TDMulti: A Tunisian Dialect-Modern Standard Arabic Multitask Corpus with a Context-Aware Cross-Attention BERT Model

Roua Torjmen, Kais Haddar

Faculty of sciences of Sfax, University of Sfax
Road Sokra Km 3, 3018 Sfax, Tunisia
rouatorjmen@gmail.com, kais.haddar@yahoo.fr

Abstract

The Tunisian dialect dominates online communication in Tunisia but remains severely under-resourced in natural language processing. We introduce the first multitask corpus of Tunisian dialect manually aligned with its equivalents in modern standard Arabic. The TDMulti corpus consists of 3,100 social media comments annotated with 12,400 labels for four interrelated tasks: hate speech detection, sentiment polarity classification, sarcasm identification, and topic category classification. The TDMulti corpus provides a new benchmark for studying pragmatic and social aspects of Tunisian dialect in relation to modern standard Arabic. To leverage this resource, we propose a deep learning model based on transformer architectures. We design three variants: a baseline multitask classifier, a cross-attention model aligning Tunisian dialect and modern standard Arabic representations, and a context-aware cross-attention mechanism with task-specific masking. We evaluate the approach using large pre-trained Arabic language models under different configurations. Results show that the context-aware cross-attention model achieves the best performance, particularly for sarcasm and hate speech detection. TDMulti is released under an open license, contributing a novel resource to advance research on Arabic dialect processing.

Keywords: Tunisian Dialect, Modern Standard Arabic, Multitask Learning, Hate Speech, Sarcasm Detection

1. Introduction

The Arabic language has a wide variety of dialects, which are widely used in everyday communication, especially on social media. Among them, the Tunisian dialect (TD), after the revolution, has gained an important position in online discourse and has become the primary means of expressing opinions, criticisms, and emotions. However, despite its prevalence, TD remains severely low-resourced in natural language processing (NLP). Most work in Arabic NLP has focused on Modern Standard Arabic (MSA).

This gap is particularly critical for tasks involving pragmatic and social phenomena such as hate speech, sentiment, and sarcasm. These tasks are challenging not only due to the scarcity of annotated resources but also because dialectal text is characterized by spelling variation, code-switching with French, and informal style. Existing datasets for Arabic dialects typically focus on a single task and rarely provide parallel alignment with MSA. To the best of our knowledge, no multitask resource exists for TD.

In this paper, we address these challenges by introducing a novel corpus of 3,100 Tunisian social media comments. Each instance is aligned with its MSA equivalent and annotated for four interrelated tasks: hate speech detection, sentiment polarity, sarcasm identification, and topic category classification. This resource enables the joint study of pragmatic and social aspects of TD in relation to MSA.

To leverage this corpus, we design a multitask model based on BERT. We investigate three model variants: (i) a baseline multitask classifier, (ii) a cross-attention model aligning TD and MSA representations, and (iii) a context-aware cross-attention model that incorporates task-specific masking for improved robustness. We evaluate our models using MARBERT and AraBERT under different settings. Experimental results show that MARBERT with context-aware cross-attention achieves state-of-the-art performance across tasks, particularly for sarcasm detection and hate speech classification.

Our contributions are threefold:

1. We present the first publicly available Tunisian dialect corpus aligned with MSA and annotated for multiple pragmatic and social tasks.
2. We propose a novel cross-attention multitask BERT model, including a context-aware variant.
3. We provide a comprehensive evaluation of Arabic BERT models on Tunisian social media.

The remainder of this paper is structured as follows. Section 2 reviews previous research on TD resources and methodological advances for hate speech, sarcasm, and offensive language detection, including classical, deep learning, and transformer-based approaches. Section 3 introduces the proposed Tunisian-MSA multitask corpus and its annotation scheme. Section 4 explains our proposed methodology, including the baseline and different configurations. Section 5 details the experimental setup, reports, analyzes the results, and discusses their implications. Finally, Section 6 concludes the paper and outlines future research directions.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3240–3249
11–16 May 2026.

2. Related Work

In recent years, there has been growing interest in detecting hate speech and offensive language, as well as sarcasm, in resource-poor dialects, including the Tunisian dialect. To date, research has focused on two main areas: the creation of databases on hate speech and offensive language, as well as sarcasm in the Tunisian dialect. The second area focuses on methodological innovations suitable for detecting the toxic dialectal language and sarcasm. Below, we review the main efforts in these two areas.

2.1 Tunisian Dialect Datasets

High-quality corpora are essential for research on hate speech and offensive language. In the context of the Tunisian dialect, only a few datasets are available.

Haddad et al. (2019) presented T-HSAB, the first public corpus for hate speech and offensive language in the Tunisian dialect. Their corpus includes about 6,000 comments. Kharrat et al. (2024) proposed the HateTune corpus in the Tunisian dialect dedicated to hate speech detection. This corpus contains over 12,000 comments labeled as Hateful or Neutral, focusing on texts in Arabic script. Trabelsi and Kouki (2024) published Bully.tn, a dataset aimed at detecting cyberbullying in the Tunisian dialect. This corpus comprises over 11,000 labeled comments from Facebook and YouTube. Gharbi et al. (2021) published TEET!, a corpus of approximately 10,000 toxic comments in Tunisian dialect. Abbes et al. (2023) collected a smaller dataset containing 2,000 Facebook comments in Tunisian dialect. Each comment is annotated for hate-related categories such as racism, sexism, and religion. Baazaoui et al. (2025) introduced a comprehensive hate speech dataset comprising 13,000 comments annotated with detailed labels such as offensive language, racism, sexism, violent language, religious intolerance, etc.

Fourati et al. (2024) introduced PoliTun, a dataset specifically designed for political analysis in TD. The dataset contains approximately 30,000 political tweets written in Arabic script. Each tweet is labeled in two key aspects: category and opinion. The category aspect consists of six distinct labels, while the opinion aspect is classified into three labels.

Mekki et al. (2022) created a sarcasm detection dataset of approximately 23,000 Tunisian social media comments written in Arabic and Latin script. Each comment is manually annotated as sarcastic or non-sarcastic.

Table 1 compares these datasets in terms of size, labels, annotator details, public availability, and data source.

Dataset	Year	Size	Labels	Annotators	Public	Source
T-HSAB (Haddad et al., 2019)	2019	~6K	Hate, Abusive, Neutral	3 annotators (2 males, 1 female; Master's & PhD students)	Yes	Social media
HateTune (Kharrat et al., 2024)	2024	~12K	Hate, Neutral	6 engineering students (5 female, 1 male, aged 20)	Yes	Social media
Bully.tn (Trabelsi & Kouki, 2024)	2024	~11K	Bullying, Non-bullying, Mixed, Hate	Not specified	Yes	Facebook, YouTube
TEET! (Gharbi et al., 2021)	2021	~10K	Hate, Abusive, Normal	Not specified	No	Social media
PoliTun Fourati et al. (2024)	2024	~30K	Category (6 labels) and Opinion (3 labels)	3 female native speakers	No	Twitter
Abbes et al. (2023)	2023	~2K	Racism, Sexism, Religion, Normal	5 PhD & postdoc researchers	No	Facebook
Mekki et al. (2022)	2022	~23K	Sarcastic, Non-sarcastic	Native Tunisian experts	No	Facebook, Twitter
Baazaoui et al. (2025)	2025	~13K	Offensive, Racism, Sexism, Violent, Religion, Sarcasm, No Hate	3 annotators	No	Social media

Table 1: Comparison of existing Tunisian dialect datasets

Although several Tunisian dialect datasets exist, they present important limitations with respect to the objectives of this work. Most resources focus on a single task (hate speech or sarcasm) and do not provide parallel alignment with MSA. Moreover, many datasets are not publicly accessible or lack multi-dimensional pragmatic annotations. Our goal is not to replace these resources but to complement them by providing (i) TD–MSA alignment, (ii) joint annotation across four interrelated pragmatic tasks, and (iii) an openly available benchmark enabling multitask modeling. This combination of properties is not covered by existing corpora and motivates the creation of TDMulti corpus.

2.2 Methodological Advances in Tunisian Dialect Hate Speech and Abuse Detection

From a methodological perspective, approaches to detecting toxic and sarcastic content in Tunisian dialect have evolved from classical machine learning to deep learning and transformer-based models. These works address challenges such as data sparsity, class imbalance, and dialectal variation.

2.2.1 Classical Machine Learning

On HateTune, Kharrat et al. (2024) compared Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Random Forests (Breiman, 2001), and XGBoost (Chen and Guestrin, 2016) using TF-IDF and manually designed features. SVM achieved the best performance (precision ≈ 0.81 and recall ≈ 0.88 for the hate class). However, these methods rely heavily on feature engineering and often generalize poorly to unseen dialect variants.

Gharbi et al. (2021) on TEET!, compared Naive Bayes, SVM, random forest, and logistic regression using unigram and bigram features. Logistic regression produced F1 scores up to ≈ 0.93 when combining feature types. These classic models highlight that simple feature sets can be strong benchmarks, but they lack deep semantic understanding and adaptability to new expressions.

2.2.2 Deep Learning Approaches

Trabelsi and Kouki (2024) explored recurrent architectures (RNN, LSTM, BiLSTM, GRU) (Graves et al., 2005; Sherstinsky, 2018) coupled with embeddings such as Word2Vec, TF- IDF, and dimensionality reduction methods. RNNs demonstrated high recall, while GRU offered better accuracy in some contexts. However, these models remain sensitive to hyperparameters and require sufficient training data to avoid overfitting.

Abbes et al. (2023) adopted a BiLSTM with an attention mechanism on BERT embeddings on their dataset, reporting an accuracy of 98.89%. The introduction of attention improved interpretability and focus on toxic signals. However, the relatively small size of the non-public dataset limits reproducibility and external validation.

2.2.3 Transformer-Based Models

More recently, work has shifted toward transformer-based models (Tenney et al., 2019; Antoun et al., 2020; Abdul-Mageed et al., 2021; Abdaoui et al., 2021), which better capture contextual and dialectal variations.

For example, Salomon et al. (2022) fine-tuned AraBERT (Antoun et al., 2020) on a mixture of T-HSAB and related datasets, reporting F1 scores as high as 0.99 for hate speech detection. However, AraBERT is primarily pre-trained on MSA, which may not fully model dialect-specific lexical nuances.

Gasmi et al. (2024) used DziriBERT (Abdaoui et al., 2021), a dialect-pre-trained transformer, for hate speech detection and addressed class imbalance via SMOTE. Their experiments yielded an accuracy of 82%, demonstrating that oversampling techniques can further improve performance in skewed data settings.

MARBERT (Abdul-Mageed et al., 2021) has demonstrated competitive performance on hate speech detection across Arabic dialects. In the study by Fourati et al. (2024), the MARBERT model achieved an F1 macro score of 63% for the Opinion Classification subtask.

Baazaoui et al. (2025) demonstrated the consistent superiority of AraBERT, achieving 89.02% recall, 90.75% precision, 90.36% F1 score, and 92.84% accuracy on the datasets, outperforming Bi-LSTM.

These studies highlight the effectiveness of transformer-based models in tasks that demand a deep understanding of linguistic and contextual nuances. Given their robustness, AraBERT and MARBERT have emerged as leading models for hate speech detection in the Tunisian context. In this paper, we adopt both MARBERT and AraBERT to evaluate their performance on our TD-MSA multitask dataset.

To our knowledge, no previous study has conducted a systematic direct comparison of multiple BERT-based models on the same Tunisian benchmark dataset. Furthermore, no study has explored architectural variations such as cross-attention or context-awareness or multitask fusion between tasks.

3. TDMulti corpus

This section describes the creation of the TDMulti corpus, a multitask TD-MSA parallel dataset designed for pragmatic and social analysis on social media. The corpus includes data collection, annotation workflow, and detailed guidelines across four interrelated

tasks: hate speech detection, sentiment polarity, sarcasm identification, and topic classification.

3.1 Data Collection and Alignment

The corpus consists of 3,100 user-generated comments collected from public Tunisian Facebook pages, YouTube channels, and discussion forums. Comments were gathered between 2023 and 2025 using topic-based sampling to ensure diverse representation of Tunisian online discourse. All comments were written in Arabic script. Non-Tunisian or duplicated content was manually filtered out. Each text was manually anonymized by removing user names, mentions, URLs, and personal identifiers. No private metadata were stored.

The MSA equivalents were not automatically generated. Instead, they were manually written by three native Tunisian annotators trained in linguistic annotation as controlled reformulations preserving semantic and pragmatic intent. Annotators followed guidelines ensuring lexical normalization without altering sarcasm, sentiment, or discourse meaning. This process should therefore be understood as human translation/alignment rather than post-hoc matching of pre-existing texts. Diacritics were automatically removed from both variants using a Python preprocessing script.

A stratified sampling approach was adopted to preserve topic and demographic balance, covering a broad range of online domains such as politics, society, economy, and entertainment.

3.2 Annotation Process

To ensure consistency in annotations, we followed a three-phase protocol inspired by best practices from recent datasets (Shiwakoti et al., 2024). In the first phase, three annotators annotated 100 randomly selected comments to test the clarity of the instructions and the feasibility of the task. Ambiguities were recorded and discussed during feedback sessions. Following feedback from the pilot phase, the second phase refined the definitions and examples of each annotation, particularly for borderline cases such as sarcasm and irony, as well as hatred and insult. An additional 200 comments were annotated to validate the revisions. The final phase resolved any remaining disagreements through group discussions moderated by an experienced linguist. The final annotation was assigned only after full consensus. This iterative procedure ensured a common interpretation of all categories and minimized variability between annotators.

Annotation was performed by three native speakers of Tunisian Arabic with backgrounds in linguistics or NLP. Each annotator worked independently before the consensus meetings. The dataset was then divided and stratified by label into 80% training sets, 10% validation sets, and 10% test sets to maintain a balanced distribution across all tasks.

3.3 Annotation Guidelines

Each TD–MSA pair was annotated for four distinct yet interrelated dimensions, following standardized definitions. Hate speech detection follows a three-level scheme consisting of neutral (no offensive content), hate (explicit discrimination or hostility), and abuse (offensive or aggressive tone without clear group targeting) categories, inspired by the annotation of Haddad et al. (2019) and Kharrat et al. (2024). Sentiment polarity is defined as Positive or Negative, capturing the comment writer's sentiment. Sarcasm detection distinguishes sarcastic from non-sarcastic comments by capturing the contrast between literal and intended meanings.

The topic classification task assigns each instance to one of twelve thematic domains reflecting Tunisian online discourse: Society, Entertainment, Economy, Politics, Law, Health, Religion, Education, Tourism, Migration, Gastronomy, and Sports.

Annotators received detailed task definitions, representative examples, and explanations to ensure consistency. Table 2 illustrates a sample instance from the corpus.

TD input	نسبة النمو والعلو الشاهق يتعاركوا اشكون ينمو اكثر ملا كذاب *nisbat al- an-numū w al-ʿulū ash-shāhiq yitʿārku, ashkūn yinmu akthar? malā kādhib* 'The growth rate and the Eiffel Tower are competing to see which one increases the most? Liar!'
MSA input	نسبة النمو والعلو الشاهق يتنافسان أيهما ينمو أكثر؟ يا لك من كاذب
Hate speech label	Abuse
Sentiment label	Negative
Sarcasm label	Sarcasm
Topic label	Economy

Table 2: Example of a comment with annotations

3.4 Inter-Annotator Agreement

Inter-annotator agreement was measured on a subset of 310 instances. Cohen's kappa (κ) scores are as follows: 0.80 for hate speech, 0.83 for sentiment, 0.78 for sarcasm, and 0.75 for topic categories.

These scores confirm strong reliability. The consensus-based resolution phase further ensured annotation uniformity and reproducibility.

3.5 Data Statistics

The TDMulti corpus contains 3,100 TD-MSA pairs and 12,400 labels across four tasks. Hate speech detection includes 1,482 neutral comments, 896 hateful comments, and 722 offensive comments. Sentiment polarity covers 1,890 negative samples and 1,210 positive samples. The sarcasm dimension includes 2,366 non-sarcastic comments and 734 sarcastic comments.

Thematic labels are distributed across twelve different domains, with 300 samples for society and entertainment, and 250 for each of the remaining categories: economics, politics, law, health, religion, education, tourism, migration, gastronomy, and sports. Figure 1 illustrates the distribution of comments for each dimension of our corpus.

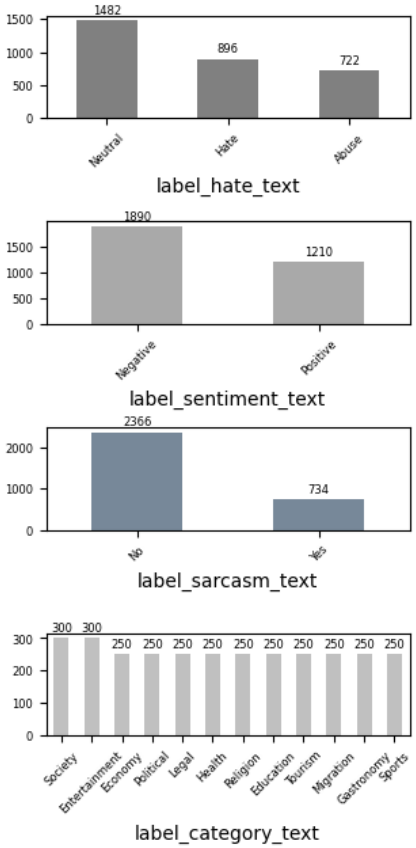

Figure1: Distribution of text categories: Hate, Sentiment, Sarcasm, and Topic

4. Methodology

We propose a multitask learning model designed to predict four tasks simultaneously: hate speech, sentiment, sarcasm, and topic category. The approach leverages two complementary Arabic BERT encoders, AraBERT and MARBERT as the main backbones to model MSA and TD representations, respectively.

AraBERT, pre-trained primarily on MSA and formal text, provides syntactic and grammatical robustness. MARBERT, trained on large-scale social media data covering multiple Arabic dialects, captures informal and dialectal nuances.

To enhance contextual understanding, we introduce two architectural extensions: Cross-attention, which facilitates semantic alignment between TD and MSA, and context-aware fusion (Baruah et al., 2020; Chauhan et al., 2019), which models discourse-level dependencies, improving the recognition of pragmatic cues such as sarcasm and emotional polarity. Figure 2 illustrates the architecture of context-aware cross-attention.

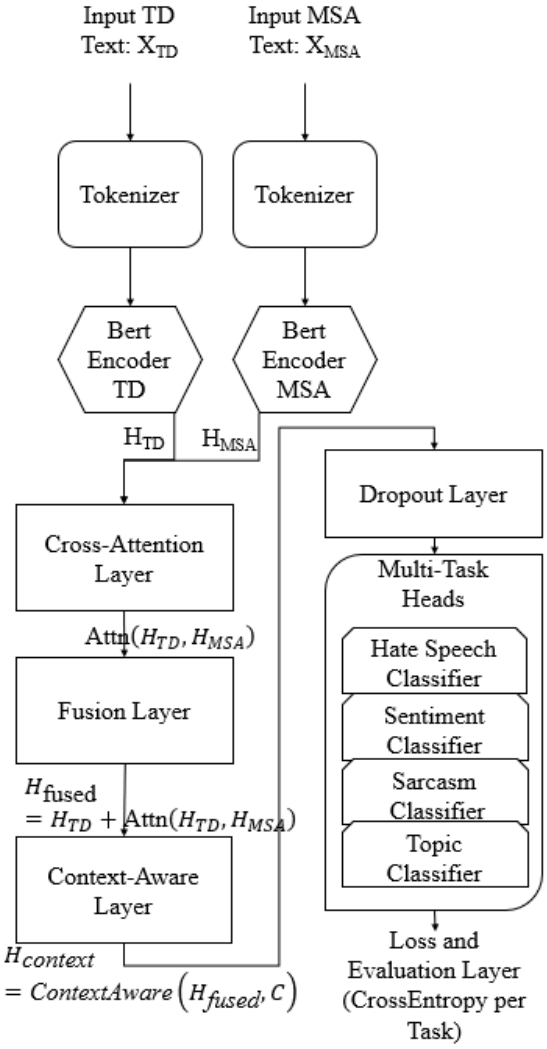

Figure 2: Context-aware cross-attention architecture

4.1 Model Architecture

Let X_{TD} and X_{MSA} denote the TD and MSA sequences, respectively. Both sequences are

independently encoded using a shared BERT backbone to produce contextual representations H_{TD} and H_{MSA}:

$$H_{TD} = \text{BERT}(X_{TD}), \quad H_{MSA} = \text{BERT}(X_{MSA})$$

The cross-attention mechanism enables semantic alignment between dialectal and standard Arabic representations. It is defined as:

$$\text{Attn}(H_{TD}, H_{MSA}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q comes from TD embeddings H_{TD} and K, V come from MSA embeddings H_{MSA}. The resulting attention output is fused with the original dialectal embeddings through residual addition:

$$H_{fused} = H_{TD} + \text{Attn}(H_{TD}, H_{MSA})$$

This fusion enables the model to leverage MSA semantics to disambiguate dialectal words, idiomatic expressions, and context-dependent meanings.

In the context-aware configuration, a context-aware layer is introduced to capture inter-utterance or inter-task dependencies. Given a context representation C, the context-aware representation is computed as:

$$H_{context} = \text{ContextAware}(H_{TD}, C)$$

The full model combines both mechanisms. The cross-attention output H_{fused} is further refined by the context-aware layer:

$$H_{context} = \text{ContextAware}(H_{fused}, C)$$

Each of the four tasks (hate speech, sentiment, sarcasm, and topic classification) has its own task-specific classification head. Each head consists of a self-attention block, followed by two feedforward layers with ReLU activation and a softmax output layer.

The overall training objective is defined as the sum of the cross-entropy losses from all tasks:

$$L_{total} = \sum_t L_t$$

where t corresponds to each task in the multitask setting.

4.2 Training and Optimization

Training employed the AdamW optimizer, a learning rate of 2×10^{-5}, a batch size of 8, seven epochs, and early stopping based on validation F1 to mitigate overfitting. Dropout layers with a rate of 0.1 were applied to task-specific and context-aware layers to further regularize the models. For training and validation, we used loss and accuracy curves. Precision, recall, and F1 score were used as metrics for each task. To include AraBERT and MARBERT, we used the Hugging Face Transformers library. For implementation, we also used the PyTorch framework. All experiments were conducted on a GPU T4 with 12 GB memory.

5. Experimental Results

In this section, we present the experimental evaluation of the proposed methodology on the TDMulti corpus. The dataset was split into 80% training, 10% validation, and 10% test sets. Given the corpus size (3,100 instances), this corresponds to 2,480 training comments, 310 validation comments, and 310 test comments. We evaluated four configurations: a baseline model, a cross-attention-only model, a context-aware-only model, and a combined cross-attention and context-aware model, using AraBERT and MARBERT as backbone encoders.

5.1 Overall Performance

Table 3 presents the learning and validation metrics for different architectural configurations. The Baseline is a standard BERT model without cross-attention or context-aware layers. It processes TD text directly, without explicitly leveraging MSA. Cross-attention only adds alignment between Tunisian Dialect and MSA. Context-aware only captures discourse-level context without cross-attention. The cross-attention + context aware model configuration combines both mechanisms for improved multitask performance.

Setting	Model	Training loss	Training accuracy	Validation loss	Validation accuracy
Cross-attention only	AraBERT	0.16	0.98	3.60	0.78
	MARBERT	**0.14**	**0.99**	**3.36**	**0.82**
Cross-attention + context-aware	AraBERT	0.26	0.97	4.38	0.77
	MARBERT	0.26	0.98	3.77	0.79
Context-aware only	AraBERT	0.24	0.98	4.36	0.76
	MARBERT	0.27	0.97	3.69	0.78
Baseline	AraBERT	0.18	0.98	3.99	0.79
	MARBERT	0.23	0.98	3.81	0.81

Table 3: Training and validation results for AraBERT and MARBERT configuration

Models incorporating cross-attention achieved higher validation accuracy than all other

configurations. The contextual component, while adding interpretive depth, slightly increased overfitting due to its greater representational complexity. Furthermore, all MARBERT configurations demonstrated higher validation accuracy compared to AraBERT.

5.2 Task-Level Performance

Table 4 presents the macro F1 scores for each individual task: hate speech detection, sentiment classification, sarcasm detection, and social topic classification for each configuration.

Setting	Model	Hate (F1)	Sentiment (F1)	Sarcasm (F1)	Topic (F1)
Cross-attention only	AraBERT	0.68	0.84	0.75	0.74
	MARBERT	**0.71**	**0.88**	0.75	**0.79**
Cross-attention + context-aware	AraBERT	0.67	0.87	0.75	0.74
	MARBERT	0.70	**0.88**	**0.78**	0.78
Context-aware only	AraBERT	0.63	0.87	0.76	0.67
	MARBERT	0.70	0.84	**0.78**	0.73
Baseline	AraBERT	0.56	0.78	0.63	0.51
	MARBERT	0.59	0.84	0.71	0.59

Table 4: Task-level macro F1-scores for different configurations

Table 5 presents the average recall, precision, and F1 scores for each setting.

Setting	Model	Recall	Precision	Avg. F1
Cross-attention only	AraBERT	0.74	0.76	0.75
	MARBERT	0.77	**0.80**	**0.78**
Cross-attention + context-aware	AraBERT	0.75	0.77	0.75
	MARBERT	**0.78**	**0.80**	**0.78**
Context-aware only	AraBERT	0.72	0.75	0.73
	MARBERT	0.76	0.77	0.76
Baseline	AraBERT	0.62	0.62	0.62
	MARBERT	0.67	0.69	0.68

Table 5: Average recall, precision, and F1-scores for different configurations

As shown in Table 4 and 5, MARBERT again outperformed AraBERT in all tasks, especially with the cross-attention module. The multitask configuration with cross-attention resulted in more balanced performance across most tasks. For sarcasm detection, the cross-attention and Context-aware configuration performed best.

5.3 Discussion

The proposed multitask model combining cross-attention and context-aware encoding demonstrated consistent performance gains across all four subtasks. The cross-attention effectively aligned dialectal expressions with MSA, supporting robust knowledge transfer without explicit normalization. These results confirm the relevance of alignment techniques for low-resource dialectal processing.

Contextual layers were particularly beneficial for discourse-dependent tasks, such as sarcasm detection, where pragmatic inference requires modeling rhetorical and dialogic structure. However, their integration occasionally reduced performance on tasks like sentiment or hate speech detection, where surface-level cues were sufficient. This suggests that uniform contextual depth across tasks may not be optimal in multitask configurations.

5.4 Error Analysis

A quantitative comparison between the context-aware cross-attention architecture (left matrices) and the cross-attention–only configuration (right matrices) confirms the qualitative observations as shown in Figure 3.

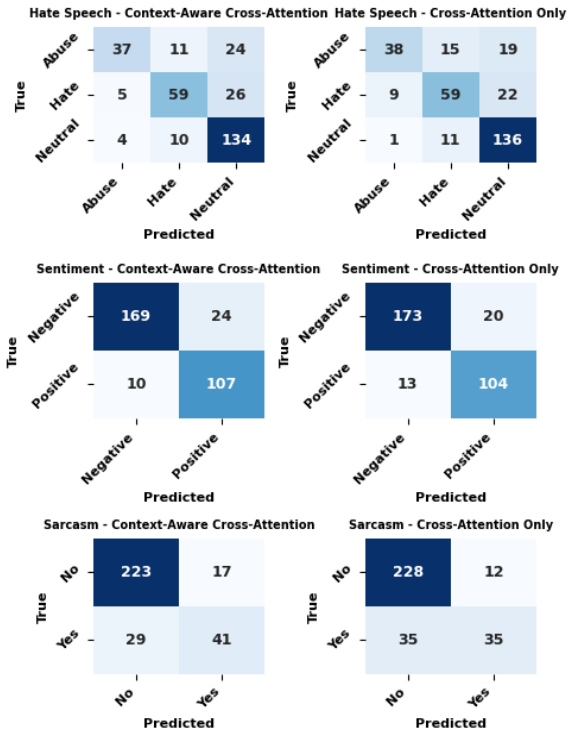

Figure 3: Comparison of confusion matrices for context-aware cross-attention (left) and cross-attention–only (right)

In the hate speech detection subtask, both systems exhibit the highest confusion between the Hate and Abuse categories; however, the inclusion of contextual alignment leads to a measurable reduction in cross-category errors and improved stability of the Neutral class. For sentiment classification, the context-aware model achieves a more balanced discrimination between Positive and Negative instances, with a lower incidence of polarity inversion errors. In sarcasm detection, contextual modeling increases sensitivity to pragmatic cues, correctly identifying a greater number of sarcastic utterances (41 versus 35) while reducing false negatives.

Topic classification showed the highest confusion, particularly between politics and law, and between entertainment and society, illustrating semantic and pragmatic overlap. In contrast, topics such as health, religion, and tourism achieved over 80% precision. These results suggest that remaining errors stem more from pragmatic ambiguity and inter-domain proximity than from lexical misunderstanding.

A qualitative analysis of misclassifications revealed recurring patterns across tasks. In hate speech detection, indirect or culturally loaded expressions are often misclassified. For instance, **حتى هو حب يعمل فيها بورقيبة صغير** *ḥattā huwa ḥabb yaʿmil fīhā Burqība ṣaghīr* 'He also wants to act like little Bourguiba' lexically appears neutral but conveys political mockery. While a base model misses such implicit abuse, the enhanced model captures the sarcastic tone through dialect-MSA alignment and contextual inference.

Sentiment errors often arise in emotionally mixed comments. In **نحب التصاور متاعك أما نكرهك** *niḥibb it-taṣāwīr mtāʿik ammā nikrahik*, 'I love your photos but I hate you', the base model overweights the positive clause, whereas the context-aware model correctly identifies the negative sentiment by modeling discourse-level polarity shifts.

Sarcasm remains challenging in the absence of explicit cues. A comment like **برافو، كهو استحق نوبل** *brāvo, kahu istaḥaq Nūbil!*, 'Bravo, he definitely deserved a Nobel!' may be interpreted literally. The context-aware model correctly captures the rhetorical exaggeration, relying on learned patterns and prior context.

Topic misclassification often occurs between overlapping domains. The comment **قانونهم يخدم كان في الانتخابات والباقي لا** *qānūnhum yikhdim kān fī l-intikhābāt, w il-bāqī lā*, 'Their laws only work during elections; the rest is just emptiness' is sometimes labeled as Law when it actually critiques political manipulation. The improved model, however, assigns it correctly to the Politics category by leveraging implied intent and discourse structure.

6. Conclusion

This paper introduced TDMulti, the first publicly available multitask corpus aligning Tunisian Dialect and Modern Standard Arabic across multiple annotation layers, designed for hate speech, sentiment, sarcasm, and topic detection. To leverage this resource, we presented a novel multitask framework combining cross-attention mechanisms with context-aware modeling, enabling joint analysis of pragmatic and social meanings in Tunisian Arabic. Our experiments demonstrate that integrating dialect-MSA alignment through cross-attention with discourse-level contextualization improves average F1 by 0.1 compared to baseline, especially for tasks requiring understanding of implicit meaning and rhetorical nuance.

The TDMulti corpus offers a valuable benchmark for future research in dialect processing, transfer learning, and low-resource Arabic NLP. Future work will focus on expanding the corpus to additional dialects, incorporating user-level and conversational context, and exploring multimodal signals such as emojis and images to better capture pragmatic intent in social media discourse.

7. Limitations

Although the TDMulti corpus is a valuable resource for Tunisian Arabic, it comprises only 3,100 social media comments and focuses exclusively on the Arabic script. Future extensions should incorporate data written in the Latin script as well as instances of code-switching to better capture the multilingual nature of Tunisian online communication. Furthermore, the predefined topic labels may not adequately represent emerging or evolving online themes. Finally, the corpus was constructed using topic-balanced sampling, which may not fully reflect the natural distributional and stylistic diversity of Tunisian online discourse.

8. Ethical Considerations

All comments were collected from public sources and anonymized by removing usernames and identifiers. Annotators were informed of potentially offensive content and had the right to withdraw at any time. The published corpus contains only anonymized text, ensuring compliance with data privacy standards.

9. Reproducibility Statement

The TDMulti corpus and all code for the experiments presented in this paper are publicly available at: https://github.com/rouatorjmen1/MultiTD-Multitask-BERT.

10. Bibliographical References

Abbes, M., Kechaou, Z., & Alimi, A. M. (2023). Deep learning approach for Tunisian hate speech detection on Facebook. In *Proceedings of the IEEE Symposium on Computers and Communications (ISCC 2023)*, pp. 739-744, Tunis, Tunisia, July. IEEE.

Abdaoui, A., Berrimi, M., Oussalah, M., & Moussaoui, A. (2021). DziriBERT: A pre-trained language model for the Algerian dialect. *arXiv preprint arXiv:2109.12346*.

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2021)*, pp. 7088-7105, Online, August. ACL.

Antoun, W., Baly, F., & Hajj, H. M. (2020). AraBERT: Transformer-based model for Arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Baazaoui, H., Cheniti-Belcadhi, L., & Zrigui, M. (2025). Hate speech detection in Tunisian dialect. In *Proceedings of the IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SERA 2025)*, pp. 2-8, Hammamet, Tunisia, May. IEEE.

Baruah, A., Das, K., Barbhuiya, F., & Dey, K. (2020). Context-aware sarcasm detection using BERT. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 83-87, Seattle, USA, July. ACL.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), pp. 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, pp. 785-794, San Francisco, USA, August. ACM.

Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019). Context-aware interactive attention for multimodal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pp. 5647-5657, Hong Kong, China, November. ACL.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), pp. 273-297.

Fourati, C., Hammami, R., Latiri, C., & Haddad, H. (2024). PoliTun: Tunisian political dataset for detecting public opinions and categories orientation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pp. 178-185, Antalya, Turkey, October. ACL.

Gasmi, S., Mezghani, A., & Kherallah, M. (2024). SMOTE for enhancing Tunisian hate speech detection on social media with machine learning. *International Journal of Hybrid Intelligent Systems (IJHIS)*, 20(4), pp. 355-368.

Gharbi, S., Arfaoui, H., Haddad, H., & Kchaou, M. (2021). TEET! Tunisian dataset for toxic speech detection. *arXiv preprint arXiv:2110.05287*.

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2005)*, pp. 799-804, Warsaw, Poland, September. Springer.

Haddad, H., Mulki, H., & Oueslati, A. (2019). T-HSAB: A Tunisian hate speech and abusive dataset. In *Proceedings of the International Conference on Arabic Language Processing (ICALP 2019)*, pp. 251-263, Hammamet, Tunisia, November. Springer.

Kharrat, O., Mohamed, F. A., Mtimet, I., Benamor, N., & Fourati, C. (2024). HateTune: Tunisian dialect hate speech detection dataset. In *Proceedings of the International Conference on Arabic Language Processing (ICALP 2023)*, pp. 63-73, Hammamet, Tunisia, November. Springer.

Mekki, A., Zribi, I., Ellouze, M., & Belguith, L. H. (2022). Sarcasm detection in Tunisian social media comments: Case of COVID-19. In *Proceedings of the International Symposium on Methodologies for Intelligent Systems (ISMIS 2022)*, pp. 45-57, Hammamet, Tunisia, October. Springer.

Salomon, P. O., Kechaou, Z., & Wali, A. (2022). Arabic hate speech detection system based on AraBERT. In *Proceedings of the International Conference on Computing, Intelligence and Communication (ICCICC 2022)*, pp. 208-213, Monastir, Tunisia, June. Springer.

Sherstinsky, A. (2018). Fundamentals of recurrent neural network (RNN) and LSTM network. *arXiv preprint arXiv:1808.03314*.

Shiwakoti, S., Thapa, S., Rauniyar, K., Shah, A., Bhandari, A., & Naseem, U. (2024). Analyzing the Dynamics of Climate Change Discourse on Twitter: A New Annotated Corpus and Multi-Aspect Classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

Resources and Evaluation (LREC-COLING 2024), pp. 984-994, Torino, Italy, May. ELRA and ICCL.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 4593-4601, Florence, Italy, July. ACL.

Trabelsi, F. B. F., & Kouki, S. (2024). Bully.tn: A cyberbullying detection dataset in Tunisian dialect. In *Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2024)*, pp. 1-8, Sousse, Tunisia, May. IEEE.