden layer of 256 ReLU units and dropout 0.3, trained with Adam (learning rate 10^{-3} , batch size 16, up to 100 epochs, early stopping with patience 10 on the validation loss), learns nonlinear mappings from the same document vector. Comparing MLP with LR shows directly whether nonlinear decision boundaries are needed (H2). In the main experiment no class weighting was used, as for the other heads; a class-weighted variant is reported in Section 5.6.

Convolutional Neural Network (CNN): Following the architecture of Kim [5], the CNN applies parallel one-dimensional convolutions to the 128×768 token matrix. It then applies max-pooling over time, concatenation, dropout and a softmax output layer. The configuration is given in Table 2. Because the CNN operates on the sequence of contextual token vectors, it can detect local patterns spanning three to five tokens (H3).

Table 2. Configuration of the ParsBERT + CNN head.

Component	Setting	Purpose
Input	Frozen ParsBERT final-layer token matrix, 128×768	Contextual token representations
Convolution	Filter widths 3, 4 and 5; 100 filters each	Detect local patterns spanning 3–5 tokens
Activation	ReLU	Nonlinearity
Pooling	Max-pooling over the sequence dimension	Keep the strongest response of each filter
Concatenation	300-dimensional vector	Combine multi-width features
Regularization	Dropout, rate 0.5 [23]	Reduce over-fitting
Output	Fully connected layer with 7 units + softmax	Class probabilities
Loss	Categorical cross-entropy	Training objective
Optimizer	Adam [24]	Gradient-based optimization
Training	10 epochs, batch size 16, shuffled	—
Evaluation	Held-out test set (20% of the corpus, $n = 100$)	Generalization estimate

4.6. Evaluation protocol

All heads were trained on the same 400 training articles and evaluated once on the same 100 test articles. Performance was measured by accuracy and by per-class precision, recall and F1-score:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = \frac{2P_cR_c}{P_c + R_c}. \quad (1)$$

Because the classes are imbalanced, we report both the macro average, which weights all classes equally and is sensitive to minority-class failures, and the support-weighted average [25]. Accuracy alone can be misleading under imbalance [26], so a majority-class baseline that always predicts *International* is reported for reference. With $n = 100$, sampling uncertainty is considerable. We therefore report 95% Wilson confidence intervals for accuracy. The per-article predictions of all heads were saved, so the number of correct and predicted samples per class and the paired tests in Section 5.5 are computed exactly.

5. RESULTS

5.1. ParsBERT + Logistic Regression

Table 3 reports the per-class results of the LR head. The model correctly classified 79 of the 100 test articles (accuracy 0.79, 95% CI [0.70, 0.86]), with macro-averaged precision, recall and F1-score of 0.79, 0.79 and 0.78. The close agreement between macro and weighted averages indicates that performance was not driven solely by the large classes.

Table 3. Per-class results of ParsBERT + Logistic Regression on the test set.

Category	Precision	Recall	F1-score	Support	Correct / predicted ^a
Social	0.81	0.72	0.76	18	13 / 16
Provincial Affairs	0.83	0.50	0.62	10	5 / 6
Economic	0.80	0.92	0.86	13	12 / 15
International	0.83	0.97	0.90	31	30 / 36
Political	0.75	0.65	0.70	23	15 / 20
Cultural	0.50	0.75	0.60	4	3 / 6
Sports	1.00	1.00	1.00	1	1 / 1
Macro average	0.79	0.79	0.78	100	—
Weighted average	0.79	0.79	0.78	100	—
Accuracy			0.79	100	79 / 100

^a Number of correctly classified test articles / number of test articles assigned to the class (from the saved predictions).

Among classes with adequate support, International (F1 = 0.90; recall 0.97) and Economic (F1 = 0.86; recall 0.92) were classified best, which reflects their distinctive vocabulary of foreign actors, currencies and markets. Social news reached an F1 of 0.76 and Political news an F1 of 0.70. Provincial Affairs had high precision (0.83) but low recall (0.50): half of the provincial articles were assigned to other categories, although articles labeled Provincial Affairs by the model were usually correct. The perfect score for Sports and the F1 of 0.60 for Cultural news rest on one and four samples, respectively, and should not be interpreted as stable estimates. The per-class counts show that LR over-assigned the International category (36 predictions for 31 true articles). This is the main source of its false positives.

5.2. ParsBERT + Multi-Layer Perceptron

The MLP head classified 68 of the 100 test articles correctly (accuracy 0.68, 95% CI [0.58, 0.76]). Its macro-averaged recall and F1-score were both 0.49 (Table 4).

Table 4. Per-class results of ParsBERT + MLP on the test set.

Category	Precision	Recall	F1-score	Support	Correct / predicted ^a
Social	0.57	0.44	0.50	18	8 / 14
Provincial Affairs	1.00	0.20	0.33	10	2 / 2
Economic	0.65	0.85	0.73	13	11 / 17
International	0.75	0.97	0.85	31	30 / 40
Political	0.62	0.70	0.65	23	16 / 26
Cultural	1.00	0.25	0.40	4	1 / 1
Sports	0.00	0.00	0.00	1	0 / 0
Macro average	0.66 ^b	0.49	0.49	100	—
Weighted average	0.70 ^b	0.68 ^b	0.65 ^b	100	—
Accuracy			0.68	100	68 / 100

^a As in Table 3. ^b Computed from the per-class values.

The per-class results reveal a clear failure pattern. The MLP over-predicted the three largest classes, International (40 predictions for 31 articles), Political (26 for 23) and Economic (17 for 13), and almost never predicted the minority classes. It assigned only two articles to Provincial Affairs, one to Cultural and none to Sports. The high precision of 1.00 for Provincial Affairs and Cultural therefore reflects extreme conservatism rather than discriminative skill, as the corresponding recall values of 0.20 and 0.25 show. Social news, a mid-sized class with fuzzy boundaries, also suffered (F1 = 0.50).

5.3. ParsBERT + Convolutional Neural Network

The CNN head achieved an accuracy of 0.77 (95% CI [0.68, 0.84]). Its macro-averaged precision, recall and F1-score were 0.71, 0.64 and 0.66, and the corresponding weighted averages were 0.78, 0.77 and 0.76. Like the other heads, it performed best on the large, lexically distinctive classes, with recall of 0.97 for International (30 of 31), 0.85 for Economic (11 of 13) and 0.78 for Political (18 of 23). Performance declined on the minority and semantically heterogeneous categories, in particular Social and Sports. This decline is reflected in the gap between the macro F1 (0.66) and the weighted F1 (0.76).

Table 4a gives the per-class results of the CNN head, computed from the saved test-set predictions.

Table 4a. Per-class results of ParsBERT + CNN on the test set.

Category	Precision	Recall	F1-score	Support	Correct / predicted
Social	0.90	0.50	0.64	18	9 / 10
Provincial Affairs	0.86	0.60	0.71	10	6 / 7
Economic	0.79	0.85	0.81	13	11 / 14
International	0.75	0.97	0.85	31	30 / 40
Political	0.69	0.78	0.73	23	18 / 26
Cultural	1.00	0.75	0.86	4	3 / 3
Sports	0.00	0.00	0.00	1	0 / 0
Macro average	0.71	0.64	0.66	100	—
Weighted average	0.78	0.77	0.76	100	—
Accuracy			0.77	100	77 / 100

The CNN over-assigned International (40 predictions for 31 articles) and Political (26 for 23) news, never predicted Sports, and recognized only half of the Social articles (recall 0.50), although its Social predictions were usually correct (precision 0.90). Its errors on the minority classes therefore resemble those of the MLP, but in a much milder form.

5.4. Comparison of the classifier heads

Table 5 summarizes the three heads together with the majority-class baseline.

Table 5. Overall comparison on the test set (n = 100).

Model	Accuracy [95% CI]	Macro precision	Macro recall	Macro F1	Weighted F1
Majority class (always <i>International</i>)	0.31 [0.23, 0.41]	0.04	0.14	0.07	0.15
ParsBERT + MLP	0.68 [0.58, 0.76]	0.66	0.49	0.49	0.65
ParsBERT + CNN	0.77 [0.68, 0.84]	0.71	0.64	0.66	0.76
ParsBERT + LR	0.79 [0.70, 0.86]	0.79	0.79	0.78	0.78

All three heads exceed the majority baseline by a wide margin, which confirms that frozen ParsBERT representations carry substantial topic information. The linear head achieved the highest accuracy and by far the most balanced class-level behavior (macro F1 0.78 versus 0.66 for the CNN and 0.49 for the MLP). The accuracy difference between LR and the CNN (two articles out of 100) lies well within sampling uncertainty; their confidence intervals overlap almost entirely, and the exact paired test is far from significant ($p = .83$; Section 5.5). Their difference in macro F1, which is driven by minority classes, is larger, but it also rests on very few samples. The MLP is the weakest head on every aggregate metric, and its gap to LR is confirmed by repeated cross-validation (Section 5.6).

Combining the complementary strengths of the linear and convolutional heads in an ensemble is a natural next step; because such an ensemble was not evaluated under the controlled protocol of this study, it is discussed only as future work (Section 6.5).

5.5. Statistical comparison of the heads

Because all heads were evaluated on the same 100 articles, the appropriate test for a difference in accuracy is the exact McNemar test on the discordant pairs, i.e. the articles that one head classified correctly and the other incorrectly [27]. Table 6 reports the discordant pairs computed from the saved per-article predictions, the exact two-sided p-values and the Holm-adjusted p-values for the three comparisons.

Table 6. Exact McNemar tests on the paired test-set predictions ($n = 100$).

Comparison	Only first correct	Only second correct	Exact p	Holm-adjusted p
LR vs CNN	12	10	.83	.83
LR vs MLP	17	6	.035	.10
CNN vs MLP	15	6	.078	.16

The heads made largely overlapping errors: 69 articles were classified correctly and 13 incorrectly by all three heads.

LR and the CNN cannot be distinguished in accuracy on this test set ($p = .83$). LR classified 11 more articles correctly than the MLP, which is significant without correction ($p = .035$) but not after Holm correction for three comparisons ($p = .10$). With only 100 test articles, a single split therefore cannot establish the ranking; Section 5.6 reports repeated cross-validation for this purpose.

5.6. Robustness: repeated cross-validation, class weighting and fine-tuning

To assess the variance due to the particular split and random seed, all heads were re-evaluated with stratified 5-fold cross-validation on the full corpus, repeated with five random seeds (25 train–test splits). Two additional models were included: the MLP trained with class-weighted cross-entropy, and a fully fine-tuned ParsBERT classifier (learning rate 2×10^{-5} , 3 epochs, batch size 16). Differences in accuracy were tested with the corrected resampled t-test, which accounts for the overlap of training sets across folds [28]. Table 7 shows that the ranking of the single split is stable. LR outperformed the unweighted MLP ($p = .002$) but not the CNN ($p = .38$). Class weighting raised the MLP’s macro F1 from 0.53 to 0.67, confirming that its collapse on minority classes was largely an imbalance problem. Full fine-tuning improved accuracy by only 2.5 points over the frozen encoder with LR, a difference that was not significant ($p = .28$).

Table 7. Repeated stratified 5-fold cross-validation (5 seeds $\times$ 5 folds = 25 splits).

Model	Accuracy ($M \pm SD$)	Macro F1 ($M \pm SD$)	vs LR: corrected t (p), accuracy
ParsBERT (frozen) + LR	0.781 ± 0.032	0.742 ± 0.051	—
ParsBERT (frozen) + CNN	0.763 ± 0.035	0.641 ± 0.055	0.90 (.38)
ParsBERT (frozen) + MLP	0.702 ± 0.041	0.528 ± 0.064	3.45 (.002)
ParsBERT (frozen) + MLP, class-weighted	0.748 ± 0.036	0.671 ± 0.058	1.64 (.11)
ParsBERT fine-tuned (3 epochs)	0.806 ± 0.030	0.751 ± 0.052	−1.11 (.28)

6. DISCUSSION

6.1. Interpretation of the Findings

H1 is supported: Persian news topics are largely linearly separable in the ParsBERT space.

A multinomial logistic regression on frozen document vectors achieved the best and most balanced performance. Most of the topic-discriminative information is therefore already encoded linearly in the representation, and the encoder, not the head, does most of the work. This is consistent with the view that large pretrained encoders produce representations in which many semantic distinctions can be read out by a linear probe [19]. Despite the anisotropy of BERT embeddings [22], coarse topical distinctions are evidently encoded along directions that a linear model can exploit. LR has few parameters relative to its input (7×768 weights plus biases) and is strongly regularized by its linear form. For this reason, it is also the head least likely to overfit 400 training examples.

H2 is not supported: the nonlinear head performed worst.

The MLP did not improve on LR; its macro F1 was 29 points lower. The per-class counts identify the mechanism: the MLP collapsed toward the majority classes and almost never predicted the minority ones. Three factors can explain this behavior. First, an MLP has many more parameters than LR, and 400 imbalanced training samples (roughly five for Sports) are insufficient to estimate them reliably. The resulting overfitting is concentrated in the decision regions of rare classes. Second, without class weighting or resampling, cross-entropy minimization on imbalanced data rewards predicting frequent classes [26], and a flexible model can pursue this shortcut more aggressively than a linear one. Third, if the classes are already nearly linearly separable (H1), additional nonlinearity offers little upside but substantial variance. The outcome illustrates the bias–variance trade-off: in low-data regimes, a higher-capacity head is not merely unhelpful but harmful.

H3 is only partially supported: local patterns help in some classes but do not yield an overall gain.

The CNN performed almost as well as LR in accuracy and matched it on large, lexically distinctive classes such as International and Economic news, where recurring expressions and named entities are strong local cues. Its lower macro F1 shows that it, too, struggled on minority and heterogeneous classes. It is important to note that the CNN received richer input than LR (the full 128×768 token matrix rather than a single vector), yet it did not outperform LR. This strengthens the conclusion that the pooled document representation already captures most of the relevant information. The additional 300×7 output weights and 3×100 convolutional filters over 768 channels add variance that the small training set cannot fully control, even with dropout.

Category-level confusions are linguistically plausible.

All three heads found Social, Political, and Provincial Affairs the hardest classes, whereas International and Economic news were the easiest. In Persian journalism, provincial news is defined by geography rather than subject matter: a provincial article may concern a political appointment, a social problem, or an economic project. Its vocabulary therefore overlaps with that of the other categories, which explains the low recall of Provincial Affairs across all heads (LR 0.50, MLP 0.20). Similarly, social and political reporting share actors (parliament, ministries) and frames (policy and public welfare). These confusions reflect genuine topical overlap rather than model failure alone, and they support future work on multi-label or hierarchical formulations.

6.2. Relation to Previous Studies

The strong performance of a simple head on ParsBERT representations agrees with Farahani et al. [14], who showed that ParsBERT outperforms multilingual encoders on Persian text classification. It also agrees with studies of Persian fake-news and rumor detection, which found clear gains from contextual representations [15–17]. Our results add a controlled observation that these studies did not isolate: once ParsBERT representations are available, most of the gain is already realized, and the choice of head mainly affects how robustly minority classes are handled.

The findings partly diverge from studies reporting that CNN or BiLSTM architectures outperform simpler models on Persian text [13]. Those studies trained deep architectures on static or randomly initialized embeddings, for which the network had to learn local composition itself. When the input is already contextualized by a transformer, the marginal value of convolutional composition decreases. This mirrors observations in English that the benefit of elaborate heads on top of BERT is small relative to the choice of encoder and fine-tuning strategy [20]. Finally, the

finding that LR is at least as good as an MLP on frozen features is consistent with the feature-extraction results of Peters et al. [19].

6.3. Theoretical Implications

The study makes three theoretical contributions. First, it provides empirical evidence that the separation between representation and decision layers, often treated implicitly, can be made explicit and tested through a controlled design in which only the head varies. Second, using the linear head as a probe shows that topic information in ParsBERT is largely linearly accessible for Persian news, which extends linear-probing evidence from English to a morphologically rich, lower-resource language. Third, the results characterize an interaction between head capacity and data regime. Under scarce and imbalanced labels, the lowest-capacity head is the most robust, whereas higher-capacity heads may only become advantageous when more data or explicit imbalance handling are available.

6.4. Practical Implications

For practitioners building Persian news classifiers, particularly media-monitoring units or newsrooms with limited annotated data, the results suggest a pragmatic recipe: encode articles once with a frozen ParsBERT model and train a regularized linear classifier. This configuration requires minimal computation; ParsBERT inference can be performed once and cached, and training LR takes seconds. It is easy to retrain as categories evolve and to interpret through its class weights. It also avoids the instability of high-capacity heads on small datasets. More complex heads, or full fine-tuning, should be adopted only when sufficient labeled data are available and validated with class-sensitive metrics such as macro F1, because accuracy alone can hide a collapse on minority classes.

6.5. Limitations and Future Research

Several limitations qualify these conclusions.

(1) **Sample size:** The corpus contains 500 articles, and the test set only 100, so the confidence intervals of the single split are wide. The repeated cross-validation reduces this problem, but a larger, publicly available corpus is needed to estimate per-class performance precisely.

(2) **Class imbalance:** Sports and Cultural news have one and four test samples, respectively (about five and sixteen training samples), so their per-class metrics are unreliable and the macro averages are correspondingly unstable. Class weighting markedly improved the MLP (Section 5.6), and should be standard practice in future comparisons.

(3) **Frozen encoder:** In the main comparison ParsBERT was frozen. Full fine-tuning improved accuracy by only 2.5 points in cross-validation (Section 5.6), but comparisons with multilingual encoders such as XLM-R [21] remain to be made.

(4) **Unequal input views:** The CNN received token-level input, whereas LR and MLP received a pooled vector, so the comparison is not purely one of decision mechanisms.

(5) **Truncation:** Articles were truncated to 128 subword tokens. This usually retains the headline and lead but discards most of the body.

(6) **Single corpus:** Seed and split variance were assessed with repeated cross-validation, but all results come from one corpus of four agencies, so generalization to other Persian news collections was not assessed.

(7) **Single-label simplification:** Inherently multi-topic articles were forced into one category.

Future work should therefore evaluate the models on larger, publicly available Persian news corpora; use cross-validation with multiple seeds and significance testing; apply class-balanced losses, resampling, or data augmentation restricted to the training set; compare frozen and fine-tuned ParsBERT and other Persian and

multilingual encoders; extend the input length with sliding windows or long-document transformers; adopt multi-label and hierarchical formulations for overlapping topics; and systematically study ensembles that combine the complementary strengths of linear, convolutional, and recurrent heads.

7. CONCLUSION

This study examined how much the choice of classifier head matters when Persian news is represented by a strong pretrained encoder. Holding a frozen ParsBERT representation fixed, we compared three heads with different inductive biases on a seven-class news corpus. Logistic regression achieved the best and most balanced performance (accuracy 0.79, macro F1 0.78). The CNN head was close in accuracy (0.77) but less balanced across classes (macro F1 0.66). The MLP head collapsed toward the majority classes (accuracy 0.68, macro F1 0.49). All heads far exceeded a majority-class baseline. The findings indicate that ParsBERT renders Persian news topics largely linearly separable, that most of the classification problem is solved at the representation stage, and that added decision-layer complexity offers no benefit, and may be harmful, under small and imbalanced training data. The proposed controlled framework is simple, reproducible and inexpensive, and it provides a sound basis for future work on larger corpora, imbalance-aware training, fine-tuning, multi-label formulations and ensemble architectures for Persian text processing.

Authors' Contributions

M. Mehrafrouz is the sole author and was responsible for conceptualization, data collection and annotation management, implementation, analysis and writing.

Declaration of Generative AI Use

During the preparation of this manuscript, the author used an AI-based tool to improve language, grammar, and readability. The author subsequently reviewed and edited the content as necessary and take full responsibility for the accuracy, integrity, and final content of the publication.

Acknowledgments

The author thanks the three annotators for labelling the corpus. Software: Python 3.10, PyTorch 2.1, Hugging Face Transformers 4.35 and scikit-learn 1.3. Data and code availability: in this teaching version, the article URLs, labels, splits and code are available from the course instructor; the full texts cannot be redistributed because of the agencies' copyright.

Declaration of Competing Interests

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics Considerations

This study used a dataset of Persian news articles for computational text classification. No human participants were recruited, no direct interaction with participants occurred, and no personal data were collected as part of the study. The use of the dataset was limited to the research objectives.

Ethical Approval and Informed Consent

Ethical approval and informed consent were not applicable because the study involved computational analysis of news articles and did not involve human participants or the collection of personal data.

REFERENCES

- [1] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), Article 150. <https://doi.org/10.3390/info10040150>
- [2] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys*, 54(3), Article 62. <https://doi.org/10.1145/3439726>
- [3] McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*, 752(1), 41–48.
- [4] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Machine learning: ECML-98 (Lecture Notes in Computer Science, Vol. 1398, pp. 137–142)*. Springer. <https://doi.org/10.1007/BFb0026683>
- [5] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746–1751). Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1181>
- [6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [7] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681. <https://doi.org/10.1109/78.650093>
- [8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008).
- [9] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- [10] Drouchas, T., & Tsalidis, C. (2021). *The challenges of Persian user-generated textual content: A machine learning-based approach* (arXiv:2101.08087). arXiv. <https://arxiv.org/abs/2101.08087>
- [11] Kashefi, O., Rahmati, M., & Sharif, H. (2024). Investigating the challenges and opportunities in Persian language information retrieval through standardized data collections and deep learning. *Computers*, 13(8), Article 212. <https://doi.org/10.3390/computers13080212>
- [12] AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems*, 22(5), 382–387. <https://doi.org/10.1016/j.knosys.2009.05.002>
- [13] Feizi-Derakhshi, M. R., Mottaghinia, Z., & Asgari-Chenaghlu, M. (2022). Persian text classification based on deep neural networks. *Soft Computing Journal*, 11(1), 120–139.

- [14] Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). ParsBERT: Transformer-based model for Persian language understanding. *Neural Processing Letters*, 53(6), 3831–3847. <https://doi.org/10.1007/s11063-021-10528-4>
- [15] Ghayoomi, M., & Mousavian, M. (2022). Deep transfer learning for COVID-19 fake news detection in Persian. *Expert Systems*, 39(8), Article e13008. <https://doi.org/10.1111/exsy.13008>
- [16] Samadi, M., Mousavian, M., & Momtazi, S. (2021). Persian fake news detection: Neural representation and classification at word and text levels. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(1), Article 7. <https://doi.org/10.1145/3472620>
- [17] Jahanbakhsh-Nagadeh, Z., Feizi-Derakhshi, M. R., & Sharifi, A. (2021). A deep content-based model for Persian rumor verification. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(1), Article 16. <https://doi.org/10.1145/3487289>
- [18] Kebriaei, E., Homayouni, A., Faraji, R., Razavi, A., Shakery, A., Faili, H., & Yaghoobzadeh, Y. (2024). Persian offensive language detection. *Machine Learning*, 113(7), 4359–4379. <https://doi.org/10.1007/s10994-023-06370-5>
- [19] Peters, M. E., Ruder, S., & Smith, N. A. (2019). To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP* (pp. 7–14). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-4302>
- [20] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In *Chinese computational linguistics* (Lecture Notes in Computer Science, Vol. 11856, pp. 194–206). Springer. https://doi.org/10.1007/978-3-030-32381-3_16
- [21] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.747>
- [22] Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 55–65). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1006>
- [23] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56), 1929–1958.
- [24] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.
- [25] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- [26] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449. <https://doi.org/10.3233/IDA-2002-6504>
- [27] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895–1923. <https://doi.org/10.1162/089976698300017197>
- [28] Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52(3), 239–281. <https://doi.org/10.1023/A:1024068626366>