Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TRNEWS-2025: A Large-Scale, Manually Annotated Turkish News Dataset for Text Classification and NLP Research

Sengul Bayrak

bayraksengul@ieee.org

Research Article

Keywords: Turkish Natural Language Processing, News Dataset, Text Classification, TRNEWS-2025

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8950363/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TRNEWS-2025: A Large-Scale, Manually Annotated Turkish News Dataset for Text Classification and NLP Research

Sengul Bayrak[1*]

[1*]Department of Computer Engineering, Istanbul Sabahattin Zaim University, Halkali Street, Istanbul, 34290, Kucukcekmece, Turkiye.

Corresponding author(s). E-mail(s): bayraksengul@ieee.org;

Abstract

This study presents TRNEWS-2025, a labeled Turkish news dataset created for use in Turkish Natural Language Processing (NLP) and machine learning research. The TRNEWS-2025 dataset comprises news articles gathered from various Turkish online news sites and media portals up to the year 2025. It has been manually labeled under nine different categories: Magazine, Politics, Sports, Arts-Culture, Health, Finance-Economy, Science-Technology, Tourism, and Environment. The dataset aims to reflect real-world news language due to its diverse range of topics and up-to-date content. To evaluate the usability and distinctiveness of the dataset, experimental analyses were performed using BERT, a modern transformer-based model, and BiLSTM+GloVe, a classic deep learning architecture. The results show that the TRNEWS-2025 dataset can be used effectively with both model paradigms. The TRNEWS-2025 dataset aims to provide a reliable and up-to-date source dataset that can be utilized in various research areas, including Turkish text classification, topic identification, explainable artificial intelligence studies, and language model evaluations. The original contribution of a Turkish dataset, such as TRNEWS-2025, significantly provides up-to-date, comprehensive, and nine-class news data.

Keywords: Turkish Natural Language Processing, News Dataset, Text Classification, TRNEWS-2025

1 Introduction

The success of methods developed in the fields of NLP and machine learning depends largely on the availability of high-quality, up-to-date, and accurately labeled datasets (Yu and Wang (2025),Boudribila et al. (2025)). Many essential NLP tasks, such as text classification, topic detection, sentiment analysis, and information retrieval, can be performed effectively thanks to reliable datasets. However, for languages with relatively few resources, such as Turkish, the number of labeled datasets with high domain diversity and up-to-date content is limited (Khan et al. (2025)). A significant portion of existing Turkish news datasets is insufficient to meet current research needs due to reasons such as a limited number of categories, outdated content, narrow subject coverage, or access restrictions. Especially in rapidly changing fields such as science and technology, health, and the environment, there is a growing need for datasets that reflect current language usage and social issues. To meet these requirements, a labeled Turkish news dataset called TRNEWS-2025 has been created. TRNEWS-2025 comprises news texts gathered from various Turkish online newspapers and media portals up to the year 2025. Each news text in the dataset is labeled under one of nine different categories: Magazine, Politics, Sports, Arts-Culture, Health, Finance-Economy, Science-Technology, Tourism, Environment. This multi-class structure ensures that the dataset provides a suitable evaluation environment for both classical machine learning algorithms and deep learning-based language models. One of the distinctive features of the TRNEWS-2025 dataset is that it was collected through a systematic and academic process, encompassing a diverse range of fields. TRNEWS-2025, which provides a current, multi-domain, and labeled Turkish news dataset, aims to serve as a valuable resource for various research areas, including Turkish text classification, topic modeling, explainable artificial intelligence studies, and comparative model performance evaluations. In this regard, the data set provides an infrastructure that supports reproducibility in Turkish NLP research and contributes to the development of new methods. A literature review on the subject was conducted under the titles Turkish News and Text Datasets, Limitations of Existing Turkish News Datasets, and Benchmark Approaches in Turkish NLP.

1.1 Related Works

The existing news and text datasets in the field of Turkish Natural Language Processing are being examined, the limitations of these datasets are discussed, and the position of the TRNEWS-2025 dataset in the literature is revealed. The aim is to clearly justify why a new and up-to-date Turkish news dataset is needed, rather than focusing on model comparisons. In recent years, large-scale and domain-specific datasets have emerged in the field of Turkish Natural Language Processing. Large-scale benchmark datasets, such as TurkishMMLU, have become available for training and professional examination tasks (Bayram et al. (2025)). In addition, medical natural language inference datasets have been developed, including TurkMedNLI (Akbasli et al. (2025)). Despite this, there is a lack of up-to-date and comprehensive Turkish datasets in the literature for multi-class language tasks such as news classification. In this context,

the TRNEWS-2025 dataset makes a significant contribution as one of the first comprehensive resources that reflect the rich distribution and multi-categorical structure of news language.

In recent years, various labeled text datasets have been created in the field of Turkish Natural Language Processing. A significant portion of these datasets focuses on fundamental NLP tasks such as text classification and sentiment analysis. The TTC-4900 dataset is one of the early studies created to classify Turkish news texts (Branwen and Emigawaty (2025)). Similarly, Turkish text datasets developed in various studies have been shaped around a limited number of categories and specific subject areas. Additionally, Turkish sentiment analysis datasets based on social media data or user comments are widely available in the literature. However, such datasets typically consist of short texts and do not adequately reflect the long contextual structure and linguistic diversity of news texts. Furthermore, some Turkish news datasets are closed access, offering limited opportunities for reproducible research. There are limited Turkish datasets in the literature. Zorarpaci emphasized that the automatic classification of Turkish news texts is an important and challenging problem in the field of Natural Language Processing, and that the agglutinative and morphologically rich structure of Turkish makes this task even more complex. This study proposes a deep learning approach that combines the Word2Vec word embedding method with the Densely Connected Convolutional Networks (DenseNet) architecture to improve Turkish news classification performance. In the proposed method, 256-dimensional Word2Vec vectors obtained from news texts were converted into 16×16 matrices and used as input to the DenseNet-BC architecture. The model was evaluated using the Turkish TTC-4900 news dataset, and a performance analysis was conducted based on precision, recall, and F1-score metrics. Experimental results show that the proposed method achieves an average F1-score of 90.02% and offers competitive performance compared to many traditional and deep learning-based methods, excluding pre-trained large language models. The findings reveal that integrating Word2Vec and DenseNet offers an effective and computationally efficient alternative for Turkish news classification problems (Zorarpaci (2025)). Kılınc et al. emphasized that the rapid increase in the number of documents on the internet makes it difficult to classify texts under the correct categories, especially on news portals. In this context, they highlighted that text classification methods, approached as supervised learning problems, have become an important need and that research in this area is limited due to the scarcity of accessible and widely available datasets for Turkish. In this study, a new and well-documented dataset called TTC-3600 was created for use in text classification tasks on Turkish news and articles. The data set is presented in file formats compatible with common text mining tools. Within the scope of the study, five different classifiers commonly used in text classification and two feature selection methods were evaluated on the TTC-3600 dataset. The experimental results show that, after preprocessing and feature selection steps, the Random Forest classifier achieved the best performance with an accuracy rate of 91.03% using the attribute-based feature selection method. The TTC-3600 data set, which is available as open access, and the experimental results obtained are expected to contribute to future comparative studies of the classification of Turkish texts (Kılınç et al. (2017)). Türk et al. introduce the BOUN Treebank, a

comprehensive Turkish dependency treebank, and a new annotation tool called BoAT, which contributes to the limited number of syntactic resources available for Turkish. Within the scope of the study, 9,761 sentences from various text types, selected from the Turkish National Corpus, were manually labeled by expert linguists according to the Universal Dependencies (UD) framework. BOUN Treebank is more comprehensive than existing Turkish UD treebanks in terms of size and diversity, reflecting the structural features of Turkish, such as its morphological richness and free word order, in a more consistent manner. The study also addresses inconsistencies in syntactic tagging for Turkish, such as embedded clauses, conjunctive structures, copula usage, compound structures, etc., in detail and proposes integrative solutions compatible with UD. The developed BoAT tool provides a practical solution, especially for agglutinative languages, with its ability to display both tree and table views, detailed configuration of morphological features, and features such as token merging/splitting. Experimental results show that dependency parsing performance using the BOUN Treebank has increased significantly compared to previous Turkish treebanks (Türk et al. (2022)). Kayakuş and Açıkgöz (2022) addressed the classification of Turkish news texts using machine learning methods, based on the importance of presenting news under the correct categories in digital journalism. In the study, an original dataset was created comprising a total of 10,500 news texts collected via RSS from five high-traffic digital news sites in Turkey, divided into three categories: economy, sports, and world. Comprehensive preprocessing, feature extraction (root words and word-level n-grams), and term weighting steps were applied to the texts. Naive Bayes and Decision Trees algorithms were used in the classification phase; performance evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The experimental results show that the Naive Bayes classifier is more successful, achieving an F1-score of 88.66% compared to the 82.96% accuracy of Decision Trees. The findings are consistent with previous studies in the field of Turkish news classification, demonstrating that statistical machine learning methods continue to offer competitive performance on Turkish texts. The study emphasizes that larger datasets and deep learning-based approaches could increase success in the future. Among the existing datasets in Turkish, TurkishMMLU presents the first comprehensive benchmark developed to evaluate the multi-domain language understanding capabilities of large Turkish language models. The dataset comprises 10,032 multiple-choice questions based on the high school curriculum, covering various subject areas and difficulty levels. Evaluations conducted on various open-source and closed-source LLMs show that model performance varies significantly depending on the subject area and question difficulty. TurkishMMLU provides a standard reference for the comparative evaluation of Turkish LLMs (Yüksel et al. (2024)). The TurkMedNLI dataset presents TurkMedNLI, the first medical Natural Language Inference (NLI) dataset for Turkish, a low-resource language. The MedNLI dataset was translated into Turkish using large language models (Llama-3.1) and the NLLB translation model; doma. Domain-specific issues, such as the expansion of medical abbreviations and conversions of units of measurement, addressed using a dedicated pipeline. The dataset has been validated using both automatic and expert physician evaluations with BERT, BERTurk, BioBERT, and BioBERTurk models and is presented as a reliable resource for Turkish biomedical NLP studies (Oğul et al.

(2025)). TR-MTEB presents TR-MTEB, the first comprehensive sentence embedding evaluation framework for Turkish, and evaluates Turkish embedding models using 26 datasets covering six different NLP tasks, including classification, clustering, retrieval, and semantic similarity. Furthermore, it has been demonstrated that Turkish-specific embedding models, trained on 34.2 million weakly supervised Turkish sentence pairs, offer performance comparable to that of multilingual models, especially in monolingual tasks. The study provides a standard evaluation infrastructure and robust baseline models for Turkish NLP (Baysan and Güngör (2025)). It presents TurkColBERT, the first comprehensive comparative benchmark for Turkish information retrieval, and systematically compares dense models with late-interaction ColBERT-based models. The results show that late-interaction models, which preserve token-level interaction due to the morphological complexity of Turkish, provide a clear advantage over dense models in terms of accuracy and efficiency. Furthermore, it has been demonstrated that Turkish information retrieval systems with high accuracy, low latency, and production suitability are possible with MUVERA indexing (Ezerceli et al. (2025)). Although Turk, such as TurkishMMLU, TurkMedNLI, and TR-MTEB, have been added to the literature in recent years, these studies have primarily focused d primarily on question-answering, natural language inference, or text embedding benchmark tasks. In contrast, the number of open-access datasets covering multi-class and current Turkish news texts remains limited. The TRNEWS-2025 dataset aims to fill this gap. The rest of the study consists of Section 2, which introduces the TR-NEWS dataset; Section 3, which applies preprocessing to the dataset; Section 4, which models the dataset; and finally, the conclusion section.

2 Description of TRNEWS-2025 Dataset

The dataset is designed for multi-class text classification problems and provides a suitable infrastructure for news classification, topic identification, text mining, and explainable artificial intelligence (XAI) studies. Furthermore, the dataset is suitable for comparatively evaluating the performance of classical machine learning such as Naive Bayes, Support Vector Machines, and Random Forests, Random Forests alongside deep learning and pre-trained language models such as BERTurk. In this regard, thanks to its current content, field diversity, and labeled structure, the data set offers a comprehensive data source that can be used for both training and research purposes in Turkish NLP studies.

2.1 Data Collection Process

The TRNEWS-2025 dataset was created from news texts obtained from open-access Turkish online news sites, digital newspapers, and media portals. The data collection process involved collecting publicly published news content using web scraping methods. In this process, only news sources that are freely accessible and do not require membership or special permission were preferred. Data collection was conducted over the period from 2023 to 2025, with the aim of ensuring that the dataset accurately reflects current language usage and contemporary social issues. News texts were collected under nine predefined categories: Magazine, Politics, Sports, Arts–Culture,

Health, Finance–Economy, Science–Technology, Tourism, and Environment. Data was obtained from different news sources to ensure content diversity and source independence (Bayrak (2025)). The following criteria were considered in the data selection process. Inclusion criteria include: the news text being in Turkish, containing the full text, clearly matching the relevant category, and not being duplicate content. Exclusion criteria are: advertising content, opinion pieces, user comments, incomplete texts, automatically generated content, and texts that do not qualify as news. Furthermore, copies of the same news story published in different sources were not included in the dataset. The dataset does not contain personal data belonging to human subjects. All news texts were obtained from publicly available sources, and no additional processing was performed to enrich them with personal names, specific identity information, or sensitive data. Therefore, the data collection process for this approval from an ethics committee committee approval. However, compliance with ethical principles and responsible data use principles has been ensured in the preparation of the dataset.

2.2 Labeling

Each news text in the dataset was labeled with one of nine predefined categories based on a content-based evaluation. A multi-labeler approach was adopted in the labeling process; each news text was reviewed by at least two different labelers. In cases where there was disagreement among annotators, the relevant examples were re-evaluated, and the final decision was made under the supervision of the academic advisor. This approach aimed to increase label accuracy and minimize subjective errors. Each news article was labeled with only one class, and the dataset has a single-labeled, multi-class structure. Annotation Guideline A written annotation guideline was prepared to ensure consistency in the annotation process for annotators. This guideline outlines the scope definition for each category, provides sample topic headings, and addresses borderline cases like intersections between economics and politics, or health and science. Annotators performed the annotation tasks according to this guideline. To assess the reliability of the annotation process, inter-annotator consistency was analyzed using Cohen's Kappa coefficient. Cohen's Kappa is a widely used metric that measures the true level of agreement between annotators, excluding random agreement. Analysis of samples independently labeled by the labelers revealed that the Cohen's Kappa coefficient was satisfactory greater than 0.60. In the literature, this value is considered to indicate a good level of labeler agreement. Samples showing low agreement were re-examined as part of the quality control process, and the final labels were updated NLP accordingly.

To assess the reliability of the labeling process, inter-labeler agreement metrics were considered. In cases of multiple labeling, the agreement rate between labelers was calculated, and examples showing low consistency were re-evaluated. This quality control step was implemented to ensure that the dataset could be used reliably in academic studies. The consistency achieved in the labeling process is dataset to be used for both be used for both or training and testing purposes.

Table 1 Class-wise Observation Numbers in the Dataset

Class	Observation Number
Magazine	315917
Politics	147261
Sports	131443
Arts–Culture	46632
Health	47861
Finance–Economy	36680
Science–Technology	30513
Tourism	29672
Environment	15610

3 Feature Engineering

The TRNEWS-2025 dataset has undergone various preprocessing steps to make it suitable for use in machine learning and deep learning models. The preprocessing process aims to reduce noise while preserving the structural integrity of the texts. The basic preprocessing steps applied are summarized below: Encoding texts in UTF-8 format Cleaning HTML tags, special characters, and unnecessary spaces Normalizing punctuation marks Converting all texts to lowercase Removing duplicate news content The preprocessing steps were performed with care to preserve the raw state of the dataset. Processes such as lemmatization, auto-generatedtop-word removal have not been directly integrated into the dataset so that users can flexibly apply them according to their own experimental setups. This approach supports the general-purpose use of the TRNEWS-2025 dataset for different modeling strategies.

Basic statistical analyses were performed on the TRNEWS-2025 data set to better reveal its structural characteristics and provide researchers with a quantitative perspective on the data. These analyses cover class distributions, text lengths, and general data characteristics.

The dataset has a multi-class structure consisting of nine different news categories. The number of examples and percentage distributions for each class are presented in Table 1. Examining the class distribution is important for evaluating the balanced structure of the dataset and possible class imbalances. This distribution shows that the dataset covers different types of news and provides a suitable structure for multi-class classification problems.

Figure 1 shows the word count distribution of news texts in the TRNEWS-2025 dataset by class.

The analysis results show that the dataset has a wide distribution ranging from short news texts to long content. This indicates that the dataset is suitable for both traditional machine learning methods and deep learning-based language models.

To better reflect the content structure of the dataset, a limited number of representative news text examples, each belonging to a different class, are provided. Table 2 shows a short excerpt from a news text selected from the dataset and the corresponding class label.

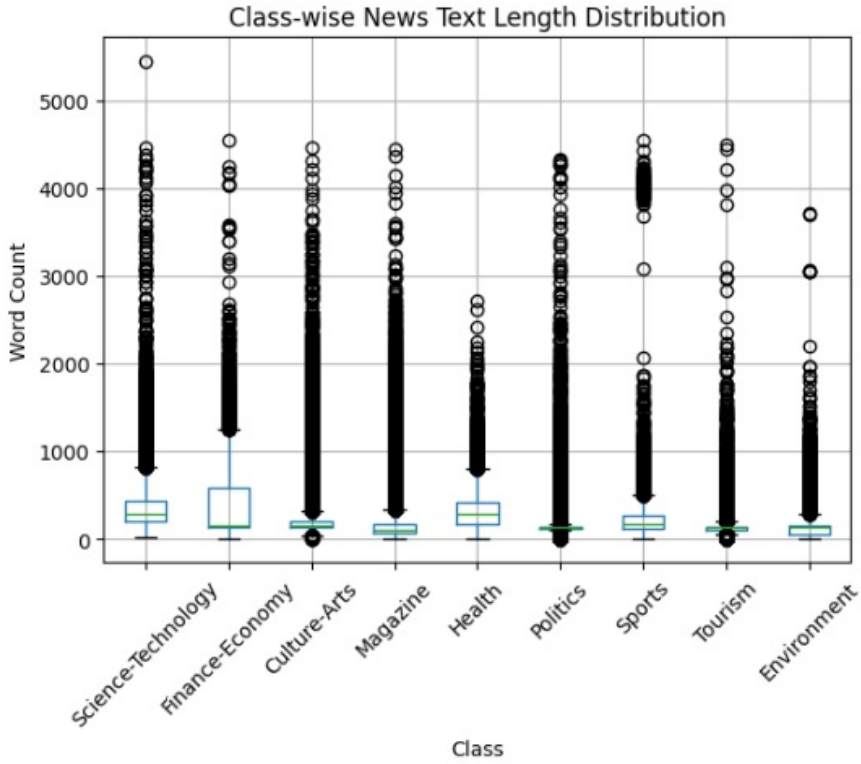


Fig. 1 Text Length Statistics by Class

This data visually supports the content diversity and labeling approach of the dataset.

4 Experimental Results with Deep Learning Models

5 Mathematical Background

This section presents the mathematical foundations of the BERT and BiLSTM–GloVe models employed in this study.

5.1 BERT Model

BERT is based on the Transformer encoder architecture and relies on the self-attention mechanism to model bidirectional contextual relationships within a sequence (Devlin et al. (2019)), (Gardazi et al. (2025)). Let an input sequence be defined as:

$$X = [x_1, x_2, \ldots, x_n], \tag{1}$$

where x_i denotes the i-th token in the sequence.

Table 2 Sample News Text and Its Class Label

Text	Class
"Birkaç yıl öncesine kadar çok sayıda balıkçıl kuşu, güpegündüz gelip burada barınıyorlardı. Bölgede hızla yayılan inşaatlardan dolayı kuşlar gelmez oldu. İnşaat ve makineleşmenin bulunduğu yerde doğal hayat sürmez. Kırklar Dağı gibi Hevsel'de de yapılaşmaya gidilecek gibi görünüyor. Fotoğraflayarak buraları ölümsüzleştiriyorsunuz, ama insan eliyle de bir yok ediliş var. Buna bir siteminiz var mı? Aslında ben sitemimi fotoğraflarla dile getiriyorum. En son çektiğim kızıl tilki ilk defa Hevsel bahçelerinde belgelendi. Sosyal medyadan çok olumlu tepkiler geldi. Bu hayvanların korunması gerektiği noktasında ortak bir bilinç oluştu. Fotoğrafını çektiğiniz o tilki hala yaşıyor mu? Şubat ayında bir tanesi öldürülmüştü. Onu fotoğraflamıştım. Haziran ayında bir yuva keşfettim. İçinde 3 yavru vardı. İki hafta sabah akşam onları izliyordum. Her anlarını fotoğrafladım. Kızıl tilkileri o şekilde belgelemek, o anları ölümsüzleştirmek benim için büyük bir mutluluktu. Bundan sonra ne tür projeleriniz var? Proje geliştirmekte sıkıntı çekiyoruz."	Arts–Culture

Each token embedding is obtained by summing token, segment, and positional embeddings (Li et al. (2025)):

$$e_i = e_i^{\text{token}} + e_i^{\text{segment}} + e_i^{\text{position}}. \tag{2}$$

The scaled dot-product self-attention mechanism is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors.

For text classification tasks, the contextual representation of the special `[CLS]` token is used:

$$\hat{y} = \text{softmax}(W h_{\text{CLS}} + b), \tag{4}$$

where h_{CLS} represents the output vector corresponding to the `[CLS]` token.

The model is trained by minimizing the cross-entropy loss function:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{5}$$

where C is the number of classes and y_c denotes the ground-truth label.

5.2 BiLSTM-GloVe Model

GloVe embeddings are learned based on global word co-occurrence statistics and are defined as (Pennington et al. (2014), Branwen and Emigawaty (2025)):

$$w_i^T \tilde{w}_j + b_i + \tilde{b}_j = \log(X_{ij}), \tag{6}$$

where X_{ij} represents the co-occurrence frequency of words i and j.

An LSTM unit is defined by the following gating equations (Schuster and Paliwal (1997)):

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{7}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{8}$$
$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c), \tag{9}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{10}$$
$$h_t = o_t \odot \tanh(c_t), \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and $\odot$ represents element-wise multiplication.

In the bidirectional setting, forward and backward hidden states are concatenated:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]. \tag{12}$$

Finally, text classification is performed using a softmax layer:

$$\hat{y} = \text{softmax}(Wh + b). \tag{13}$$

6 Experimental Results

The experimental results show that transformer-based models such as BERT achieve the highest performance on the TRNEWS-2025 dataset. Classical recurrent neural network architectures combined with pretrained word embeddings also provide competitive results, demonstrating the suitability of the dataset for both modern and traditional NLP models. When analyzing the accuracy and loss graphs presented, it is observed that the model's validation performance is approximately 86%. For a nine-class Turkish news dataset with high domain diversity, this value indicates strong and satisfactory performance. This result demonstrates that the TRNEWS-2025 dataset can be effectively used with transformer-based language models. However, it is observed that the training accuracy and training loss curves do not reflect the expected typical learning behavior based on epochs. Specifically, the decline in training accuracy and the training loss approaching zero stem from the model being trained with a limited number of epochs or the training process not being monitored in detail on an epoch basis. Therefore, for the BERT model, validation performance was used as the basis for evaluation instead of epoch-based convergence analysis. In dataset papers, the goal is to demonstrate the dataset's compatibility and discriminative power with different model types rather than reporting the highest performance. In this context, the obtained validation accuracy value demonstrates that the TRNEWS-2025 dataset provides a suitable and reliable benchmark environment for modern language models.

Table 3 shows a comparison of the performance and computational costs of the BERT and Bi-LSTM + GloVe models in the TRNEWS-2025 dataset. Bi-LSTM + GloVe has a significantly lower training and testing time and more accurate results,

Algorithm 1 News Classification Pipeline on TRNEWS-2025 Using BERT and BiLSTM–GloVe

Require: TRNEWS-2025 texts $D = \{d_i\}$, labels $Y = \{y_i\}$; model type $m \in \{\text{BERT}, \text{BiLSTM–GloVe}\}$

Ensure: Trained model θ_m and test metrics $(Acc, P, R, F1)_{test}$

Data Preparation and Preprocessing
$D \Leftarrow$ load texts
$Y \Leftarrow$ load labels
$D \Leftarrow$ normalize texts
$(D_{train}, Y_{train}), (D_{val}, Y_{val}), (D_{test}, Y_{test}) \Leftarrow$ split dataset
Model-Specific Input Representation
if $m =$ BERT **then**
    $(input_ids, attention_mask) \Leftarrow \text{Tokenizer}_{\text{BERT}}(D)$
    $(input_ids, attention_mask) \Leftarrow$ pad/truncate to length L
else if $m =$ BiLSTM–GloVe **then**
    $V \Leftarrow \text{Vocab}(D_{train})$
    $S \Leftarrow \text{Sequence}(D, V)$
    $S \Leftarrow$ pad/truncate to length L
end if
Training
$\theta_m \Leftarrow$ initialize parameters
$e \Leftarrow 1$
while $e \leq E$ **do**
    for all mini-batch $b \in \mathcal{B}(D_{train})$ **do**
        compute forward pass on b
        $\mathcal{L} \Leftarrow$ compute loss
        $\theta_m \Leftarrow$ update parameters (Adam/AdamW)
    end for
    validate on (D_{val}, Y_{val}) and save best / early-stop
    $e \Leftarrow e + 1$
end while
Testing and Reporting
$(Acc, P, R, F1)_{test} \Leftarrow$ evaluate on (D_{test}, Y_{test})
return θ_m, $(Acc, P, R, F1)_{test}$

while BERT offers more balanced overall performance, especially when dealing with class imbalance.

Figure 2 presents the loss curves for the BERT and BiLSTM–GloVe models. BERT shows a rapid decline in training loss, while validation loss remains nearly constant, which indicates limited generalization within a small number of epochs. In comparison, the BiLSTM–GloVe model shows a gradual and consistent decline in both training and validation losses, which indicates a more stable learning process and better generalization ability.

Figure 3 shows a comparison of the confusion matrices for the BERT and BiLSTM–GloVe models on the TRNEWS-2025 dataset. Both models obtain high

Table 3 Performance comparison of BERT and Bi-LSTM + GloVe models on the TRNEWS-2025 dataset

Model	LR	Opt.	Ep.	Loss	Acc.	Prec.	Rec.	F1	Train	Test
BERT	–	AdamW	–	0.421	0.869	0.797	0.729	0.753	6710	611
Bi-LSTM + GloVe	0.001	RMS	8	0.263	0.916	–	–	–	1207	27

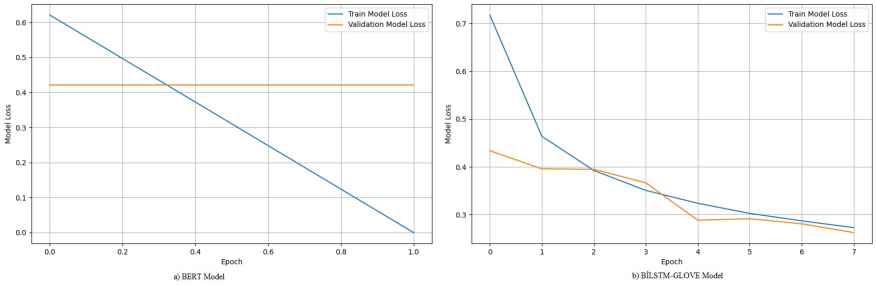


Fig. 2 Loss curves for the BERT and BiLSTM–GloVe models

classification accuracy for majority classes such as Magazine, Sports, and Politics. However, BERT shows higher confusion among semantically overlapping categories, particularly in the Environment class, which is frequently misclassified as Politics or Finance-Economy. In contrast, the BiLSTM–GloVe model shows a more balanced error distribution among medium-sized classes and, despite its lower contextual modeling capacity, demonstrates better stability under class imbalance.

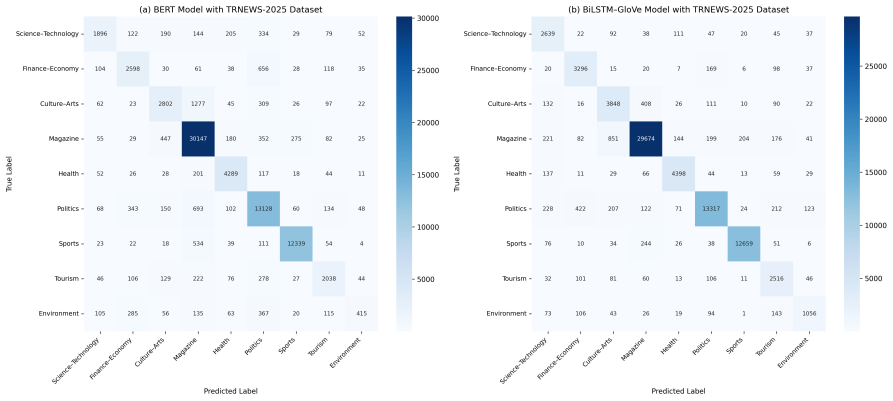


Fig. 3 Confusion Matrices for BERT and BiLSTM-GLOVE Models

7 Conclusion

In this study, the usability and distinguishability of the TRNEWS-2025 labeled Turkish news dataset were evaluated using two different deep learning approaches. In the experimental analyses, BERT, a modern transformer-based model, was compared with

BiLSTM+GloVe, a classical deep learning model. The results obtained show that both models can work successfully on the TRNEWS-2025 dataset. The BERT model achieved high classification performance, particularly in the Magazine, Sports, Health, and Politics classes, by effectively using context-based language models. This result shows that the dataset provides a suitable benchmark for transformer-based models. The BiLSTM+GloVe model, on the other hand, demonstrated a stable learning process thanks to the close correlation between the training and validation metrics. The balanced increase in accuracy, precision, recall, and F1-score values indicates that the model can generalize without overfitting. This demonstrates that the TRNEWS-2025 dataset can be effectively used with classical deep learning architectures. Class-based analyses and confusion matrices reveal similar class confusions in both models. The confusions between the Culture–Arts and Magazine classes and the Environment and Politics/Economy classes, in particular, stem from natural semantic overlaps in news content. This finding indicates that the dataset reflects real-world news language rather than an artificially separated structure. In conclusion, the experimental results obtained demonstrate that the TRNEWS-2025 dataset is compatible with both modern transformer-based and classical deep learning models and can be used as a reliable reference dataset for Turkish Natural Language Processing research.

Acknowledgements. TRNEWS-2025 dataset was compiled with the contributions of students and under academic supervision as part of the BIM 432 – Natural Language Processing course conducted in the Computer Engineering and Software Engineering departments at Istanbul Sabahattin Zaim University during the 2023-2024 Spring academic year, was subsequently improved, organized, and expanded for research purposes. I would also thank my student Alper Karaca for his help in preparing the data set.

Declarations

- Funding The author has not received any financial support for this research.
- Conflict of interest/Competing interests Author declares no conflict of interest
- Ethics approval and consent to participate Ethical approval was not required for this paper.
- Consent for publication Not applicable.
- Data availability The TRNEWS-2025 dataset used in this study is publicly available via IEEE DataPort at: https://doi.org/10.21227/vasv-dj22
- Materials availability All materials generated or analyzed during this study are included in this published article or are available from the corresponding author upon reasonable request.
- Code availability The source code used in this study is available from the corresponding author upon reasonable request.
- Author contribution The author conceptualized the study, designed the methodology, developed the dataset, implemented the models, analyzed the results, and wrote the manuscript.

References

Akbasli IT, Birbilen AZ, Teksam O (2025) Leveraging large language models to mimic domain expert labeling in unstructured text-based electronic healthcare records in non-english languages. BMC Medical Informatics and Decision Making 25(1):154

Bayrak S (2025) Trnews-2025: A multiclass turkish news text dataset. IEEE DataPort, https://doi.org/10.21227/vasv-dj22

Bayram MA, Fincan AA, Gümüş AS, et al (2025) Tokenization standards for linguistic integrity: Turkish as a benchmark. arXiv preprint arXiv:2502.07057

Baysan MS, Güngör T (2025) Tr-mteb: A comprehensive benchmark and embedding model suite for turkish sentence representations. In: Findings of the Association for Computational Linguistics: EMNLP 2025, pp 8867–8887

Boudribila A, Tajer A, Boulghasoul Z (2025) Intelligent extraction of manufacturing system components from natural language using transformer models. International Journal of Production Research 63(24):10608–10632

Branwen D, Emigawaty E (2025) Lightweight BiLSTM-attention model using GloVe for multi-class mental health classification on reddit. Journal of Applied Informatics and Computing 9(5):2899–2911

Devlin J, Chang MW, Lee K, et al (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, pp 4171–4186

Ezerceli Ö, Hussieni ME, Taş S, et al (2025) Turkcolbert: A benchmark of dense and late-interaction models for turkish information retrieval. arXiv preprint arXiv:2511.16528

Gardazi NM, Daud A, Malik MK, et al (2025) Bert applications in natural language processing: a review. Artificial Intelligence Review 58(6)

Kayakuş M, Açıkgöz FY (2022) Classification of news texts by categories using machine learning methods. Alphanumeric Journal 10(2):155–166

Khan N, Ab Rahman AAH, Rizvi SM, et al (2025) Systematic literature review of machine learning models and applications for text recognition. IEEE Access

Kılınç D, Özçift A, Bozyigit F, et al (2017) Ttc-3600: A new benchmark dataset for turkish text categorization. Journal of Information Science 43(2):174–185

Li S, Liu K, Chen X (2025) A context-aware personalized recommendation framework integrating user clustering and BERT-based sentiment analysis. Journal of Computer, Signal, and System Research 2(6):100–108

Oğul İÜ, Soygazi F, Bostanoğlu BE (2025) Turkmednli: a turkish medical natural language inference dataset through large language model based translation. PeerJ Computer Science 11:e2662

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of EMNLP 2014, pp 1532–1543

Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45(11):2673–2681

Türk U, Atmaca F, Özateş ŞB, et al (2022) Resources for turkish dependency parsing: Introducing the boun treebank and the boat annotation tool. Language Resources and Evaluation 56(1):259–307

Yu X, Wang J (2025) Sentiment analysis of classical chinese literature: An unsupervised deep learning model with bert and graph attention networks. PLOS ONE 20(9):e0330919

Yüksel A, Köksal A, Senel LK, et al (2024) Turkishmmlu: Measuring massive multitask language understanding in turkish. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp 7035–7055

Zorarpaci E (2025) An analysis of densely connected convolutional networks integrated with word2vec embedding for turkish news classification. In: 2025 10th International Conference on Computer Science and Engineering (UBMK). IEEE, pp 293–298