Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

End-to-End Fine-Tuning of DeBERTa-Base for Stance Detection

Nabil Daffa As'ad Saputra [1], Muljono [1], Abdul Karim [2], and De Rosal Ignatius Moses Setiadi [3,*]

[1] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia; e-mail : nabil.asad678@gmail.com; muljono@dsn.dinus.ac.id
[2] Cerebrovascular Disease Research Center and Department of Artificial Intelligence Convergence, Hallym University, Chuncheon 24252, South Korea; e-mail : abdullkarim@korea.ac.kr
[3] Research Group for Quantum Computing and Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia; e-mail : moses@dsn.dinus.ac.id
* Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: Stance detection plays an important role in contemporary news analysis by identifying the argumentative relationship between a claim and its associated textual context. In the era of algorithm-driven media, news articles often convey implicit support, opposition, or neutral discussion of specific claims, making stance analysis essential for detecting media bias and researching misinformation. However, accurately modeling such relations remains challenging due to long document lengths, implicit stance expressions, and complex discourse structures. This study evaluates an end-to-end Transformer-based stance detection approach that fine-tunes the DeBERTa-base language model on news text using the Fake News Challenge Stage 1 (FNC-1) dataset under a stance-relevant formulation. The proposed framework updates all parameters of the pre-trained model directly during training, avoiding hand-crafted feature engineering and auxiliary classifiers. Claim–context pairs are jointly encoded and formulated as a three-class stance classification task (agree, disagree, discuss), following the exclusion of unrelated instances to focus on argumentative relations. To ensure robust evaluation under class imbalance, model performance is assessed on a held-out test set using standard classification metrics. Experimental results on the test data show that the proposed approach achieves 96.28% accuracy and 96.23% F1-score, indicating balanced precision–recall performance across stance categories. These findings suggest that a carefully configured end-to-end fine-tuning strategy based on DeBERTa-base is effective for capturing argumentative relations in news text within a three-class stance-relevant setting, providing a reliable and reproducible solution for document-level stance detection without relying on complex architectural modifications or feature engineering.

Keywords: DeBERTa; End-to-End Learning; Fake News Detection; Natural Language Processing; News Analysis; Stance Detection; Text Classification; Transformer Fine-Tuning.

Received: October, 10th 2025
Revised: February, 5th 2026
Accepted: February, 6th 2026
Published: February, 6th 2026
Curr. Ver.: February, 6th 2026

1. Introduction

The increasingly rapid and massive flow of information in the digital era has transformed how people consume news, while simultaneously increasing the risk of misinformation and disinformation. The Reuters Institute Digital News Report indicates that more than 60% of internet users access news through social media and algorithm-driven platforms, where content visibility is strongly influenced by interaction metrics such as clicks, comments, and views [1]. These mechanisms encourage content producers to adjust news narratives to align with the majority audience's preferences to maximize engagement, shaping a text's stance on an issue not solely by facts but also by economic and algorithmic pressures. As a result, argumentative bias is often embedded in language that appears neutral, particularly in news coverage of sensitive social and political issues [2], [3].

In this context, information disseminated through online media must be evaluated not only in terms of factual accuracy but also with respect to the stance constructed by the text

toward a given claim. Stance detection has therefore become an important component of modern news analysis systems, as it enables the identification of whether a text supports, opposes, or merely discusses a particular statement [4], [5]. Unlike sentiment analysis, which focuses on emotional expression, stance detection emphasizes argumentative relations and a text's position on an issue, making it more relevant for bias detection, public opinion analysis, and monitoring narrative dynamics in algorithm-driven media ecosystems.

Early approaches to text mining, including stance detection, evolved from traditional text classification methods relying on statistical and linguistic feature engineering, such as Bag-of-Words, n-grams, and TF-IDF. These methods represent text as word-frequency vectors without considering sequential context or semantic relations among tokens [6]. Such representations were combined with classification algorithms, including Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression [7]–[11], which are relatively effective for small datasets and short texts. However, they exhibit limitations when applied to long documents with complex argumentative structures and implicitly expressed stances. The primary limitation of these approaches lies in their inability to capture long-range dependencies and contextual meaning that often determine a text's position toward a claim.

Advances in deep learning introduced neural models based on word embeddings, such as Word2Vec and GloVe, enabling continuous semantic representations of words, followed by Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) architectures that explicitly model word sequences [12]–[15]. While these approaches improved stance detection performance compared to statistical methods, they still face challenges related to training efficiency and gradient stability when applied to long texts with complex argumentative structures [16]. Since the introduction of the Transformer architecture in 2017, natural language processing paradigms have gradually shifted from sequence-based models to self-attention-based approaches. This shift became more pronounced after 2020, when Transformer-only fine-tuning emerged as the dominant approach across various NLP tasks. Transformer variants such as BERT, RoBERTa, ALBERT, and DeBERTa have since demonstrated consistent performance improvements in text classification, text analysis, and stance detection, particularly for long texts with implicit argumentative relations, outperforming feature-based and conventional RNN approaches [7], [17]–[19].

A key advantage of Transformer architectures in NLP, particularly for stance detection, lies in their ability to model global relationships among tokens in parallel, enabling more comprehensive learning of semantic and argumentative relations in long texts compared to locally sequential approaches. The self-attention mechanism allows models to capture long-range dependencies that often determine a text's stance toward a claim, especially when support or opposition is expressed implicitly through complex discourse structures [20], [21]. DeBERTa further strengthens this capability through its disentangled attention mechanism, which separates content and positional representations, enabling more precise learning of semantic interactions among tokens without distortion from additive positional information, as in earlier Transformer models [22], [23]. These characteristics make DeBERTa more stable in capturing argumentative nuances in long texts with non-linear discourse structures, which represent a major challenge in stance detection tasks [24], [25].

In addition to architectural characteristics, Transformer performance in stance classification tasks is strongly influenced by training configuration strategies adopted during fine-tuning. Prior studies have shown that optimizers designed for pre-trained models, conservative learning rate settings, and adaptive learning rate scheduling play an important role in maintaining gradient stability and preventing degradation of semantic representations during training [26], [27]. A combination of architectural strengths and appropriate training configurations has established modern Transformers as a reliable approach for modeling argumentative relations in news text.

In this context, this study evaluates an end-to-end Transformer fine-tuning approach using DeBERTa-base for stance detection in news text, adopting a stance-relevant formulation that emphasizes argumentative relations between claims and textual context rather than topic relevance. Unlike feature-engineered or complex hybrid architectures, all parameters of the pre-trained model are updated directly during training, allowing contextual representations to fully adapt to the argumentative relations between claims and contexts within the news domain. The main contributions of this study are summarized as follows:

- Evaluating an end-to-end fine-tuning approach based on DeBERTa-base for stance detection on news text, without relying on additional feature engineering or hybrid architectures.
- Providing stable and balanced performance evaluation through the application of stratified cross-validation on datasets with imbalanced class distributions, using standard classification metrics.
- Offering empirical analysis of model prediction behavior, including confusion matrices and prediction probability distributions, to support the interpretation of stance classification results.

The remainder of this paper is organized as follows. Section 2 presents the theoretical background and related work on stance detection and Transformer models. Section 3 describes the proposed methodology, including the pipeline, model architecture, and training strategy. Section 4 reports experimental results and discussion, while Section 5 concludes the paper by summarizing the main findings and outlining limitations and directions for future research.

2. Background and Related Work

2.1. Theoretical Background

2.1.1. Stance Detection

Stance detection is a classification task in natural language processing that aims to determine a text's stance toward a given claim or topic. It has been widely studied in the contexts of news analysis, public opinion mining, and misinformation detection. Early studies define stance as an argumentative relationship between a claim and its surrounding context, in which the text may express support, opposition, or neutrality with respect to the claim. This formulation has been adopted in several benchmark datasets and shared tasks, including the Fake News Challenge (FNC) [28]–[30] and SemEval stance detection tasks. [31]–[33]. Formally, stance detection can be modeled as a mapping function:

$$f: (c, x) \rightarrow y, \quad y \in \{y_1, y_2, \ldots, y_k\} \tag{1}$$

where c denotes the claim (e.g., a headline), x represents the contextual text (e.g., an article body), and y is the predicted stance label drawn from a predefined set of classes. This formulation has been consistently used across machine learning and deep learning approaches, including neural and Transformer-based models, to capture the semantic relationship between claims and contexts.

Unlike sentiment analysis, which focuses on the emotional polarity of a single text, stance detection explicitly models inter-text argumentative relations. As a result, stance detection is particularly relevant to fake news detection and media bias analysis, where the stance toward a claim is often expressed implicitly rather than through overt emotional language. Prior studies identify several key challenges in stance detection, including implicit stance expression, long-range dependencies in lengthy texts, and linguistic ambiguity, especially in news articles written in a neutral tone but conveying subtle argumentative bias.

These characteristics are illustrated in Figure 1, which presents a conceptual diagram of stance detection as a relational task between a claim and its contextual evidence. The figure emphasizes that stance labels are not determined by isolated texts but by their semantic and argumentative interaction, reflecting the core structure of benchmark datasets such as FNC-1, where headlines function as claims and article bodies provide contextual grounding.

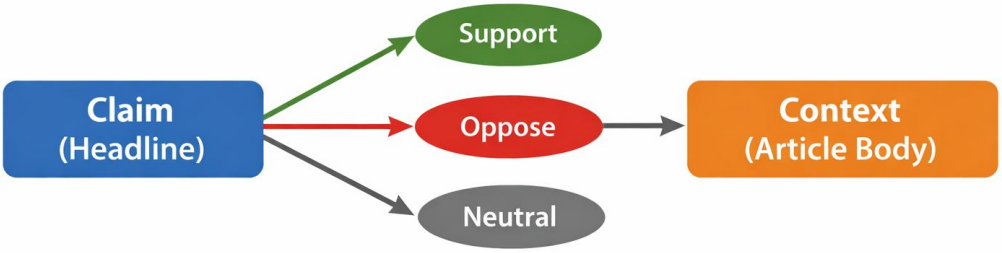

Figure 1. Conceptual diagram of stance detection with the relationship between claims and context.

2.1.2. Transformer-based Language Models

Recent advances in stance detection have been driven by Transformer-based representation learning, which has demonstrated substantial improvements over traditional RNN- and CNN-based architectures in a wide range of text classification tasks. The key innovation of the Transformer architecture lies in the self-attention mechanism, which enables the model to capture global dependencies among tokens without relying on sequential processing [34]–[36]. Mathematically, self-attention is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\intercal}}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k adalah denotes the dimensionality of the key vectors. This formulation allows each token to attend to all other tokens based on contextual relevance dynamically.

In stance detection, self-attention enables models to directly model semantic interactions between claims and their supporting or opposing arguments, even when these elements are distributed across distant segments of the text. Empirical studies have shown that Transformer-based models such as BERT and RoBERTa consistently outperform traditional feature-based and recurrent models, particularly on tasks involving long documents and complex argumentative structures.

The role of self-attention in modeling claim–context interactions is illustrated in Figure 2, which depicts how tokens from the claim and the contextual text jointly interact within a shared representation space. This visualization highlights the ability of Transformer models to integrate information across the entire input sequence, thereby addressing the long-range dependency challenges inherent in stance detection.

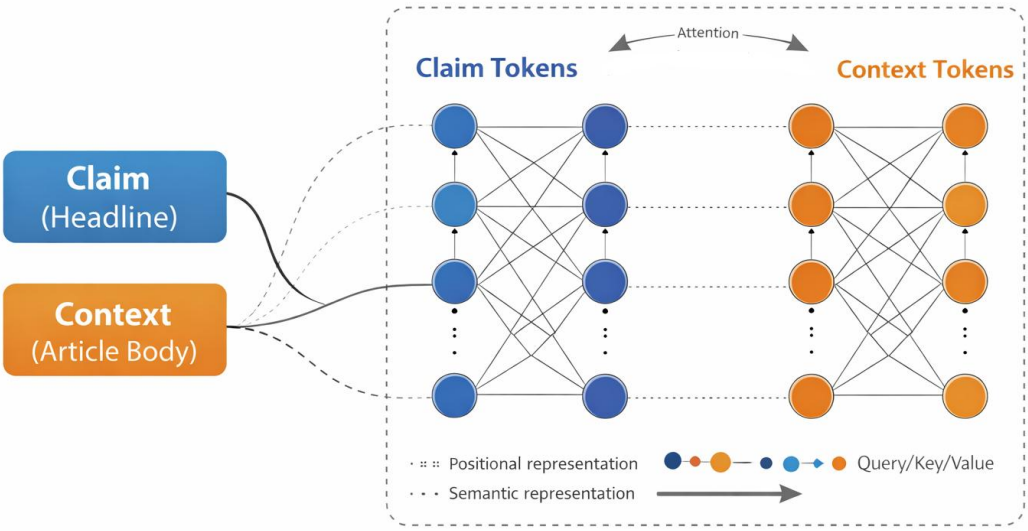

Figure 2. Illustration of the self-attention mechanism in the Transformer and the interaction of claims and context.

2.1.3. DeBERTa and Disentangled Attention

Further developments in Transformer architectures have led to the introduction of DeBERTa, which incorporates a disentangled attention mechanism designed to improve contextual representation learning. Unlike conventional Transformers that combine token content and positional information additively, DeBERTa explicitly separates these components, enabling a more precise modeling of semantic and positional relationships. Conceptually, the attention score between tokens in DeBERTa is computed by independently considering content and relative position representations:

$$A_{i,j} = h_i^{\intercal} W_{cc} h_j + h_i^{\intercal} W_{cp} p_{i,j} + p_{i,j}^{\intercal} W_{pc} h_j \tag{3}$$

where h_i and h_j denote token content representations, and $p_{i,j}$ represents the relative positional embedding between tokens i and j. This disentanglement allows the model to preserve positional sensitivity while maintaining robust semantic representations.

The conceptual difference between conventional self-attention and disentangled attention is illustrated in Figure 3. While conventional self-attention merges content and positional cues into a single representation, disentangled attention models separate these aspects, resulting in improved stability for long texts and non-linear discourse structures. Such characteristics are particularly relevant for stance detection, where the argumentative relationship between a claim and its context may depend on relative position and discourse structure rather than local token proximity alone [22], [24], [25].

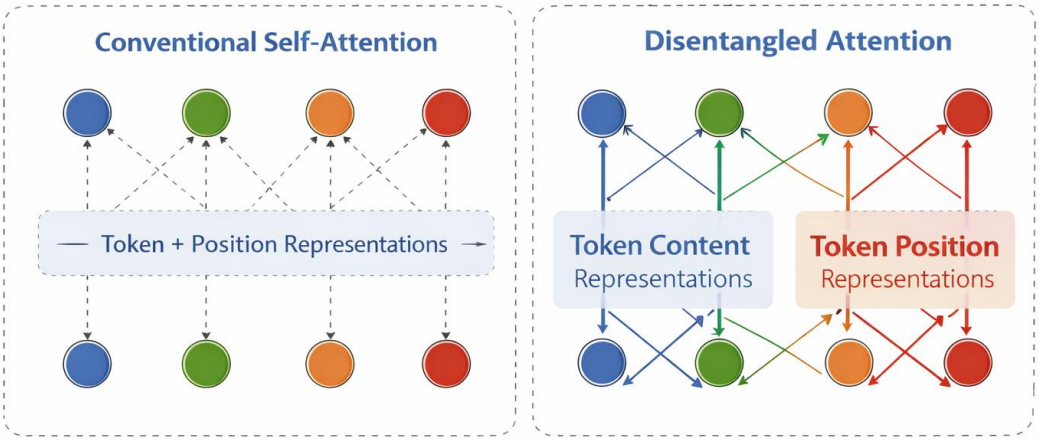

Figure 3. A conceptual comparison illustration of conventional self-attention and disentangled attention.

2.1.4. Stance Classification Formulation

In most Transformer-based stance detection frameworks, a global representation of the input text is obtained from the special $[CLS]$ token produced by the encoder. This representation is then projected into the stance label space using a linear transformation followed by a softmax function:

$$z = Wh_{[CLS]} + b \tag{4}$$

$$P(y = k|x) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \tag{5}$$

where $h_{[CLS]}$ denotes the global contextual representation, K is the number of stance classes, and $P(y = k|x)$ is the predicted probability for class k. Model parameters are optimized by minimizing the cross-entropy loss:

$$\mathcal{L} = -\sum_{k=1}^{K} y_k \log P(y_k) \tag{6}$$

This formulation provides a consistent and widely adopted theoretical foundation for evaluating Transformer-based stance detection models in an end-to-end learning setting, without reliance on handcrafted features or external classifiers.

2.2. Related Works

Stance detection has been extensively studied in the context of fake news analysis, where determining the relationship between a claim and its supporting or opposing evidence is a central challenge. Among several benchmark datasets proposed in the literature, the Fake News Challenge Stage 1 dataset has become one of the most widely adopted testbeds for evaluating document-level stance detection methods. Prior research on this benchmark reflects the broader evolution of stance detection approaches, ranging from feature-based

machine learning models to neural architectures and, more recently, Transformer-based language models.

Early studies predominantly relied on feature engineering and traditional machine learning techniques. Hanselowski et al. [37] conducted a large-scale retrospective analysis of top-performing systems on this benchmark. They showed that most high-ranking approaches depended heavily on handcrafted lexical features and surface-level similarity measures rather than deep semantic modeling. Their analysis also revealed limitations in the original evaluation metric under severe class imbalance, demonstrating that strong aggregate scores do not necessarily correspond to robust stance understanding, particularly for minority classes. Within this line of work, Altheneyan and Alhadlaq [6] represent a strong and influential machine learning–based baseline on the FNC-1 dataset. They proposed a machine learning–based framework combining extensive text preprocessing, lexical feature extraction (e.g., n-grams and TF-IDF variants), and stacked ensemble classifiers implemented in a distributed setting. Importantly, their approach adopts a stance-relevant dataset formulation by separating unrelated instances and further analyzing the agree, disagree, and discuss categories, thereby focusing on argumentative relations between claims and articles. While their system achieved competitive performance, it remained highly dependent on manual feature engineering and on scalable ML pipelines rather than on end-to-end semantic representations.

Subsequent research explored neural network–based models to better capture semantic interactions between claims and contexts. Abedalla et al. [15] investigated several deep learning architectures, including CNNs, Bi-LSTMs, and attention-based models, treating stance detection as a core subtask within fake news detection. Their results demonstrated that neural sequence models consistently outperform traditional machine learning baselines, while also highlighting persistent challenges related to class imbalance and evaluation protocols. Along similar lines, Conforti et al. [38] reframed stance detection as a cross-level task, explicitly addressing the asymmetry between short claims and long news articles. By employing hierarchical neural architectures with conditional and co-matching attention, their approach demonstrated that sentence-level modeling and document-structure awareness are crucial for capturing long-range semantic relations and subtle stance distinctions.

More recent studies have shifted toward explicit end-to-end semantic modeling, aiming to reduce reliance on handcrafted features. Mohtarami et al. [19] are particularly relevant in this regard, as they proposed an end-to-end stance detection approach based on memory networks that uses iterative inference to model global interactions between headlines and article bodies. Their work demonstrated that directly modeling claim–context relations at the semantic level leads to substantial improvements over feature-based pipelines, especially in capturing implicit and nuanced argumentative cues. This line of research strongly motivated the transition toward representation learning–based stance detection methods.

The latest advances leverage Transformer-based language models and transfer learning. Slovikovskaya and Attardi [39] systematically evaluated fine-tuning strategies using BERT, XLNet, and RoBERTa, comparing feature-based pipelines with fully end-to-end Transformer models. Their findings showed that fine-tuned Transformers substantially outperform earlier approaches on macro-level and class-wise metrics, especially for the challenging disagree class, establishing contextualized language models as the dominant paradigm for stance detection on this benchmark.

Despite these advances, existing studies differ significantly in model complexity, feature dependencies, and evaluation protocols. In particular, strong baselines such as [6] rely on extensive feature engineering, while end-to-end semantic models such as [19] introduce additional architectural components to capture claim–context interactions. Meanwhile, recent Transformer-based studies often emphasize performance gains without explicitly examining whether simpler, fully end-to-end fine-tuning strategies are sufficient. This leaves an open research gap regarding the effectiveness of straightforward, reproducible end-to-end Transformer fine-tuning—without auxiliary features or specialized architectural augmentations—under a stance-relevant formulation of widely used benchmarks. In benchmark-driven stance detection research, the Fake News Challenge Stage 1 (FNC-1) dataset is one of the most commonly adopted evaluation benchmarks. While the original formulation defines four classes, prior studies have shown that the unrelated class primarily reflects topic relevance rather than argumentative stance and dominates evaluation under severe class imbalance. Following this stance-relevant perspective, the present study adopts a three-class setting (agree, disagree, discuss), and all comparisons and conclusions are interpreted within this scoped benchmark.

3. Proposed Method

This section presents the proposed end-to-end stance detection framework, which fine-tunes a DeBERTa-base Transformer model. Following the theoretical formulation and empirical context discussed in Sections 1 and 2, the proposed method focuses on a minimal yet principled design that directly models the semantic relationship between a claim and its contextual evidence without introducing handcrafted features or auxiliary classifiers. An overview of the methodological pipeline is illustrated in Figure 4.

3.1. Methodological Pipeline

The proposed methodological pipeline consists of a sequential workflow encompassing dataset preparation, tokenization, model training, and evaluation, all implemented consistently with the experimental codebase.

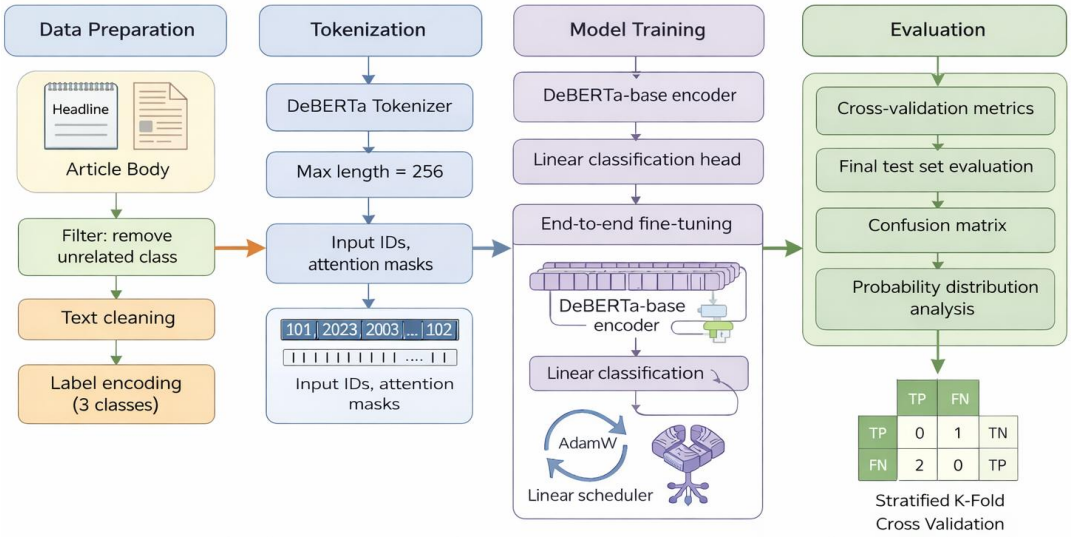

Figure 4. Overview of the methodological pipeline for the proposed stance detection framework, from dataset preparation to final evaluation.

The pipeline begins with dataset preparation, where headline–article pairs are filtered to retain stance-relevant instances and reformulated as a three-class classification problem. The processed texts are then tokenized with the DeBERTa tokenizer and fed to an end-to-end Transformer-based model. Model training is performed using stratified cross-validation to ensure balanced evaluation across stance categories, followed by final testing and comprehensive performance analysis. This pipeline is intentionally designed to emphasize reproducibility and methodological clarity, avoiding additional feature engineering, heuristic rules, or task-specific preprocessing. All pipeline stages are implemented directly in the experimental codebase without manual intervention, ensuring consistency and full replicability of the proposed approach.

3.2. Dataset Preparation and Preprocessing

The experiments are conducted on a benchmark dataset for document-level stance detection, where each instance consists of a claim (headline) and its corresponding contextual text (news article body). In line with common practice, instances labeled as unrelated are excluded to focus on stance-relevant relationships, resulting in a three-class classification problem. Text preprocessing is intentionally lightweight to preserve semantic content. Headline and article body texts are lowercased, cleaned from URLs and non-alphanumeric symbols, and concatenated into a single input sequence. This combined representation allows the model to jointly encode the claim and its context while maintaining their semantic dependencies. Tokenization is performed using the DeBERTa tokenizer with a fixed maximum sequence length of 256 tokens. The tokenizer produces input IDs and attention masks, which serve as the sole inputs to the Transformer encoder.

3.3. Model Architecture

The proposed stance detection framework adopts a Transformer-only, end-to-end architecture based on DeBERTa-base, consisting of a pre-trained DeBERTa encoder followed by a lightweight classification head. The overall architecture and data flow are illustrated in Figure 5. In this architecture, the model's input is a pair of textual segments: a claim (the headline) and its corresponding context (the news article body). These two segments are concatenated into a single input sequence following the standard Transformer input format: $[CLS]$ claim $[SEP]$ context $[SEP]$. The claim represents the statement whose stance is to be determined, while the context provides the supporting, opposing, or neutral evidence. This formulation directly reflects the relational nature of stance detection and is consistent with the task definition introduced in Section 2.

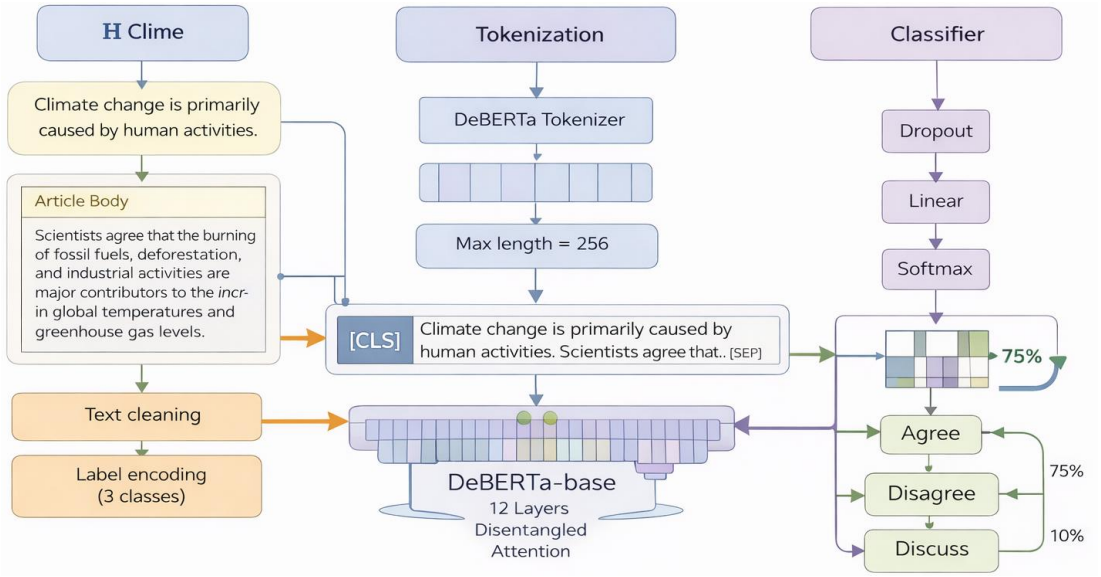

Figure 5. Architecture of the proposed end-to-end DeBERTa-based stance detection model. The contextual representation of the $[CLS]$ token is used for stance classification.

As shown in Figure 5, the combined input sequence is first processed by the DeBERTa tokenizer, which converts the text into token IDs and corresponding attention masks. These tokens are then passed to the DeBERTa-base encoder, which consists of 12 Transformer layers, each with 12 attention heads and a hidden size of 768. Through the disentangled attention mechanism, the encoder models token content and relative positional information separately, enabling more stable and expressive contextual representations, particularly for long, structurally complex texts.

The encoder produces contextualized embeddings for all tokens in the input sequence. Among these, the embedding corresponding to the special $[CLS]$ token from the final encoder layer is treated as a global representation of the claim–context pair. This representation is assumed to summarize the overall semantic and argumentative relationship between the claim and the context, rather than representing either segment in isolation.

The $[CLS]$ representation is then passed through a dropout layer, which serves as a regularization mechanism during training, followed by a fully connected linear layer that maps the 768-dimensional vector into a vector of size K, where K denotes the number of stance classes. In this study, $K = 3$ corresponding to the agree, disagree, and discuss labels. Finally, a softmax function is applied to transform the logits into a probability distribution over stance classes.

It is important to note that the values shown in Figure 5 are conceptual and symbolic, intended solely to visualize the model's data flow and decision process. They do not represent actual prediction outputs or empirical probabilities. The figure aims to clarify how claim and context information is jointly encoded and how the final stance decision is produced, rather than to provide example results. This architecture deliberately avoids additional modules, feature fusion mechanisms, or external classifiers. By relying exclusively on the representational

capacity of the DeBERTa encoder and a minimal classification head, the proposed design emphasizes simplicity, transparency, and reproducibility, while remaining fully aligned with the end-to-end learning paradigm adopted in modern Transformer-based stance detection systems.

3.4. Training Strategy

The proposed stance detection model is trained using an end-to-end fine-tuning strategy, in which all parameters of the pre-trained DeBERTa-base encoder and the classification head are jointly updated. This design choice aims to fully leverage contextual representations learned during large-scale pre-training while avoiding the complexity of feature engineering or hybrid modeling pipelines. To ensure transparency, reproducibility, and fair comparison, all training configurations are fixed across experiments, and no hyperparameter search or tuning procedure is performed. Following standard practice for robust and leakage-free evaluation, the dataset is first split into a held-out test set (20%) and a training pool (80%) using stratified sampling with a fixed random seed. The held-out test set is created once, prior to any model training, and remains completely isolated throughout the experimental process. It is not used for cross-validation, model selection, or training, and is reserved exclusively for final performance evaluation. To estimate performance stability and variability under class imbalance, Stratified K-Fold Cross Validation is applied exclusively to the training pool. In this procedure, the training data are partitioned into five folds with preserved class proportions, and each fold is trained independently. Importantly, each headline–article pair is treated as a unique instance, and no instance appears in more than one split, ensuring that evaluation results reflect genuine generalization rather than overlap between training and validation data. This protocol minimizes the risk of information leakage and enables a reliable assessment of model robustness.

Table 1. Training configuration for the proposed DeBERTa-based stance detection model.

Component	Configuration
Pre-trained model	DeBERTa-base
Fine-tuning strategy	End-to-end (all layers updated)
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Learning rate scheduler	Linear decay with warm-up
Warm-up proportion	10% of total training steps
Gradient clipping	Max norm = 1.0
Batch size	16
Maximum sequence length	256
Epochs (cross-validation)	2 epochs per fold
Epochs (final model)	3 epochs
Cross-validation strategy	Stratified K-Fold
Number of folds	5
Loss function	Cross-entropy loss
Random seed	42

Model optimization is performed using the AdamW optimizer, widely adopted for fine-tuning large pre-trained language models due to its effective weight decay. A conservative learning rate is employed to prevent disruption of the contextual representations learned during pre-training. To further stabilize training, a linear learning rate scheduler with a warm-up phase is applied, gradually increasing the learning rate at the early stage of training before decaying it linearly over the remaining training steps.

Gradient clipping is used to control the magnitude of gradients and improve numerical stability. Training is conducted with a moderate batch size and a fixed maximum input sequence length, balancing computational efficiency and contextual coverage for long news articles. During cross-validation, the number of training epochs per fold is intentionally

limited to provide a conservative estimate of generalization performance while maintaining computational tractability. After cross-validation, a final model is trained from scratch on the full training pool using a slightly longer training schedule, and its performance is evaluated once on the held-out test set following the same preprocessing and configuration. All experiments are conducted with fixed random seeds to guarantee reproducibility. A summary of the training and evaluation configuration used throughout the experiments is reported in Table 1.

3.5. Evaluation Protocol

Model performance is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Metrics are computed both in a weighted manner and per class to account for class imbalance. Evaluation is performed at two levels: (i) across all validation folds to assess stability and generalization, and (ii) on a held-out test set to provide an unbiased estimate of final model performance. In addition to aggregate metrics, confusion matrices and predicted probability distributions are analyzed to provide deeper insights into model behavior across stance categories. This evaluation protocol emphasizes balanced and interpretable performance assessment, ensuring that improvements are not driven solely by dominant classes.

4. Results and Discussion

This section presents the experimental results and discusses the performance characteristics of the proposed DeBERTa-based stance detection framework. The analysis focuses on dataset characteristics, model performance under cross-validation and test conditions, error patterns across stance categories, and comparison with representative prior studies.

4.1. Dataset Characteristics and Class Distribution

The experiments in this study are conducted on the Fake News Challenge Stage 1 (FNC-1) dataset, which is publicly available at https://github.com/FakeNewsChallenge/fnc-1.The dataset consists of headline–article pairs annotated with stance labels that describe the relationship between a claim (headline) and its corresponding news article body.

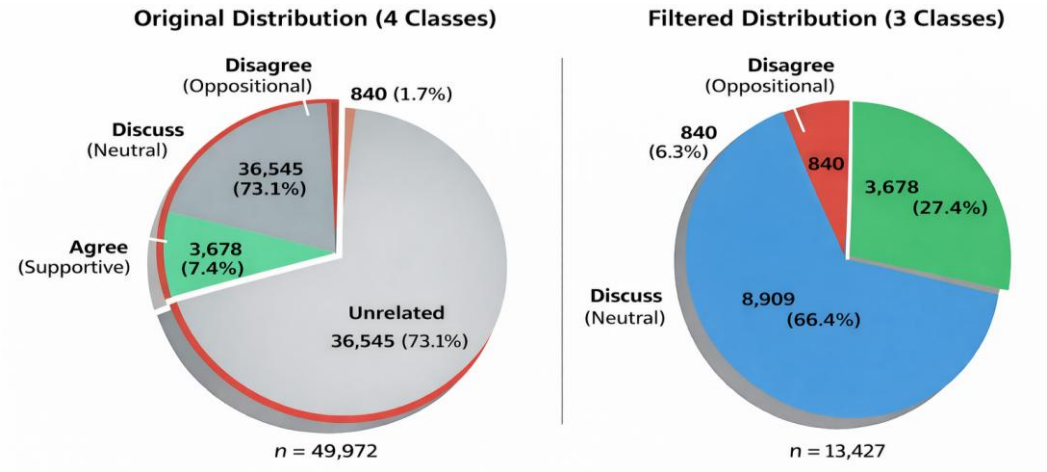

Figure 6. Class distribution of the FNC-1 dataset before and after stance-relevant filtering.

As illustrated in Figure 6 (left), the original dataset exhibits a severe class imbalance. The unrelated class dominates the distribution with 36,545 instances (73.1%), while the disagree class accounts for only 840 instances (1.7%). The remaining samples belong to the discuss class (8,909 instances, 17.8%) and the agree class (3,678 instances, 7.4%). This extreme skew poses a major challenge for stance detection, as high overall accuracy can be achieved by exploiting topic-relatedness rather than learning genuine argumentative relations between claims and articles. Moreover, the overwhelming prevalence of unrelated instances substantially increases the computational cost during training while providing limited information for learning stance-specific representations.

To focus the analysis on stance-relevant interactions and avoid conflating stance classification with topic-relatedness, all instances labeled as unrelated are removed prior to model training and evaluation. This filtering choice simultaneously reduces unnecessary computational overhead and enables a more efficient allocation of model capacity toward learning argumentative relations of interest. The resulting three-class dataset contains 13,427 instances, composed of discuss (66.4%), agree (27.4%), and disagree (6.3%) classes, as shown in Figure 6 (right). This stance-relevant formulation is consistent with prior work that explicitly separates topic relevance from argumentative stance when analyzing the FNC-1 dataset. Although class imbalance remains, the filtered distribution better reflects a realistic stance detection scenario, where neutral or discussion-oriented texts are naturally more frequent than explicit opposition.

The before–and–after comparison in Figure 6 highlights the motivation for this filtering choice and clarifies that all subsequent experiments and comparisons in this study are conducted under the three-class stance-relevant setting. In addition to improving methodological focus, this design choice contributes to a more computationally tractable and reproducible experimental setup, particularly for document-level Transformer fine-tuning. This comparison further motivates the use of Stratified K-Fold Cross Validation in subsequent experiments, ensuring that class proportions are preserved across training and evaluation splits despite the skewed distribution.

4.2. Experimental Setup and Training Overview

To ensure transparency and reproducibility, all experiments are conducted within a fixed, explicitly defined computational setup. Text preprocessing is intentionally lightweight and limited to basic cleaning operations, allowing the model to learn stance-related representations directly from the raw claim–context pairs without task-specific normalization or handcrafted feature augmentation. Model training follows an end-to-end fine-tuning strategy using the AdamW optimizer with a learning rate of 2×10^{-5}, which is commonly used to stabilize gradient updates in pre-trained Transformer models. All parameters of the DeBERTa-base encoder and the classification head are updated jointly, without freezing layers or performing hyperparameter search. To ensure fair evaluation under class imbalance, stratified cross-validation is applied consistently across all experiments.

All training and evaluation procedures are implemented using the HuggingFace Transformers ecosystem with PyTorch Lightning as the training framework, and executed on an NVIDIA Tesla T4 GPU. This controlled experimental environment ensures that the reported results are attributable to the modeling approach rather than variations in hardware configuration or optimization settings. This standardized setup provides a reliable foundation for the quantitative performance analysis presented in the following section.

4.3. Cross-Validation Results and Stability Analysis

Model performance is first evaluated using 5-fold Stratified Cross-Validation to assess robustness and stability in the imbalanced three-class stance detection setting. Quantitative results for each validation fold are summarized in Table 2, using accuracy, precision, recall, and F1-score as evaluation metrics.

Table 2. Cross-validation performance of the proposed DeBERTa-base model across five folds.

Fold	Accuracy	Precision	Recall	F1-Score
Fold 1	0.9311	0.9305	0.9311	0.9307
Fold 2	0.9395	0.9398	0.9395	0.9387
Fold 3	0.9372	0.9381	0.9372	0.9374
Fold 4	0.9353	0.9345	0.9353	0.9344
Fold 5	0.9437	0.9433	0.9437	0.9429
Average ± Std	0.9373 ± 0.0042	0.9372 ± 0.0044	0.9373 ± 0.0042	0.9368 ± 0.0041

As reported in Table 2, the proposed model achieves consistently strong performance across all validation folds, with average accuracy and F1-score around 0.94. The close correspondence among precision, recall, and F1-score indicates balanced classification behavior,

suggesting that the model does not rely disproportionately on the majority class. Performance variance across folds is low, with standard deviations below 0.005 for all metrics. This indicates that the model's predictions are stable and robust across different data partitions, despite the dataset's inherent class imbalance. Minor fluctuations across folds are expected due to differences in the distribution of minority classes, particularly the oppositional category. The cross-validation results demonstrate that the proposed end-to-end DeBERTa-based framework achieves reproducible, reliable performance under stratified evaluation. This stability provides a solid basis for assessing the model's generalization to unseen data, which is further examined using a held-out test set in the following section.

4.4. Held-out Test Set Performance and Class-wise Analysis

Figure 7 compares the average performance from 5-fold stratified cross-validation (Section 4.3) with results on an independent held-out test set. Under cross-validation, the proposed model demonstrates stable performance across all evaluation metrics, with average accuracy and weighted F1-score of approximately 0.94. As discussed previously, this evaluation primarily reflects performance stability and variance across different data partitions, rather than final generalization performance.

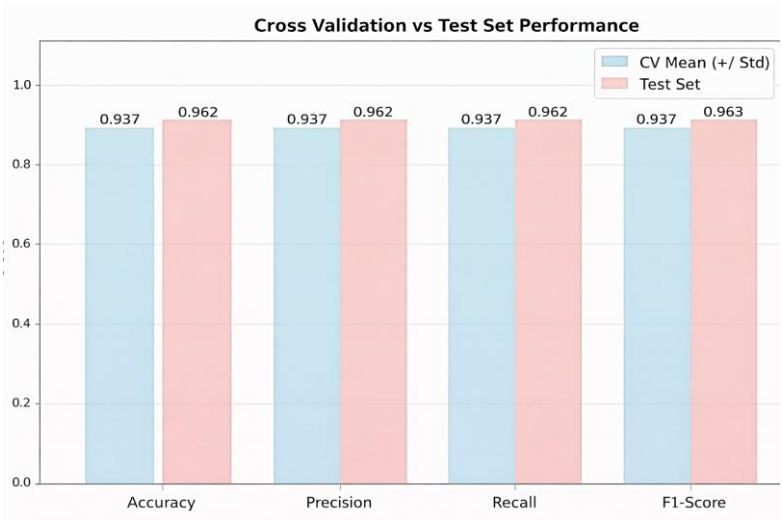

Figure 7. Comparison of cross-validation and test set performance

As illustrated in Figure 7, performance on the held-out test set is consistently higher than the cross-validation averages across accuracy, precision, recall, and F1-score. This difference can be explained by two complementary factors. First, cross-validation typically provides a more conservative estimate of performance because each fold may contain less favorable distributions of minority classes, particularly the Oppositional category[40], [41]. Second, the final model evaluated on the held-out test set is trained on the entire training pool using a slightly longer training schedule (three epochs) compared to the two epochs per fold used during cross-validation, allowing the model to benefit from increased data exposure and additional optimization steps. Such behavior is consistent with established evaluation practices and does not, by itself, indicate overfitting or information leakage when data splits are properly isolated. The observed performance gap between cross-validation and held-out test evaluation remains limited (approximately 2–3%) and is consistent across all reported metrics. According to prior methodological studies, small and systematic differences of this magnitude are generally expected in stratified evaluation settings, especially when final models are trained using more data and slightly longer optimization schedules [41], [42]. The close correspondence between validation and test performance therefore suggests that the learned decision boundaries are stable and transferable to unseen data.

To provide a more detailed assessment of generalization behavior under class imbalance, Table 3 reports precision, recall, and F1-score for each stance category on the held-out test set, together with macro-averaged and weighted-averaged results. While weighted metrics reflect overall performance dominated by the majority classes, macro-level evaluation assigns

equal importance to each stance category and is therefore more informative for assessing minority-class behavior in the three-class stance-relevant setting.

Table 3. Class-wise precision, recall, and F1-score on the held-out test set.

Class	Precision	Recall	F1-score	Support
Oppositional	0.9247	0.8036	0.8599	168
Neutral	0.9766	0.9832	0.9799	1782
Supportive	0.9370	0.9497	0.9433	736
Macro Average	0.9461	0.9122	0.9277	–
Weighted Average	0.9625	0.9628	0.9623	–

As shown in Table 3, the proposed model achieves the strongest performance on the Neutral and Supportive classes, which together account for the majority of instances in the dataset. Performance on the Oppositional class remains comparatively lower due to its limited representation (approximately 6% of the data). Nevertheless, recall for the Oppositional class reaches approximately 80%, indicating reasonable sensitivity to oppositional stances. When interpreted alongside the macro-F1 score of 0.9277, these results indicate that the model achieves a reasonably balanced trade-off across stance categories under severe class imbalance, rather than uniform class-wise performance. Taken together, the evidence from cross-validation stability (Section 4.3), held-out test evaluation, class-wise analysis (Table 3), and the comparison illustrated in Figure 7 demonstrates that the proposed end-to-end DeBERTa-based framework exhibits robust and reproducible behavior under stratified evaluation. The evaluation protocol and observed results align with best practices for assessing Transformer-based text classification models on imbalanced stance detection benchmarks

4.5. Confusion Matrix and Probability Distribution Analysis

To further analyze class-wise performance and prediction behavior, the model's outputs on the test set are examined using confusion matrices and softmax probability distributions, as illustrated in Figures 8 and 9.

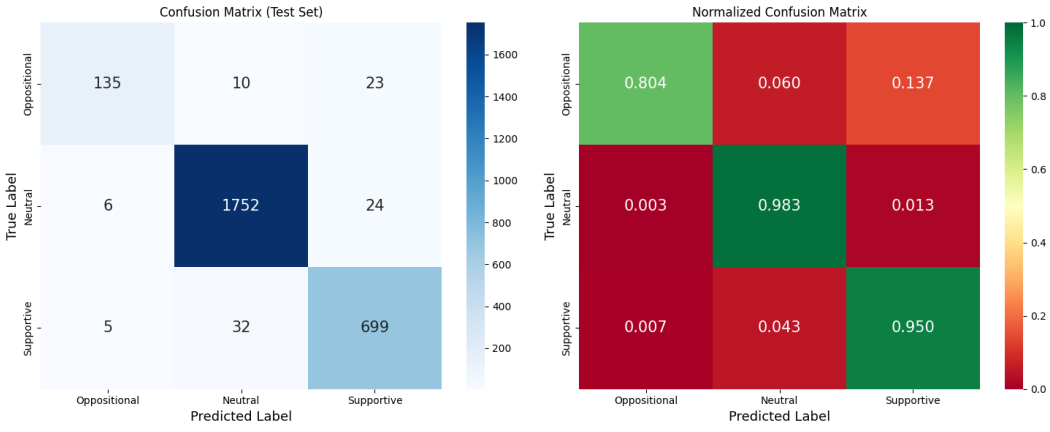

Figure 8. Confusion matrix (**a**) and normalized confusion matrix; (**b**) for the DeBERTa-based stance detection model on the test set.

Figure 8 presents the confusion matrix and its normalized counterpart for the three stance classes: Oppositional, Neutral, and Supportive. The diagonal entries indicate correct classifications, while off-diagonal entries represent misclassifications. As shown in Figure 8(a), the model correctly classifies 1,752 Neutral, 699 Supportive, and 135 Oppositional instances. Misclassifications are relatively limited and primarily occur between semantically adjacent classes. For example, a small number of Oppositional instances are predicted as Neutral (10 cases), and some Supportive instances are misclassified as Neutral (32 cases), reflecting the inherent ambiguity between neutral discussion and weakly polarized stances.

The normalized confusion matrix in Figure 8(b) highlights class-wise recall performance. The Neutral class achieves the highest recall (98.3%), followed by Supportive (95.0%) and Oppositional (80.4%). The comparatively lower recall for the Oppositional class can be attributed to its smaller sample size and the tendency of oppositional language to be expressed implicitly or through nuanced argumentation, which often overlaps with neutral discourse. This observation is consistent with prior findings that stance detection systems struggle most with minority and implicitly expressed opposition classes.

To complement the confusion matrix analysis, Figure 9 presents the distribution of softmax probability scores for each stance class on the test set. The histograms provide insight into the confidence characteristics of the model's predictions across different stance categories. As shown in Figure 9(b), the Neutral class exhibits a strong concentration of predicted probabilities near 1.0, with a mean probability of 0.667. This distribution indicates high model confidence and well-defined decision boundaries for neutral statements, which constitute the majority class in the dataset. In contrast, the Supportive class in Figure 9(c) shows a broader probability distribution with a moderate mean value of 0.278, reflecting a wider range of confidence levels and increased semantic variability in supportive expressions. The Oppositional class, illustrated in Figure 9(a), displays a markedly lower mean probability of 0.055, with most predictions clustered near zero and only a small fraction receiving high confidence scores. This behavior is expected given the relatively limited number of oppositional samples and their frequent semantic overlap with neutral or weakly supportive statements.

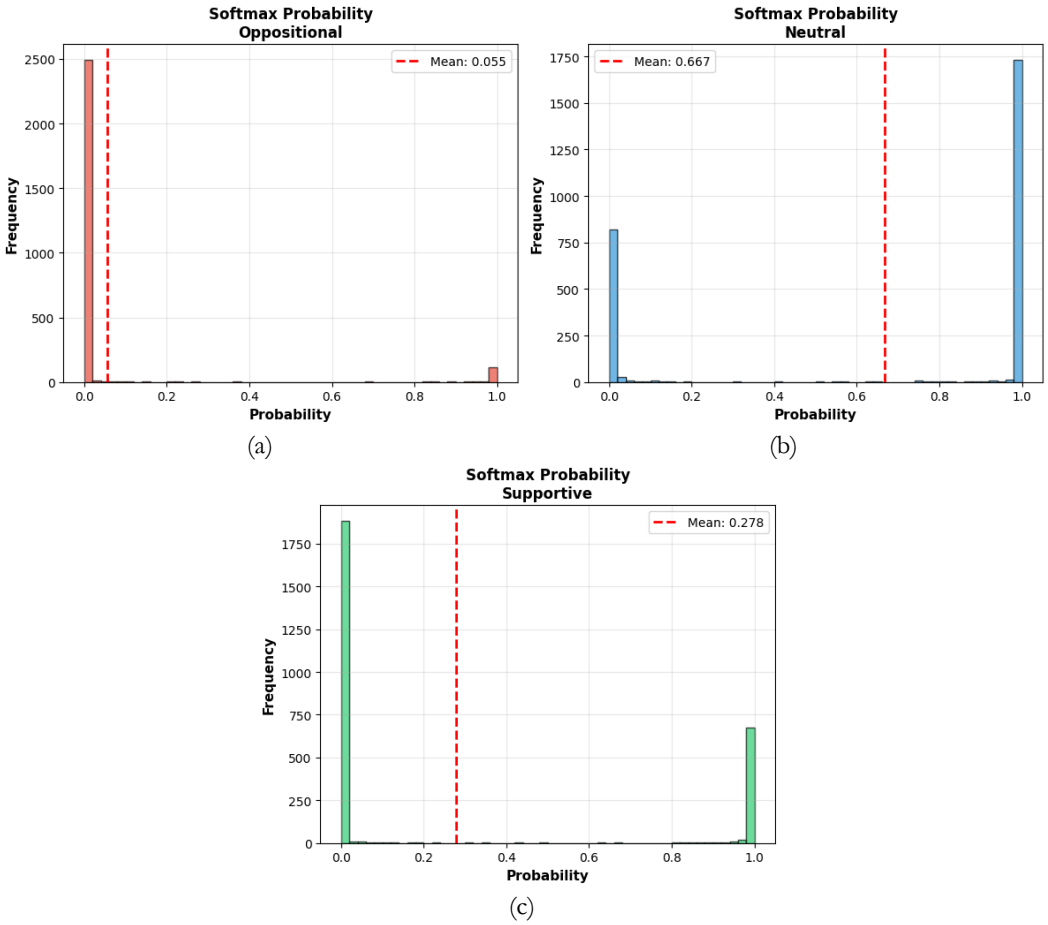

Figure 9. Distribution of softmax probability scores for each stance class on the test set: (**a**) Oppositional; (**b**) Neutral; and (**c**) Supportive.

The probability distributions demonstrate that the model does not produce uniformly flat or ambiguous predictions. Instead, it converges toward confident class assignments when sufficient semantic evidence is present, while appropriately reflecting uncertainty in more challenging cases. When considered jointly with the confusion matrix results shown in Figure 8, these findings indicate that the proposed DeBERTa-based stance detection model

effectively internalizes the semantic structure of claim–context relations and maintains balanced decision-making across stance classes despite inherent dataset imbalance.

4.6. Comparative Evaluation with Prior Studies and Other End-to-End Models

Table 4 places the proposed approach in the context of representative prior studies and baseline models evaluated on the same benchmark. All comparisons are conducted under the same three-class stance-relevant formulation of the FNC-1 dataset and follow a consistent preprocessing and evaluation protocol. Compared to the feature-engineered ensemble approach of Altheneyan and Alhadlaq [6], the proposed DeBERTa-base model achieves substantial improvements across all evaluation metrics, with an absolute gain of approximately +4% in F1-score and +3% in accuracy. This performance gap highlights the limitations of feature-heavy pipelines. It suggests that direct end-to-end contextual representation learning is more effective for capturing implicit stance relations in long news texts.

When compared with the Transformer-based method of Karande et al. [19], which represents a strong stance-related neural baseline, the proposed model still demonstrates consistent gains, achieving higher precision, recall, F1-score, and accuracy. While Karande et al. incorporate stance as part of a broader credibility analysis pipeline rather than as a standalone stance classification task, the observed improvements indicate that end-to-end fine-tuning of modern Transformers can achieve competitive, and often superior, performance in a stance-relevant evaluation setting.

In addition to prior studies, Table 4 also reports results for commonly used Transformer baselines, including BERT, RoBERTa, and ALBERT, which are widely adopted in stance detection and text classification research. These models are fine-tuned under the same experimental conditions as the proposed approach, including identical preprocessing, input formulation, and number of training epochs, ensuring a fair and controlled comparison. Among these baselines, RoBERTa achieves strong performance, reflecting its robust pre-training strategy, while BERT and ALBERT exhibit comparatively lower but stable results.

Table 4. Comparison with prior studies and other end-to-end models.

Ref	Precision	Recall	F1-Score	Accuracy
Altheneyan and Alhadlaq [6]	92.03	92.45	92.25	93.45
Karande et al. [19]	94.81	94.81	95.89	95.85
BERT	93.56	93.71	93.62	93.71
RoBERTa	95.23	95.23	95.22	95.23
AlBERT	92.92	92.93	92.86	92.93
DeBERTa (proposed)	96.25	96.28	96.23	96.28

Notably, the proposed DeBERTa-base model consistently outperforms all Transformer baselines considered, including RoBERTa, across all reported metrics. These gains are achieved without architectural augmentation, feature fusion, or auxiliary training objectives, and are primarily attributed to DeBERTa's improved representational capacity, particularly its disentangled attention mechanism and enhanced contextual encoding. This observation aligns with prior findings that DeBERTa is more effective in modeling long-range semantic and argumentative dependencies, which are critical for document-level stance detection.

From a methodological perspective, these results suggest that model simplicity and performance are not necessarily competing objectives. Rather than introducing additional complexity through ensembles or task-specific modules, the results demonstrate that a carefully configured, end-to-end fine-tuning strategy can fully leverage the strengths of modern Transformer architectures. This is particularly relevant for stance detection, where robustness, reproducibility, and transparency are essential for both research comparability and real-world deployment.

More broadly, the comparison highlights a shift in stance detection research from feature-centric optimization toward representation-centric learning, where the quality of contextual embeddings plays a dominant role. While the proposed approach does not claim universal superiority across all possible stance-detection settings, the consistent improvements observed under a controlled, stance-relevant evaluation protocol indicate that modern Transformer architectures—when fine-tuned in a principled, reproducible manner—are sufficient

to achieve strong, reliable performance on document-level stance-detection benchmarks. This finding supports the broader argument that future research should prioritize transparent end-to-end modeling strategies over increasingly complex feature engineering pipelines.

5. Conclusions

This study evaluated an end-to-end Transformer fine-tuning approach based on DeBERTa-base for stance detection in news text. The experimental results demonstrate that directly updating all parameters of a pre-trained Transformer, without additional feature engineering or hybrid architectures, is sufficient to capture argumentative relations between claims and their contextual evidence. Under stratified evaluation, the proposed model exhibits stable and balanced overall performance, indicating its ability to generalize effectively in the presence of class imbalance and implicitly expressed stances. The findings confirm that a simplified, end-to-end fine-tuning strategy can serve as a reliable alternative to more complex stance detection pipelines. While the model achieves strong overall performance, its effectiveness under class imbalance should be interpreted with caution: the oppositional (disagree) class remains highly underrepresented, and although recall for this class reaches approximately 80%, the results indicate that the approach is reasonably robust given severe class imbalance, rather than fully insensitive to it.

From a methodological standpoint, adopting a three-class stance-relevant formulation also yields a more focused and tractable learning problem. By excluding instances labeled as unrelated, the model is encouraged to concentrate on genuine argumentative relations between claims and contextual evidence, while reducing unnecessary complexity introduced by topic-related but non-argumentative pairs. This design choice not only clarifies the scope of stance analysis but also supports a more efficient and reproducible experimental setup for document-level Transformer fine-tuning, without altering the fundamental nature of the stance detection task. From a practical perspective, the proposed framework contributes to the development of stance detection systems that are transparent, reproducible, and easy to deploy. By emphasizing methodological simplicity and standard evaluation protocols, this study provides a practical baseline that can be readily adopted or extended in related research on media bias analysis and misinformation detection. It is important to note that the reported results are specific to the three-class stance-relevant formulation adopted in this study and should not be directly generalized to the original four-class FNC-1 setting without further investigation.

Nevertheless, this work has several limitations. The evaluation is conducted on a single benchmark dataset and focuses on a three-class stance formulation, which may limit generalizability across domains or more fine-grained stance categories. In addition, the study does not compare different Transformer architectures or investigate the impact of alternative fine-tuning strategies. Future research may extend this work by evaluating cross-domain generalization, incorporating multilingual datasets, or examining parameter-efficient fine-tuning methods to improve further scalability, robustness, and applicability in real-world scenarios.

Author Contributions: Conceptualization: N.D.A.S. and D.R.I.M.S.; Methodology: N.D.A.S. and D.R.I.M.S; Software: N.D.A.S.; Validation: D.R.I.M.S., M. and A.K.; Formal analysis: D.R.I.M.S., M. and A.K.; Investigation: D.R.I.M.S., M. and A.K.; Resources: N.D.A.S.; Data curation: N.D.A.S.; Writing—original draft preparation: N.D.A.S.; Writing—review and editing: D.R.I.M.S., M. and A.K.; Visualization: D.R.I.M.S.; Supervision: D.R.I.M.S.; Project administration: M.; Funding acquisition: Nil. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are publicly available. The experiments were conducted using the Fake News Challenge Stage 1 (FNC-1) dataset, available at: https://github.com/FakeNewsChallenge/fnc-1. No new datasets were generated during the current study. All preprocessing, model training, and evaluation procedures were performed on this publicly available dataset, ensuring transparency and reproducibility of the reported results.

Acknowledgments: The authors would like to acknowledge the use of artificial intelligence–based tools to support this study. Specifically, AI-assisted tools were employed to refine the clarity, structure, and language of the manuscript during the writing and editing process. In addition, AI tools were used to assist in the generation of illustrative figures for conceptual visualization purposes. All technical content, experimental design, analysis, and interpretation of results remain the responsibility of the authors. No external administrative, financial, or material support was received beyond what has been stated in the Funding section.

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

References

[1] N. Newman, "Overview and key findings of the 2025 Digital News Report," *Reuters Institute*, 2025. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025/dnr-executive-summary (accessed Jan. 05, 2026).

[2] J. T. Feezell, J. K. Wagner, and M. Conroy, "Exploring the effects of algorithm-driven news sources on political behavior and polarization," *Comput. Human Behav.*, vol. 116, p. 106626, Mar. 2021, doi: 10.1016/j.chb.2020.106626.

[3] S. Valenzuela, M. Piña, and J. Ramírez, "Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing," *J. Commun.*, vol. 67, no. 5, pp. 803–826, Oct. 2017, doi: 10.1111/jcom.12325.

[4] X. Feng, J. Luo, Y. Yang, D. El Baz, and L. Shi, "Health Misinformation Detection: Approaches, Challenges and Opportunities," *Inq. J. Heal. Care Organ. Provision, Financ.*, vol. 62, Sep. 2025, doi: 10.1177/00469580251384784.

[5] L. Pangtey, A. Bhatnagar, S. Bansal, S. S. Dar, and N. Kumar, "Large Language Models Meet Stance Detection: A Survey of Tasks, Methods, Applications, Challenges and Future Directions," *arXiv*. Jan. 19, 2026. [Online]. Available: http://arxiv.org/abs/2505.08464

[6] A. Altheneyan and A. Alhadlaq, "Big Data ML-Based Fake News Detection Using Distributed Learning," *IEEE Access*, vol. 11, no. March, pp. 29447–29463, 2023, doi: 10.1109/ACCESS.2023.3260763.

[7] A. Angdresey, L. Sitanayah, and I. L. H. Tangka, "Sentiment Analysis for Political Debates on YouTube Comments using BERT Labeling, Random Oversampling, and Multinomial Naïve Bayes," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 342–354, Jan. 2025, doi: 10.62411/jcta.11668.

[8] A. Bahmani, "Fusion of Statistical and Stylistic Text Features with SVM for Persian Sentiment Analysis," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 534–548, Dec. 2025, doi: 10.62411/faith.3048-3719-287.

[9] D. Küçük and F. Can, "Stance Detection," *ACM Comput. Surv.*, vol. 53, no. 1, pp. 1–37, Jan. 2021, doi: 10.1145/3369026.

[10] D. R. I. M. Setiadi, D. Marutho, and N. A. Setiyanto, "Comprehensive Exploration of Machine and Deep Learning Classification Methods for Aspect-Based Sentiment Analysis with Latent Dirichlet Allocation Topic Modeling," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 12–22, May 2024, doi: 10.62411/faith.2024-3.

[11] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, p. 799, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14744.

[12] D. R. I. M. Setiadi, W. Warto, A. R. Muslikh, K. Nugroho, and A. N. Safriandono, "Aspect-Based Sentiment Analysis on E-commerce Reviews using BiGRU and Bi-Directional Attention Flow," *J. Comput. Theor. Appl.*, vol. 2, no. 4, pp. 470–480, Apr. 2025, doi: 10.62411/jcta.12376.

[13] N. Alturayeif, H. Luqman, and M. Ahmed, "A systematic review of machine learning techniques for stance detection and its applications," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 5113–5144, Mar. 2023, doi: 10.1007/s00521-023-08285-7.

[14] G. Rajendran, B. Chitturi, and P. Poornachandran, "Stance-In-Depth Deep Neural Approach to Stance Classification," *Procedia Comput. Sci.*, vol. 132, pp. 1646–1653, 2018, doi: 10.1016/j.procs.2018.05.132.

[15] A. Abedalla, A. Al-Sadi, and M. Abdullah, "A Closer Look at Fake News Detection: A Deep Learning Perspective," in *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, Oct. 2019, no. October 2019, pp. 24–28. doi: 10.1145/3369114.3369149.

[16] J. Zhu, "Comparative study of sequence-to-sequence models: From RNNs to transformers," *Appl. Comput. Eng.*, vol. 42, no. 1, pp. 67–75, Feb. 2024, doi: 10.54254/2755-2721/42/20230687.

[17] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *J. Big Data*, vol. 11, no. 1, p. 25, Feb. 2024, doi: 10.1186/s40537-023-00842-0.

[18] P. H. Hussan and S. M. Mangj, "BERTPHIURL : A Teacher-Student Learning Approach Using DistilRoBERTa and RoBERTa for Detecting Phishing Cyber URLs," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, 2025, doi: 10.62411/faith.3048-3719-71.

[19] H. Karande, R. Walambe, V. Benjamin, K. Kotecha, and T. Raghu, "Stance detection with BERT embeddings for credibility analysis of information on social media," *PeerJ Comput. Sci.*, vol. 7, p. e467, Apr. 2021, doi: 10.7717/peerj-cs.467.

[20] K. Jain, F. Doshi, and L. Kurup, "Stance Detection Using Transformer Architectures and Temporal Convolutional Networks," in *Advances in Intelligent Systems and Computing*, 2021, pp. 437–447. doi: 10.1007/978-981-15-4409-5_40.

[21] M. Matero, N. Soni, N. Balasubramanian, and H. A. Schwartz, "MeLT: Message-Level Transformer with Masked Document Representations as Pre-Training for Stance Detection," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2959–2966. doi: 10.18653/v1/2021.findings-emnlp.253.

[22] W. Lu, D. Ming, X. Mao, J. Wang, Z. Zhao, and Y. Cheng, "A DeBERTa-Based Semantic Conversion Model for Spatiotemporal Questions in Natural Language," *Appl. Sci.*, vol. 15, no. 3, p. 1073, Jan. 2025, doi: 10.3390/app15031073.

[23] F. Leng, F. Li, Y. Bao, T. Zhang, and G. Yu, "DABC: A Named Entity Recognition Method Incorporating Attention Mechanisms," *Mathematics*, vol. 12, no. 13, p. 1992, Jun. 2024, doi: 10.3390/math12131992.

[24] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," *arXiv*. Oct. 06, 2021. [Online]. Available: http://arxiv.org/abs/2006.03654

[25] C. Lu and Y. Shen, "Disentangled Abstract Meaning Representation Attention Augmented DeBERTa Model," in *Proceedings of the 2024 2nd International Conference on Information Education and Artificial Intelligence*, Dec. 2024, pp. 821–826. doi: 10.1145/3724504.3724638.

[26] L. Liu, X. Liu, J. Gao, W. Chen, and J. Han, "Understanding the Difficulty of Training Transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5747–5763. doi: 10.18653/v1/2020.emnlp-main.463.

[27] Y. Tay *et al.*, "Scale Efficiently: Insights from Pre-training and Fine-tuning Transformers," *arXiv*. Jan. 30, 2022. [Online]. Available: http://arxiv.org/abs/2109.10686

[28] M. H. Hoti, F. Qorrolli, and F. Spahija, "Enhancing Fake News Detection via Stance Analysis: Leveraging Advanced NLP Techniques and Machine Learning Models," *Int. J. Interact. Mob. Technol.*, vol. 19, no. 11, pp. 39–50, Jun. 2025, doi: 10.3991/ijim.v19i11.55007.

[29] R.-H. Hsu, T.-W. Hsu, and Y.-Y. Chen, "Addressing Imbalanced Data in Stance Detection for Improved Fake News Detection," in *2025 IEEE 8th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, Aug. 2025, pp. 144–150. doi: 10.1109/MIPR67560.2025.00031.

[30] M. M. Lawal and A. Abdulrauf, "Fake News Detection Using Bi-LSTM Architecture: A Deep Learning Approach on the ISOT Dataset," *J. Comput. Theor. Appl.*, vol. 3, no. 2, pp. 104–114, Sep. 2025, doi: 10.62411/jcta.14235.

[31] R. Muthusami, K. Saritha, K. S. Rao, P. Sugapriya, and G. Saveetha, "Interpretable stance detection in social media via topic-guided transformers," *Discov. Artif. Intell.*, vol. 5, no. 1, p. 355, Nov. 2025, doi: 10.1007/s44163-025-00635-9.

[32] B. Zhang, G. Dai, J. Ma, H. Lin, and H. Huang, "Large Language Model Enhanced Fuzzy Logic Fusion Framework for Stance Detection," in *Communications in Computer and Information Science*, 2025, pp. 130–144. doi: 10.1007/978-981-96-5084-2_9.

[33] J. Ma, C. Wang, H. Xing, D. Zhao, and Y. Zhang, "Chain of Stance: Stance Detection with Large Language Models," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2025, pp. 82–94. doi: 10.1007/978-981-97-9443-0_7.

[34] L. M. Pham and H. Cao the, "LNLF-BERT: Transformer for Long Document Classification With Multiple Attention Levels," *IEEE Access*, vol. 12, pp. 165348–165358, 2024, doi: 10.1109/ACCESS.2024.3492102.

[35] S. Kumar and A. Solanki, "An abstractive text summarization technique using transformer model with self-attention mechanism," *Neural Comput. Appl.*, vol. 35, no. 25, pp. 18603–18622, Sep. 2023, doi: 10.1007/s00521-023-08687-7.

[36] B. R. K, B. E. Babu, S. P. Racharla, P. Pavani, Y. Balagoni, and M. Ajmeera, "A Performance Evaluation of Transformer Models and Recurrent Neural Networks Models in Efficient Text Classification Tasks," in *2025 8th International Conference on Computing Methodologies and Communication (ICCMC)*, Jul. 2025, pp. 1171–1177. doi: 10.1109/ICCMC65190.2025.11140627.

[37] A. Hanselowski *et al.*, "A retrospective analysis of the fake news challenge stance detection task," *COLING 2018 - 27th Int. Conf. Comput. Linguist. Proc.*, pp. 1859–1874, 2018.

[38] C. Conforti, M. T. Pilehvar, and N. Collier, "Towards Automatic Fake News Detection: Cross-Level Stance Detection in News Articles," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018, pp. 40–49. doi: 10.18653/v1/W18-5507.

[39] V. Slovikovskaya, "Transfer Learning from Transformers to Fake News Challenge Stance Detection (FNC-1) Task," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, Oct. 2019, no. May, pp. 1211–1218. [Online]. Available: http://arxiv.org/abs/1910.14353

[40] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications," *Healthc. Inform. Res.*, vol. 27, no. 3, pp. 189–199, Jul. 2021, doi: 10.4258/hir.2021.27.3.189.

[41] J. J. Eertink, M. W. Heymans, G. J. C. Zwezerijnen, J. M. Zijlstra, H. C. W. de Vet, and R. Boellaard, "External validation: a simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical prediction models using PET data from DLBCL patients," *EJNMMI Res.*, vol. 12, no. 1, p. 58, Sep. 2022, doi: 10.1186/s13550-022-00931-w.

[42] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," *J. Anal. Test.*, vol. 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2.