The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 5, October 2025
https://doi.org/10.33022/ijcs.v14i5.4984

# Architectural Evolution of Transformer Models in NLP: A Comparative Survey of Recent Developments

**Diyar Waysi Naaman[1], Berivan Tahir Ahmed[2], Ibrahim Mahmood Ibrahim[3]**
diyar457@gmail.com[1], berivantahir86@gmail.com[2], ibrahim.mahmood@auas.edu.krd[3]
[1]Ministry of Education, General Directory of Education in Duhok, Kurdistan Region, Iraq
[2,3]Akre University for Applied Science -Department of Computer Networks and Information Security, Iraq

**Article Information**

Received : 25 Aug 2025
Revised : 6 Sep 2025
Accepted : 7 Oct 2025

**Keywords**

XLM-RoBERTa, XLM-R, BERT, Transfer Model, Natural Language Processing

**Abstract**

This comprehensive literature review examines the impact and advancements of Cross-lingual Language Model - Robustly Optimized Bidirectional Encoder Representations from Transformers-based Multilingual (XLM-RoBERTa) on multilingual natural language processing, tracking the evolution of its impact from example incorporating over a hundred research works from the years 2020 through 2025. Because language technologies are on the rise, XLM-RoBERTa has become synonymous with a well-known cross-lingual model as it outdoes its predecessors with the novel pre-training method on 2.5TB of multilingual corpora from 100 languages.

This model demonstrates impressive zero-shot cross-lingual transfer dominance with more than 85% accuracy gains over older models in low-resource languages, as well as competitive results in monolingual benchmarks. This review integrates the results from XLM-RoBERTa's architectural evolution, pre-training and masked language modelling scheme, and its performance in numerous benchmarks, including but not limited to named entity recognition, question answering, text classification, sentiment analysis, and detection of hate speech. With other multilingual models like mBERT, XLM, and mT5, we conduct extensive comparative evaluations to expose the critical architectural advantages behind the continual 5-6% annual performance gain since 2020.

The results highlight the importance of XLM-RoBERTa for advancements in language-agnostic representations and for narrowing the gap in performance between high-resource and low-resource languages, which has far-reaching consequences for global access to language technologies. Nevertheless, the critique reveals serious shortcomings in computational resource requirements and in morphologically rich languages, inviting further research in the multilingual transformer development.

## A. Introduction

Recent years have witnessed remarkable advancements in natural language processing (NLP), largely driven by transformer-based architectures and self-supervised pre-training paradigms. While models such as BERT [1] and RoBERTa [2] have revolutionized monolingual NLP capabilities, their application across diverse languages remains challenged by linguistic diversity and resource disparities. XLM-RoBERTa [3] represents a significant milestone in addressing these challenges through cross-lingual model pre-training at unprecedented scale.

In an increasingly interconnected global landscape, the ability to process and understand text across linguistic boundaries has become paramount. Traditional approaches to multilingual NLP have often relied on language-specific models or translation systems, resulting in fragmented ecosystems that reinforce disparities between high-resource and low-resource languages. XLM-RoBERTa challenges this paradigm by offering a unified model capable of representing 100 languages within a shared embedding space, thereby facilitating knowledge transfer across languages.

This literature review aims to comprehensively analyze XLM-RoBERTa's architecture, methodology, and performance across diverse NLP tasks. Through systematic examination of empirical studies, we seek to address the following research questions:

1. How does XLM-RoBERTa's architecture and pre-training methodology differ from preceding multilingual models?
2. What performance improvements does XLM-RoBERTa demonstrate across diverse NLP tasks and language families?
3. To what extent does XLM-RoBERTa bridge the performance gap between high-resource and low-resource languages?
4. What limitations and challenges persist in XLM-RoBERTa's application to multilingual NLP?

The significance of this review lies in its synthesis of fragmented research findings into a coherent narrative that elucidates XLM-RoBERTa's contributions to multilingual NLP. By critically analyzing its strengths and limitations, we aim to provide researchers and practitioners with comprehensive insights into this influential model's capabilities and potential applications.

The remainder of this paper is organized as follows: Section B provides essential background on transformer architectures and multilingual language models; Section C details our methodology for literature selection and analysis; Section D present XLM-RoBERTa and the idea of pretreating model; Section E provide the performance metrics and its equations; Section F presents our comprehensive literature review with the table of comparison between different presented paper work; Section G present the discussion part with the proper analysis with related figures; Section H present the Limitation and Challenges of discussed approach; Section I provide recommendation for future work and finally Section J present the conclusion part.

## B. Background

### 1. Transformer Architecture and Self-Attention Mechanisms

The transformer architecture, introduced by Vaswani et al. [4], represents a paradigm shift in sequence modeling by eliminating recurrence and convolutions in favor of attention mechanisms. At its core, the self-attention mechanism enables the model to weigh the importance of different words in a sequence when representing each word, thereby capturing long-range dependencies more effectively than recurrent architectures. This mechanism can be formalized as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Where Q, K, and V represent query, key, and value matrices derived from input embedding, and d_k is the dimension of the key vectors. Through multi-head attention, transformers process information across multiple representation subspaces, enhancing the model's capacity to capture diverse linguistic patterns.

## 2. Evolution of Pre-trained Language Models

Pre-trained language models have evolved rapidly since ELMo [5] demonstrated the efficacy of contextual word representations. BERT [1] introduced bidirectional context modeling through masked language modeling (MLM), where the model predicts randomly masked tokens based on their surrounding context. RoBERTa [2] refined BERT's methodology by removing the next sentence prediction objective, adopting dynamic masking, and substantially increasing training data and batch sizes, resulting in significant performance improvements.

## 3. Multilingual NLP Models

Early approaches to multilingual NLP relied on parallel corpora and explicit cross-lingual supervision. Multilingual BERT (mBERT) demonstrated that a single model trained on concatenated monolingual corpora from 104 languages could develop cross-lingual representations without explicit alignment. XLM [6] enhanced this approach by incorporating translation language modeling (TLM), which leverages parallel data to align representations across languages.

## 4. XLM-RoBERTa: Foundations and Innovation

XLM-RoBERTa [3] represents a convergence of RoBERTa's training methodology with multilingual pre-training at unprecedented scale. Unlike its predecessors, XLM-RoBERTa was trained on 2.5TB of newly created clean CommonCrawl data spanning 100 languages, making it the largest multilingual corpus utilized for pre-training. By adopting RoBERTa's refined training approach while eliminating the TLM objective in favor of scaled MLM pre-training, XLM-RoBERTa established new benchmarks for cross-lingual transfer.

The model's innovation lies not merely in scaling existing approaches but in demonstrating that robust cross-lingual representations can emerge from masked language modeling alone when applied to sufficiently diverse multilingual data. This finding challenges previous assumptions about the necessity of parallel data or explicit cross-lingual objectives for effective representation alignment across languages.

Recent work by Abdullah et al.[7] demonstrates the adaptability of RoBERTa through a fine-tuning strategy for Named Entity Recognition in Central-Kurdish, a low-resource language. By modifying the architecture with zero-initialized attention and training on a newly constructed corpus, the study establishes a new performance benchmark for Kurdish NLP, further validating the role of large multilingual models in overcoming linguistic disparities.

This is further supported by findings from SemEval-2023, where Höfer and Mottahedin [8] reported that while XLM-RoBERTa offers strong cross-lingual capabilities, it underperformed compared to monolingual models on English-specific NER tasks.

## C. Methodology

This review employed a systematic approach to identify relevant literature examining XLM-RoBERTa and comparative multilingual models. The search strategy encompassed multiple academic databases including ACL Anthology, IEEE Xplore, arXiv, Google Scholar, and Semantic Scholar. Primary search terms included "XLM-RoBERTa," "cross-lingual language models," "multilingual transformers," and "comparative multilingual NLP models."

Studies were selected based on the following inclusion criteria , Peer-reviewed journal articles, conference proceedings, or preprints with substantial technical content with publications dated between January 2020 (following XLM-RoBERTa's introduction) and May 2025 which contain comparative analyses between XLM-RoBERTa and other multilingual models where research focusing on architectural modifications, fine-tuning approaches, or application domains

The extracted data was organized thematically to address our research questions, with particular attention to cross-cutting themes such as cross-lingual transfer, language resource disparities, and domain-specific applications. Quantitative results were synthesized through comparative tables, while qualitative findings were integrated into a narrative synthesis highlighting convergent and divergent perspectives on XLM-RoBERTa's capabilities.

## D. XLM-RoBERTa and Pre-training

### 1. Model Architecture

XLM-RoBERTa maintains the transformer-based architecture established by its predecessors, consisting of 12 layers, 768 hidden dimensions, and 12 attention heads in its base configuration, resulting in approximately 270 million parameters. The large variant expands to 24 layers, 1024 hidden dimensions, and 16 attention heads, with approximately 550 million parameters. This architectural consistency with RoBERTa facilitates direct performance comparisons while highlighting the impact of multilingual pre-training data.

The model utilizes a shared subword vocabulary of 250,000 tokens generated using Sentence Piece tokenization with byte-pair encoding (BPE). This vocabulary size represents a significant expansion compared to mBERT (110K tokens) and XLM (200K tokens), accommodating the increased linguistic diversity of 100 languages. Notably, XLM-RoBERTa does not employ language embedding or identifiers during pre-training, relying instead on the model's capacity to implicitly distinguish languages through contextual patterns [9].

### 2. Pre-training Methodology

XLM-RoBERTa's pre-training methodology differs significantly from its predecessors in both scale and approach. The model was trained exclusively with the masked language modeling (MLM) objective on 2.5TB of data extracted from CommonCrawl and filtered through rigorous quality control measures. This represents a substantial increase from mBERT's Wikipedia-based training data (approximately 60GB) and even RoBERTa's 160GB corpus.

Following RoBERTa's training protocol, XLM-RoBERTa employed dynamic masking, where masking patterns are generated on-the-fly rather than during data preprocessing. Training utilized a peak learning rate of 0.0007 with linear warm-up over 10,000 steps, followed by polynomial decay. The model was trained with a batch size of 8192 sequences for 1.5 million steps, equivalent to approximately 300 epochs over the training data.

### 3. Comparative Pre-training Approaches

Comparative analyses reveal key distinctions between XLM-RoBERTa's pre-training approach and those of other multilingual models:

**Table 1.** Key distinction between XLM-RoBERTa with other Models

| Model | Training Data | Languages | Objectives | Tokenization | Special Features |
|---|---|---|---|---|---|
| mBERT | Wikipedia (60GB) | 104 | MLM, NSP | WordPiece | Language-agnostic |
| XLM | Wikipedia (100GB) | 15-100 | MLM, TLM, CLM | BPE | Language embeddings |
| XLM-R | CommonCrawl (2.5TB) | 100 | MLM only | SentencePiece | Balanced sampling |
| mBART | CC25 (1TB) | 25 | Seq2Seq denoising | SentencePiece | Encoder-decoder |
| mT5 | mC4 (8TB) | 101 | Span corruption | SentencePiece | Encoder-decoder |

Notably, XLM-RoBERTa demonstrates that explicit cross-lingual objectives like translation language modeling (TLM) are not essential for effective cross-lingual transfer when sufficient multilingual data is employed. This finding challenges earlier assumptions about the necessity of parallel data for cross-lingual alignment, suggesting that shared subword vocabulary and large-scale pre-training can naturally induce cross-lingual representation.

## E. Performance Metrics

The performance metrics used in the effective analysis such as Precision, Accuracy, F1-Score, and Recall are described bellow with their mathematical representation [9].

### 1. Accuracy

Accuracy Metric is defined as ration of correct classification outcome to the total number of outcomes and is mathematically denoted as:

$$Accuracy = \frac{TP_w + TN_w}{TP_w + TN_w + FP_w + FN_w} \quad (2)$$

**2. Precision**

The precision measure the positive instance during classification and is mathematically represented as:

$$Precision = \frac{TP_w}{TP_w + FP_w} \quad (3)$$

**3. F1-Score**

The F1-Score computes the average of call and precision, which is mathematically denoted as:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

**4. Recall**

Recall measure the ration of the true positive to the total of true positive and false negative, which is mathematically denoted as:

$$Recall = \frac{TP_w}{TP_w + FN_w} \quad (5)$$

Where TP_w denotes the true positive, TN_w denoted the true negative, FP_w denoted the false positive and FN_w denotes the false negative .

## F. Literature Review

Haq et al. [10] addresses the critical challenge of fake news detection in Malayalam YouTube comments through an innovative Multi-Pooling Feature Fusion (MPFF) approach. Utilizing XLM-RoBERTa, the study tackles both binary and multiclass classification tasks, achieving a macro-averaged F1 score of 0.874 in binary classification and 0.628 in multiclass categorization. The research stands out by implementing a sophisticated feature extraction strategy combining [CLS] token, mean, and max pooling techniques, demonstrating superior performance across traditional machine learning, deep learning, and transformer-based models. The work provides crucial insights into handling fake news detection in low-resource languages, particularly highlighting the complexities of identifying nuanced misinformation in Malayalam social media discourse.

Kodali et al. [11] presented a novel approach to hate speech detection and target identification in Devanagari-script languages, specifically focusing on Hindi and Nepali, published in the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL). The researchers developed a hybrid Attention BiLSTM-XLM-RoBERTa architecture, leveraging XLM-RoBERTa's multilingual embeddings pre-trained on 100 languages and fine-tuned on a dataset of Nepali and Hindi texts. The model achieved impressive performance, securing a Macro F1 score of 0.7481 for hate speech detection (Task B) and 0.6715 for target identification (Task C), demonstrating its effectiveness in capturing complex linguistic nuances. The most significant methodological contribution lies in the integration of an attention mechanism with BiLSTM and XLM-RoBERTa

embeddings, enabling the model to focus on critical contextual cues and sequential dependencies unique to Devanagari-scripted languages. By addressing the gap in multilingual hate speech analysis for low-resource language contexts, the research provides a robust framework for automated content moderation across diverse linguistic environments.

Manukonda & Kodali [12] presented an innovative approach to language identification across Devanagari-scripted languages, published in the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL). The researchers developed a hybrid Attention BiLSTM-XLM-RoBERTa model that effectively distinguishes between closely related languages including Nepali, Marathi, Sanskrit, Bhojpuri, and Hindi. Utilizing XLM-RoBERTa base embeddings pre-trained on 100 languages and fine-tuned on a comprehensive dataset of 52,416 training samples, the model achieved an exceptional performance with a 0.9974 Macro F1-score on the test set. The most significant methodological contribution lies in the integration of an attention mechanism with BiLSTM and XLM-RoBERTa embeddings, enabling the model to capture nuanced linguistic features across these closely related Devanagari-script languages. By addressing the complex challenges of language identification in multilingual contexts, the research provides a robust framework for context-aware natural language understanding systems, ultimately ranking 5th in the competition with minimal differences from the top entries.

Azadi et al. [13] conducted a groundbreaking study on bilingual sexism classification published in the EXIST Lab at CLEF 2024, focusing on detecting and characterizing sexist content in English and Spanish social media tweets. Utilizing XLM-RoBERTa, a multilingual transformer model pre-trained on 100 languages, the researchers fine-tuned the model on a dataset of over 10,000 tweets annotated by six crowdsourcing annotators. The study addressed two primary tasks: sexism identification and intention detection, achieving 4th place in the soft-soft evaluation for sexism identification and 2nd place for source intention. The most significant methodological innovation lies in the novel approach of incorporating the "learning with disagreements" paradigm, which captures multiple annotator perspectives by using majority voting and including diverse annotator votes. By leveraging XLM-RoBERTa's robust multilingual capabilities and exploring few-shot learning with GPT-3.5, the research provides critical insights into automated sexism detection across

Ali Al-Laith's [14] research paper, presented at the Joint Workshop on Financial Technology and Natural Language Processing, investigates the effectiveness of large language models (LLMs) for financial causality detection across English and Spanish datasets. The study employed a comprehensive approach using both generative (GPT-4o) and discriminative (XLM-RoBERTa and BERT) models to address a hybrid question-answering task involving extractive answers from financial disclosures. By evaluating performance using Semantic Answer Similarity (SAS) and Exact Match (EM) metrics, the research highlighted the complementary strengths of fine-tuned pre-trained language models and generative approaches. The XLM-RoBERTa-large model emerged as the top performer, achieving 5th place in English and 4th place in Spanish, while

demonstrating the potential of generative models like GPT-4o in few-shot learning scenarios.

Aamir et al. [15] introduced an innovative transformer-based approach to topic modeling for Urdu, a low-resource language with complex linguistic characteristics. By integrating BERTopic, XLM-R, and GPT frameworks, the researchers developed a sophisticated method for extracting nuanced themes from Urdu text. Utilizing the newly created LUCTM-24 dataset of 10,000 documents, the study demonstrated significant improvements over traditional topic modeling techniques like Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). The proposed approach achieved a 0.05 improvement in coherence and a diversity score of 0.87, highlighting the potential of transformer models in capturing the complex linguistic nuances of morphologically rich languages.

Ortame et al. [16] presented a comprehensive approach to hate speech detection on social media, focusing on tweets related to migrants and ethnic minorities in Italian. The study explored two innovative methodological approaches: an attention-based Bidirectional LSTM (AT-BiLSTM) and fine-tuned XLM-RoBERTa models. By analyzing 20.4 million tweets and utilizing two distinct datasets (EVALITA and IstatHate), the researchers developed a sophisticated Hate Speech Index (HSI) that captures the dynamics of hate speech over time. The best-performing model, XLM-RoBERTa-large, demonstrated superior performance in detecting nuanced hate speech, with the ability to correlate hate speech peaks with significant social events.

Chu Duong Huy Phuoc's [17] research paper, presented at the CoMeDi Shared Task Workshop, addresses the challenge of predicting disagreement rankings in multilingual word-in-context judgments. The study leverages the paraphrase-xlm-r-multilingual-v1 model, which is based on XLM-RoBERTa, to generate semantic embeddings for context pairs. By combining these embeddings with a deep neural regression model featuring batch normalization and dropout, the researchers developed a sophisticated approach to capturing semantic disagreements across seven languages. The approach achieved competitive performance, ranking 3rd out of 7 teams in the evaluation phase, with a notable average Spearman correlation of 0.124. The research highlights the potential of multilingual embeddings and robust neural architectures in handling complex semantic similarity tasks.

Yadagiri, Krishna, and Pakray [18] conducted a comprehensive study on AI-generated text detection, published in the 1st Workshop on GenAI Content Detection (GenAIDetect), focusing on leveraging transformer-based models for multilingual essay authenticity verification. The researchers employed DistilBERT for English and XLM-RoBERTa for Arabic text classification, utilizing a benchmark dataset comprising essays from various AI models like GPT-3.5-Turbo, GPT-4, Gemini, and LLaMa. Using a binary classification approach, the team achieved a recall of 0.825, ranking 18th in the English sub-task with a 0.77 accuracy and 20th in the Arabic sub-task with a 0.59 accuracy. The most significant methodological contribution lies in the integration of transformer-based models with additional linguistic features, such as average line length, vocabulary richness, and stop word frequency, to enhance AI-generated text detection. By demonstrating the potential of transformer architectures in distinguishing between human-authored and AI-

generated content, the research provides a foundational approach for academic integrity verification across multilingual contexts.

Nwaiwu and Jongsawat [19] conducted a comprehensive study examining XLM-RoBERTa's performance in hate speech detection for code-switched Spanglish text, published as a preprint on Preprints.org. The researchers utilized the SemEval2020 challenge dataset, which comprises 11,999 training, 3,000 validation, and 6,500 test samples of Spanish-English code-switched social media content. Employing a multilingual transformer model pre-trained on extensive multilingual datasets, they achieved exceptional performance metrics, with XLM-RoBERTa attaining 96.14% accuracy, 96.16% precision, 96.14% recall, and a 96.12% F1-score in detecting hate speech. The study's primary innovation lies in its rigorous comparative analysis between transformer-based and traditional machine learning models, demonstrating XLM-RoBERTa's superior capability in capturing linguistic nuances within complex code-switched environments. By incorporating hate lexicons and implementing comprehensive preprocessing techniques, the research provides critical insights into multilingual natural language processing challenges, particularly in hate speech detection across linguistically dynamic contexts.

Mathur and Shrivastava [20] conducted a comprehensive study on context-aware transformer models for ambiguous word classification in code-mixed sentiment analysis, published in the Journal of Information Systems Engineering and Management. The researchers employed four transformer-based models (XLM-RoBERTa, IndicBERT, DistilBERT, and TinyBERT) to address the challenge of word sense disambiguation in multilingual contexts. Utilizing extensive data preprocessing techniques including stopword removal, stemming, and lemmatization, they evaluated the models' performance on ambiguous word classification across various metrics. The most significant methodological contribution lies in the comparative analysis of transformer models, with DistilBERT emerging as the top performer, achieving an impressive 88.75% accuracy, 92.87% precision, 88.75% recall, and a 90.39% F1-score. XLM-RoBERTa, while effective with multilingual corpora, showed moderate precision and high recall (87.04% accuracy), highlighting the nuanced capabilities of different transformer architectures in handling linguistic ambiguity. The research provides critical insights into the potential of context-aware models for improving sentiment analysis in code-mixed language environments.

Ragab et al. [21] presented a comprehensive study at the 1st International Workshop on Nakba Narratives as Language Resources, focusing on multilingual propaganda detection using transformer-based models including mBERT, XLM-RoBERTa, and mT5. The research employed a balanced dataset of 13,500 Facebook posts spanning five languages (Arabic, English, Hebrew, French, and Hindi), with a particular focus on the framing of the Israeli War on Gaza. The mT5 model demonstrated exceptional performance, achieving an outstanding accuracy of 99.61% and an F1-score of 0.9961, highlighting its text-to-text framework's ability to effectively capture linguistic nuances across multiple languages. The study emphasized the critical role of data balancing techniques, showing significant performance improvements for mBERT and XLM-RoBERTa when addressing class imbalances, with final accuracies of 92.0% and 89.51% respectively.

Putra and Yulianti [22] conducted a comprehensive study on closed-domain question answering (QA) for educational websites, focusing on the Universitas Indonesia website. The research explored the effectiveness of transfer learning strategies using three BERT-based models: IndoBERT, RoBERTa, and XLM-RoBERTa. By leveraging the SQuAD dataset for transfer learning, the researchers demonstrated significant improvements in QA model performance. The XLM-RoBERTa base model emerged as the top performer, achieving an F1 score of 61.72% and highlighting the potential of transfer learning in improving information retrieval from complex educational platforms.

Yuxin Cai's [23] research, published in the Transactions on Computer Science and Intelligent Systems Research, presents a novel approach to Aspect Category Sentiment Analysis (ACSA) for Cantonese restaurant reviews using XLM-RoBERTa. The study focuses on a multilingual transformer model applied to a low-resource language, utilizing a custom-constructed dataset of 7,473 Cantonese restaurant reviews from OpenRice. The model was fine-tuned with specific configurations, including an Adam optimizer, a learning rate of 1e-5, and training over 5 epochs, demonstrating remarkable performance across five aspect categories: food, service, ambience, price, and timeliness. The XLM-RoBERTa model achieved a superior weighted accuracy of 75.17%, significantly outperforming baseline models such as SVM, Naive Bayes, and BERT, and highlighted the model's potential for sentiment analysis in linguistically complex, low-resource languages. The most significant contribution lies in creating the largest Cantonese review dataset for ACSA and demonstrating the effectiveness of transformer-based models in capturing nuanced linguistic features.

Elisa Di Nuovo, Emmanuel Cartier, and Bertrand De Longueville's [24] research paper addresses the critical challenge of multilingual sentiment analysis by introducing XLM-RLnews-8, a model based on XLM-RoBERTa-Large. The study focuses on domain-adapting the model to news articles across the 24 official European Union languages and fine-tuning it for tripartite sentiment analysis. By leveraging a novel multilingual news dataset and the Unified Multilingual Sentiment Analysis Benchmark (UMSAB), the researchers developed a model that demonstrates competitive performance across different languages and domains. The research contributes to the field by highlighting the importance of domain adaptation, multilingual capabilities, and sophisticated sentiment analysis techniques.

Yulianti et al. [25] presented a groundbreaking study on Legal Entity Recognition (LER) in Indonesian court decision documents, introducing a novel dataset called IndoLER and exploring transformer-based models for named entity recognition. The research focused on a comprehensive dataset of approximately 1,000 criminal court decision documents, annotated with 20 fine-grained legal entities, totaling around 6 million words and 25,000 annotated entities. The study evaluated multiple transformer models, including multilingual (M-BERT, XLM-RoBERTa) and monolingual (IndoBERT, IndoRoBERTa) approaches. The XLM-RoBERTa large model emerged as the top performer, achieving an impressive F1-score of 0.9295, outperforming previous baseline models like BiLSTM and BiLSTM-CRF by 7.9% and 2.64% respectively. The research not only advances named entity

recognition in Indonesian legal documents but also provides a publicly available dataset for future research.

Badawi et al. [26] introduce KurdiSent, the first comprehensive, manually annotated sentiment analysis dataset for the Kurdish language, addressing a critical gap in natural language processing resources for low-resource languages. The dataset comprises 12,309 tweets collected during the COVID-19 pandemic, categorized across five domains: social, art, health, technology, and news, with sentiments classified as positive, negative, or neutral. The research methodology involved rigorous data preprocessing, including removal of non-Kurdish characters, special characters, and elongated words, and employed three academic experts for manual annotation. The annotation process utilized the Doccano tool, achieving a Kappa coefficient ranging from 0.78 to 0.89, indicating strong inter-annotator agreement. Experimental evaluation using various machine learning and deep learning models demonstrated the dataset's effectiveness, with the XLM-R model achieving the highest accuracy of 85% across sentiment classification tasks. The most significant methodological contribution is the creation of a structured, publicly accessible corpus that provides a foundational resource for sentiment analysis research in the Kurdish language, particularly the Sorani dialect, which is the most widely spoken variant.

Jennifer D. et al. [27] conducted a comprehensive review of natural language processing (NLP) learning models, with a particular focus on sentiment analysis and sentence classification. The study analyzed 40 research papers, with a significant emphasis on BERT (Bidirectional Encoder Representations from Transformers) and its variants. Out of the examined papers, 21 utilized BERT for text classification, highlighting its dominance in NLP tasks. The research provides an in-depth exploration of advanced sentiment analysis approaches, comparing BERT with other language models and investigating proprietary BERT-based models across various applications including business, finance, social media analysis, and emotional understanding.

Dorkin and Sirts [28] presented an innovative approach to applying transformer-based models to ancient and historical languages through the adapters framework. The research addressed the challenge of applying modern language models to low-resource historical languages by developing a uniform, computationally lightweight method using parameter-efficient fine-tuning. Utilizing XLM-RoBERTa, the researchers created a flexible system that could be applied across 16 different ancient and historical languages, achieving notable performance in tasks such as morphological annotation, POS-tagging, lemmatization, and gap-filling. Their submission secured second place overall, with a first-place finish in word-level gap-filling, demonstrating the feasibility of adapting pre-trained language models to historical linguistic contexts.

Masaling and Suhartono [29] present a novel approach to structured sentiment analysis (SSA) utilizing pre-trained RoBERTa and XLM-RoBERTa models to extract and analyze opinion tuples across diverse datasets. The study focuses on addressing limitations in traditional sentiment analysis methods by developing a comprehensive system that can extract holder, target, and expression components while simultaneously classifying sentiment polarity. The experimental setup involved three datasets: OpeNEREN, MPQA, and DSUnis, with models configured

using 12 layers, 768 hidden units, and 12 attention heads. The XLM-RoBERTa model demonstrated exceptional performance on the OpeNEREN dataset, achieving a sentiment graph F1 (SF1) score of 64.6%, while the RoBERTa model showed consistent performance across MPQA (25.3% SF1) and DSUnis (29.9% SF1) datasets. The most significant methodological contribution was the development of a dual-module approach combining a node extractor and edge predictor, which enables comprehensive sentiment tuple extraction and relationship identification, providing a more nuanced understanding of sentiment structures beyond traditional classification methods.

Katyshev, Anikin, and Zubankov [30] investigated the application of XLM-RoBERTa in efficiently constructing knowledge bases, specifically focusing on ontology creation using Russian language programming books, published as a book chapter. The researchers employed a multilingual transformer model with training parameters including a learning rate of 1e-5, batch size of 16, and maximum sequence length of 512 tokens, fine-tuned over 3 epochs. Demonstrating superior performance, XLM-RoBERTa achieved an impressive F1-score of 0.87, outperforming other state-of-the-art models like BERT-base Multilingual, XLM, mBART, and RuGPT-3 in concept and relationship extraction. The study's primary innovation lies in leveraging the model's bidirectional transformer architecture and multilingual pre-training to automate ontology construction, highlighting its potential for efficient knowledge base development across different domains and languages. The research critically examines both the advantages of XLM-RoBERTa, such as language independence and contextual understanding, and its limitations, including computational intensity and domain adaptation challenges.

Davide Colla, Matteo Delsanto, and Elisa Di Nuovo's [31] research paper, presented at the NLP4CALL workshop, addresses the challenge of multilingual grammatical error detection across five languages (Czech, English, German, Italian, and Swedish). The authors employed a token classification approach using XLM-RoBERTa, developing a binary classification model to identify correct or incorrect tokens in sentences. Their methodology involved fine-tuning five separate language models and exploring two experimental settings: training on the provided training set and training on combined training and development sets. The research demonstrated the effectiveness of contextual representations in detecting grammatical errors, with the proposed system achieving the highest scores on five out of six test sets. A key innovation was the approach's ability to leverage contextual information to detect a broader range of error types compared to traditional token-counting methods.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain's [32] research paper, presented at the Third Workshop on Speech and Language Technologies for Dravidian Languages, addresses the critical challenge of fake news detection in the Malayalam language. The study leverages the XLM-RoBERTa base model to develop an innovative approach for distinguishing between genuine and fake news articles. By fine-tuning the multilingual model on a Malayalam dataset and employing Bayesian optimization for hyperparameter tuning, the researchers achieved a remarkable macro-averaged F-Score of 87%. This approach not only demonstrates the potential of transformer-based models in handling low-resource

languages but also highlights the importance of language-specific modeling for effective fake news detection.

Edgar Andrés Santamaría's [33] research paper, presented at the 17th International Workshop on Semantic Evaluation (SemEval-2023), addresses the challenging task of fine-grained multilingual named entity recognition (NER) using XLM-RoBERTa. The study focuses on the MultiCoNER shared task, which involves a complex taxonomy of 36 tags across 12 languages. By implementing a sequence labeling fine-tuning approach, the researcher developed a baseline system that leverages transfer learning from pre-trained Named Entity Recognition models and cross-lingual knowledge. The approach demonstrated varying performance across languages, with the most notable challenge being a fine-grained entity taxonomy and simulated errors in the test set, highlighting the complexity of multilingual NER tasks.

Rahul Mehta and Vasudeva Varma's [34] research paper, presented at SemEval-2023, addresses the challenging task of multilingual complex named entity recognition (NER) using XLM-RoBERTa. The study focuses on the MultiCoNER II shared task, which involves identifying named entities across 12 languages with a complex taxonomy of 30 entity types. By fine-tuning the XLM-RoBERTa base model on each language dataset, the researchers demonstrated the model's effectiveness in handling complex named entities across diverse linguistic contexts. The approach leveraged cross-lingual representation learning, achieving notable performance across different languages and entity types, with particularly strong results in the Creative Work category.

Hämmerl, Libovický, and Fraser [35] conducted a groundbreaking study on combining static and contextual multilingual embeddings, published as an arXiv preprint, focusing on improving cross-lingual representation learning using XLM-RoBERTa. The researchers extracted static embeddings for 40 languages from XLM-R, validating them through cross-lingual word retrieval and alignment using VecMap, and then applied a novel continued pre-training approach to XLM-R. Their methodology leveraged a unique alignment technique that does not require parallel text, instead using an alignment loss that combines masked language modeling (MLM) with either Mean Squared Error (MSE) or Deep Canonical Correlation Analysis (DCCA). The most significant methodological contribution lies in the approach of indirectly transferring alignment knowledge from well-aligned static embeddings to the contextual model. Evaluated across multiple complex semantic tasks like question answering, sequence labeling, and sentence retrieval, the proposed method demonstrated improved cross-lingual performance, particularly when using the DCCA alignment loss, achieving an average performance improvement across various downstream tasks and highlighting the potential for enhancing multilingual representation learning without relying on parallel corpora.

Sirusstara et al. [36] conducted a comprehensive study on clickbait headline detection in Indonesian news sites, published as a conference paper, focusing on comparing the performance of various transformer-based models, particularly RoBERTa variants. The researchers utilized the CLICK-ID dataset containing 6,632 annotated news headlines, experimenting with four pre-trained models including IndoBERT and Cahya Wirawan's RoBERTa models. Using a methodology involving

exploratory data analysis, preprocessing, and model fine-tuning, they achieved impressive results, with cahya/XLM-RoBERTa-large and indobenchmark/IndoBERT-p1 performing exceptionally well, reaching approximately 92% accuracy. The most significant methodological contribution lies in the comparative analysis of transformer models' performance in detecting clickbait headlines, highlighting the nuanced differences between various pre-trained models. The study not only provides insights into clickbait detection techniques but also demonstrates the effectiveness of transfer learning and transformer architectures in natural language processing tasks for the Indonesian language, ultimately recommending IndoBERT-p1 as the most resource-efficient model while acknowledging XLM-RoBERTa's consistent performance across validation and unseen datasets.

Usama Yaseen and Stefan Langer's [37] research paper, presented at the Scientific Document Understanding workshop (SDU@AAAI-22), addresses the challenge of multilingual acronym extraction across six languages (Danish, English, French, Spanish, Persian, and Vietnamese) in scientific and legal domains. The authors employed a BiLSTM-CRF model with XLM-RoBERTa embeddings, introducing a novel domain adaptive pretraining approach to improve multilingual acronym extraction performance. Their key methodology involves pretraining the XLM-RoBERTa model on the task-specific corpus to better adapt the contextual representations to scientific and legal domains. The research demonstrates the effectiveness of a unified multilingual model, achieving competitive performance across languages, with a notable finding that training across multiple languages enables more effective cross-lingual transfer compared to language-specific models.

Nair et al. [38] present ColBERT-X, a cross-language dense retrieval model published in the domain of information retrieval, specifically addressing cross-language information retrieval (CLIR) challenges. The research leverages XLM-RoBERTa (XLM-R), a large multilingual transformer encoder capable of processing multiple languages, with a parameter size of approximately 550 million parameters. The experimental setup focused on retrieving documents across seven languages (Chinese, Persian, French, German, Italian, Russian, and Spanish) using the MS MARCO passage ranking dataset for training and HC4 and CLEF newswire collections for evaluation. The authors introduced two innovative training strategies: zero-shot and translate-train, demonstrating significant improvements over traditional BM25 query translation baselines, with Mean Average Precision (MAP) gains ranging from 4% to 15% across different language pairs. The most significant methodological contribution is the generalization of the ColBERT retrieval approach to support cross-language information retrieval by utilizing a multilingual encoder and novel transfer learning techniques.

Thet and Pa [39] conducted a comprehensive study on enhancing sentence classification performance by exploring RoBERTa's intermediate layers for transfer learning in the context of Myanmar language sentiment analysis. The research addresses the critical challenge of limited labeled data in low-resource language domains by investigating alternative pooling strategies for leveraging pre-trained language models. The experimental setup utilized a dataset of 72K sentences crawled from Myanmar Celebrity Facebook comments, with

experiments conducted using two pre-trained models: language-specific MyanBERTa and multilingual XLM-RoBERTa-base. The authors proposed two innovative pooling approaches - LSTM pooling and weighted pooling - to extract semantic knowledge from RoBERTa's intermediate layers, moving beyond traditional approaches that only utilize the final output layer. The most significant methodological contribution was demonstrating that different pooling strategies can effectively capture nuanced representations across intermediate layers. Experimental results revealed that the LSTM pooling method improved accuracy for the language-specific model, while the weighted pooling strategy with XLM-RoBERTa-base outperformed other approaches, achieving the highest classification performance.

Pannach and Donicke [40] conducted a comprehensive study on German verbal idiom disambiguation, published at the KONVENS 2021 Shared Task, focusing on leveraging XLM-RoBERTa for distinguishing between literal and figurative uses of verbal idioms. The researchers employed a sophisticated approach that involved semi-automatically extending the training data, collecting additional examples from Wiktionary and web sources to address the limited dataset. Utilizing XLM-RoBERTa, they achieved a remarkable 0.7622 F1-score, significantly outperforming the baseline linear SVM model's 0.5696 F1-score. The most significant methodological contribution lies in the novel data collection strategy and the application of a powerful transformer model to the nuanced task of idiom disambiguation. Interestingly, the study revealed that the XLM-RoBERTa model performed best with the original training data, demonstrating the model's inherent capability to capture contextual subtleties in linguistic expressions. The research provides critical insights into the potential of advanced transformer models for handling complex linguistic phenomena, particularly in distinguishing between literal and figurative language use.

Bing Li, Yujie He, and Wenjin Xu's [41] research paper, published on arXiv, introduces a novel approach to cross-lingual Named Entity Recognition (NER) using XLM-RoBERTa. The authors leverage parallel corpora and a sophisticated entity alignment model to transfer NER knowledge across languages with minimal human annotation. Their methodology focuses on projecting named entities from English to target languages (German, Spanish, Dutch, and Chinese) using an alignment model built on XLM-RoBERTa. The approach demonstrates competitive performance, particularly for Chinese, with F1 score improvements ranging from 2.4% to 6.4% over zero-shot transfer baselines. The most significant methodological innovation lies in their entity alignment technique, which transforms the cross-lingual transfer problem into a token classification task, effectively addressing previous limitations in translation-based approaches by maintaining natural language nuances and fluency.

Stefan Ziehe, Franziska Pannach, and Aravind Krishnan's [42] research paper, presented at the First Workshop on Language Technology for Equality, Diversity and Inclusion, addresses the novel task of hope speech detection across English, Malayalam, and Tamil languages. The study explores machine learning techniques for identifying positive and supportive language in social media comments, with a focus on leveraging XLM-RoBERTa's multilingual capabilities. By implementing baseline models (SVM and Complement Naive Bayes) and a cross-lingual transfer

learning approach, the researchers demonstrated the effectiveness of transformer-based models in detecting hope speech. The XLM-RoBERTa model achieved top rankings for English and Malayalam, highlighting the potential of advanced language models in promoting positive online communication.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau's [43] research paper explores the impact of scaling multilingual transformer models for masked language modeling. The study introduces two significant models, XLM-RXL (3.5B parameters) and XLM-RXXL (10.7B parameters), which demonstrate remarkable improvements in cross-lingual understanding and performance across multiple languages. By scaling the XLM-R model's capacity, the researchers achieved a 1.8% to 2.4% average accuracy improvement on XNLI benchmarks and outperformed the RoBERTa-Large model on English GLUE tasks while handling 99 additional languages. The research provides crucial insights into the potential of larger-capacity multilingual models to excel in both high-resource and low-resource language contexts.

Xie et al. [44] presented a sophisticated approach to multilingual and cross-lingual word-in-context (WiC) disambiguation at SemEval-2021 Task 2, focusing on determining whether a target word maintains the same meaning across two different contexts. The research leveraged the XLM-RoBERTa transformer model, implementing innovative techniques to enhance performance across both multilingual (same-language) and cross-lingual (different-language) tasks. By introducing special input tagging, target word embedding concatenation, and advanced training strategies like WordNet data augmentation, adversarial training, and pseudo-labeling, the team achieved championship performance in four cross-lingual tasks and competitive rankings in multilingual tasks. The XLM-RoBERTa Large model, combined with these techniques, demonstrated remarkable adaptability in capturing nuanced word meanings across diverse linguistic contexts.

Conneau et al. (2020) present XLM-RoBERTa (XLM-R), a groundbreaking multilingual masked language model trained on 100 languages using over two terabytes of filtered CommonCrawl data. The model comprises two variants: XLM-R Base with 270 million parameters and XLM-R with 550 million parameters, utilizing a Sentence Piece tokenizer and a large vocabulary of 250,000 tokens. The experimental setup focused on cross-lingual understanding tasks, including cross-lingual natural language inference (XNLI), named entity recognition, and question answering, with evaluation across multiple language benchmarks. The model demonstrated remarkable performance improvements, including a 14.6% average accuracy gain on XNLI, 13% average F1 score improvement on MLQA, and significant gains for low-resource languages like Swahili (15.7% accuracy improvement) and Urdu (11.4% accuracy improvement). The most significant methodological contribution was addressing the "curse of multilinguality" by scaling model capacity and utilizing extensive multilingual training data, proving it possible to create a single large model that performs competitively across multiple languages without sacrificing per-language performance.

**Table 2.** Summary of the work Performed by most of the research reviewed in this paper

| Authors | Year | Model Specifications | Benchmark Datasets | Performance Metrics | Identified Limitations |
|---|---|---|---|---|---|
| Haq et al. [10] | 2025 | XLM-RoBERTa base with Multi-Pooling Feature Fusion (MPFF) approach | 5,091 texts for binary classification; 2,100 texts for multiclass classification; Malayalam YouTube comments | Binary Classification: F1 Score 0.874 (Ranked 6th); Multiclass Classification: F1 Score 0.628 (Ranked 1st) | Limited dataset size; Class imbalance; Challenges in nuanced category detection; Difficulty with subtle misinformation |
| Kodali et al. [11] | 2025 | Hybrid Attention BiLSTM-XLM-RoBERTa; XLM-RoBERTa base embeddings; BiLSTM layer (hidden size 256, 2 layers); Attention mechanism | Devanagari-script languages (Hindi and Nepali); Task B: 19,019 training samples (16,805 Non-Hate, 2,214 Hate); Task C: 2,214 training samples | Task B (Hate Speech Detection): Macro F1 Score: 0.7481; Task C (Target Identification): Macro F1 Score: 0.6715 | Limited computational resources; Complexity of XLM-RoBERTa architecture; Challenges in capturing nuanced linguistic variations; Need for more extensive fine-tuning data |
| Manukonda & Kodali [12] | 2025 | Hybrid Attention BiLSTM-XLM-RoBERTa; XLM-RoBERTa base embeddings (768-dimensional); BiLSTM layer (hidden size 256, 2 layers); Attention mechanism | 5 Devanagari-scripted languages; Total training samples: 52,416 (Nepali: 12,543; Marathi: 11,034; Sanskrit: 10,996; Bhojpuri: 10,184; Hindi: 7,659) | Test Set Accuracy: 0.9974; Test Set Macro F1-Score: 0.9976; Ranked 5th in competition | Computational constraints limiting model size; Restricted to XLM-RoBERTa base model; Limited fine-tuning data for masked language model; Potential generalizability challenges |
| Azadi et al. [13] | 2025 | XLM-RoBERTa (Multilingual Transformer); Pre-trained on 100 languages; Fine-tuned on bilingual dataset | EXIST 2024 Dataset; 10,000+ tweets in English and Spanish; 6 annotators per tweet; Train: 6,920 tweets; Development: 1,038 tweets; Test: 2,076 tweets | Task 1 (Sexism Identification): 4th place in soft-soft evaluation, ICM-Hard Norm: 0.78, F1-Score: 0.78; Task 2 (Intention Detection): 2nd place in soft-soft evaluation, ICM-Hard Norm: 0.56, F1-Score: 0.48 | Limited performance on minority class samples; Challenges in capturing nuanced sexist intentions; Need for more sophisticated few-shot learning approaches; Potential annotator bias |
| Al-Laith [14] | 2025 | XLM-RoBERTa (base and large); BERT-base-multilingual-cased; GPT-4o generative model; Multilingual approach (English and Spanish) | FinCausal 2025 shared task; Financial disclosures; Extractive question-answering task | Semantic Answer Similarity (SAS): English: 0.96, Spanish: 0.98; Exact Match (EM): English: 0.7615, Spanish: 0.8084 | Performance variations across languages; Generative models' lower exact matching precision; Dependence on prompting techniques |

| Authors | Year | Model Specifications | Benchmark Datasets | Performance Metrics | Identified Limitations |
|---|---|---|---|---|---|
| Aamir et al. [15] | 2025 | BERTopic; XLM-R; GPT | LUCTM-24 (10,000 Urdu documents; 5,000,000 total words; Average 500 words per document; Unique vocabulary: 120,000 words) | Coherence Score: BERTopic + XLM-R + GPT: 0.62, LDA: 0.57, NMF: 0.55; Diversity Score: BERTopic + XLM-R + GPT: 0.87, LDA: 0.76, NMF: 0.72 | Computational complexity of transformer models; Limited annotated resources for Urdu; Difficulty in handling extreme linguistic variations; Need for domain-specific fine-tuning |
| Ortame et al. [16] | 2025 | Attention-based BiLSTM (AT-BiLSTM); XLM-RoBERTa (Base and Large) | 20.4 million Tweets (2018-2023); EVALITA 2020 HaSpeeDe 2 dataset; Custom IstatHate dataset | Macro F1 Score Across Test Sets: XLM-RoBERTa Large: Up to 0.811; AT-BiLSTM: Up to 0.773 | Computational complexity of large models; Subjectivity in hate speech labeling; Limited generalizability across different linguistic contexts; Potential bias in training data; Difficulty in capturing subtle forms of hate speech |
| Phuoc [17] | 2025 | Paraphrase-XLM-R-multilingual-v1; XLM-RoBERTa base model; Deep neural regression architecture; Multilingual approach (7 languages) | DWUG dataset; Languages: Chinese, English, German, Norwegian, Russian, Spanish, Swedish; Word-in-Context (WiC) task | Spearman's Rank Correlation; Average score: 0.124; Ranked 3rd out of 7 teams | Challenges with Latin-based languages; Reliance on embedding quality; Limited modeling of underlying disagreement causes |
| Yadagiri et al. [18] | 2025 | DistilBERT (English Sub-task); XLM-RoBERTa (Arabic Sub-task); Pre-trained on 100 languages; Additional semantic feature extraction | English Dataset: Train: 629 Human, 1467 AI; Dev: 1235 Human, 391 AI; Arabic Dataset: Train: 1145 Human, 925 AI; Dev: 182 Human, 299 AI; AI Models: GPT-3.5-Turbo, GPT-4, Gemini, LLaMa | English Sub-task: Accuracy: 0.77, Precision: 0.784, Recall: 0.82, F1-Score: 0.77; Arabic Sub-task: Accuracy: 0.59, Precision: 0.55, Recall: 0.56, F1-Score: 0.55 | XLM-RoBERTa not explicitly fine-tuned for Arabic; Performance variations across languages; Challenges in deep understanding of Arabic syntax and semantics |
| Nwaiwu & Jongsawat [19] | 2025 | XLM-RoBERTa (Multilingual Transformer) | SemEval2020 Challenge Dataset; Total Samples: 21,499 (11,999 training, 3,000 validation, | Accuracy: 96.14%; Precision: 96.16%; Recall: 96.14%; F1-Score: 96.12% | Challenges with slang and idiomatic expressions; Difficulties in handling semantic ambiguities; Performance variations across different |

| Authors | Year | Model Specifications | Benchmark Datasets | Performance Metrics | Identified Limitations |
|---|---|---|---|---|---|
| | | | 6,500 test); Code-switched Spanglish social media text | | transformer architectures |
| Mathur & Shrivastava [20] | 2025 | XLM-RoBERTa, IndicBERT, DistilBERT, TinyBERT | Not specified in detail | XLM-RoBERTa: Accuracy: 87.04%, Precision: 75.77%, Recall: 87.04%, F1-Score: 81.02%; DistilBERT: Accuracy: 88.75%, Precision: 92.87%, Recall: 88.75%, F1-Score: 90.39% | XLM-RoBERTa limited in precision; Challenges in capturing full contextual nuances; Variability in performance across different models |
| Ragab et al. [21] | 2025 | Multilingual transformer models: mBERT, XLM-RoBERTa, mT5 | 13,500 Facebook posts; 5 languages: Arabic, English, Hebrew, French, Hindi; Focused on Israeli War on Gaza narrative | Without Balancing: XLM-RoBERTa: 69.70% accuracy, F1-score 0.5595; With Balancing: XLM-RoBERTa: 89.51% accuracy, F1-score 0.8934 | Limited to five languages; Challenges in detecting subtle propaganda elements; Potential overfitting in XLM-RoBERTa |
| Putra & Yulianti [22] | 2025 | IndoBERT (base), RoBERTa (base), XLM-RoBERTa (base) | Edu-QA Dataset: 126 documents; 2,692 total question-answer pairs; Focused on Universitas Indonesia website | Transfer Learning Scenario: XLM-RoBERTa: F1-Score 61.72%; Non-Transfer Learning: XLM-RoBERTa: F1-Score 56.79% | Focused on a single educational website; Limited to base versions of transformer models |
| Cai [23] | 2024 | XLM-RoBERTa (multilingual transformer) | Custom OpenRice Cantonese Restaurant Review Dataset (7,473 reviews) | Weighted Accuracy: 75.17%; Weighted F1-Score: 74.58%; Weighted Precision: 75.13%; Per-category performance ranging from 69.03% to 79.80% | Increased computational complexity with longer token lengths; Potential ambiguity in classifying unmentioned aspect categories; Varied performance across different aspect categories |
| Di Nuovo et al. [24] | 2024 | XLM-RoBERTa-Large base; Domain-adapted to news articles; Fine-tuned for tripartite sentiment analysis; Covers 24 EU languages | UMSAB (Multilingual Twitter Dataset); IMDb Film Reviews; Multilingual News Headlines; EMM Multilingual News | Weighted F1 Scores: UMSAB Test Set: 70.61 (average); IMDb Dataset: 0.868; News Headlines: 0.697-0.729 | Performance variations across languages; Domain adaptation challenges; Dataset bias and class imbalance; Limited performance in some language contexts |

| Authors | Year | Model Specifications | Benchmark Datasets | Performance Metrics | Identified Limitations |
|---|---|---|---|---|---|
| | | | Dataset | | |
| Yulianti et al. [25] | 2024 | M-BERT, XLM-RoBERTa (base and large), IndoBERT (base and large), Indonesian RoBERTa | IndoLER dataset: 993 criminal court decision documents; 20 fine-grained legal entities; Approximately 6 million words; 24,845 annotated entities | XLM-RoBERTa Large: F1-score: 0.9295, Precision: 0.9124, Recall: 0.9472; Improvement Over Baselines: 7.9% over BiLSTM, 2.64% over BiLSTM-CRF | Focused only on criminal court documents; Potential bias due to data imbalance; Limited to Indonesian language |
| Badawi et al. [26] | 2024 | XLM-R, MT5, BERT, SVM, CNN-LSTM, CNN-RNN, Naive Bayes, Logistic Regression, Decision Tree, Random Forest | Total tweets: 12,309; Language: Kurdish (Sorani dialect); Collection period: COVID-19 pandemic; Sentiment classes: Positive, Neutral, Negative; Categories: Social, Art, Health, Technology, News | XLM-R (Accuracy: 85%), Precision: 0.85, 0.89, 0.85, Recall: 0.85, 0.89, 0.86 | Lack of Kurdish sentiment analysis datasets; Limited NLP resources for Kurdish language; Dialect variations; Subjective sentiment interpretation |
| Jennifer D. et al. [27] | 2024 | BERT-based models including XLM-RoBERTa | Various datasets including Twitter Sentiment, Amazon Product Reviews, IMDB Movie Reviews, SemEval Datasets | Accuracy in Sentiment Classification: Up to 92-98% for BERT-based models | Limited fine-grained sentiment datasets; Computational expense; Out-of-vocabulary words; Maximum token limitations |
| Dorkin & Sirts [28] | 2024 | XLM-RoBERTa (Multilingual RoBERTa) with Adapters Framework | 16 Ancient and Historical Languages including Ancient Greek, Classical Latin, Old Church Slavonic, Gothic, Vedic Sanskrit | Second Place Overall; First Place in Word-Level Gap Filling | Challenges with underrepresented scripts; Limited performance on some languages; Difficulty with highly diverse writing systems |
| Masaling & Suhartono [29] | 2024 | RoBERTa, XLM-RoBERTa: 12 layers, 768 hidden units, 12 attention | OpeNEREN (Hotel reviews), MPQA (News text), DSUnis | OpeNEREN Dataset: XLM-RoBERTa SF1: 64.6%, RoBERTa | Varying performance across different datasets; Challenges with uneven tuple distribution; |

| Authors | Year | Model Specifications | Benchmark Datasets | Performance Metrics | Identified Limitations |
|---|---|---|---|---|---|
| | | heads; Trained with AdamW optimizer; Learning rate: 1e-4 | (Online university reviews) | SF1: 59.9%; MPQA Dataset: RoBERTa SF1: 25.3%; DSUnis Dataset: RoBERTa SF1: 29.9% | Performance variations between RoBERTa and XLM-RoBERTa |
| Katyshev et al. [30] | 2023 | XLM-RoBERTa (Multilingual Transformer); Learning Rate: 1e-5; Batch Size: 16; Max Sequence Length: 512 tokens; Fine-tuned for 3 epochs | Russian language programming books; 70-30% train-test split; Domain: Programming and computer science | Precision: 0.89; Recall: 0.86; F1-Score: 0.87; Compared against BERT-base M, XLM, mBART, RuGPT-3 | High computational requirements; Model complexity; Potential challenges in domain adaptation; Sensitivity to noise in input data |
| Colla et al. [31] | 2023 | XLM-RoBERTa large model; Token classification architecture; Multilingual approach covering 5 languages | MultiGED shared task corpus; Languages: Czech, English, German, Italian, Swedish; Domains: Learner language texts | F0.5 score (precision-weighted); Achieved highest scores on 5/6 test sets; Significant improvement over baseline methods | Performance varies across languages; Challenges with low-resource languages; Difficulty with English REALEC dataset |
| Raja et al. [32] | 2023 | XLM-RoBERTa base model; Multilingual transformer architecture; Fine-tuned for Malayalam language | Malayalam fake news dataset; Social media platforms; Balanced dataset with original and fake news | Macro-averaged F-Score: 87%; Ranked 2nd in the shared task; Compared with mBERT and DistilBERT | Limited research on Malayalam fake news detection; Challenges in handling linguistic nuances; Dependence on dataset quality |
| Santamaría [33] | 2023 | XLM-RoBERTa base model; Sequence labeling approach; Fine-tuned on MultiCoNER dataset | 12 languages; Fine-grained taxonomy with 36 tags; Divided into 6 main groups; Includes scientific and complex entity recognition | Development Set Results: Highest: Ukrainian (F1: 0.631), Lowest: Chinese (F1: 0.488); Test Set Results: Highest: Swedish (F1: 27.96), Lowest: Chinese (F1: 8.06) | Fine-grained taxonomy challenges; Simulated errors in test set; Varying performance across languages; Low-context entity recognition |
| Mehta & Varma [34] | 2023 | XLM-RoBERTa base model; Fine-tuned on 12 language datasets; 30 complex named entity types | MultiCoNER v2 dataset; 12 languages including Bangla, Chinese, English, Farsi, French, German, Hindi, | Development Set F1-Scores: Highest: Hindi (69.04%), Lowest: Chinese (25.50%); Test Set Highlights: Hindi: 63.29%, | Varying performance across languages; Challenges with low-resource languages; Difficulty with complex entity types; Performance drop in test set |

| Authors | Year | Model Specifications | Benchmark Datasets | Performance Metrics | Identified Limitations |
|---|---|---|---|---|---|
| | | | Italian, Portuguese, Spanish, Swedish, Ukrainian; Includes corrupted and uncorrupted test sets | English: 52.08%, German: 55.54%, Spanish: 54.81% | |
| Hämmerl et al. [35] | 2022 | XLM-R base model (270M parameters); Static embeddings extracted from layer 6; 40 languages covered; Alignment techniques: VecMap, DCCA | Multilingual datasets across: Question Answering (XQuAD, TyDiQA), Sequence Labeling (PAN-X, UD-POS), Sentence Retrieval (Tatoeba) | Downstream Task Average Improvement: Base XLM-R: 60.62, +MLM: 62.86, +fasttextDCCA: 63.45, +X2S-MADCCA: 64.96 | Performance variations across language families; Sensitivity of certain tasks to embedding changes; Computational resource requirements; Potential bias towards high-resource languages |
| Sirusstara et al. [36] | 2022 | IndoBERT-p1, cahya/XLM-RoBERTa-large, cahya/BERT-base, cahya/RoBERTa-base, Multinomial Naive Bayes (baseline) | CLICK-ID Dataset; Total Headlines: 6,632; Training Set: 4:1 split; Includes clickbait and non-clickbait labels | Unseen Set Results: IndoBERT-p1: 92.32% F1-score, cahya/XLM-RoBERTa-large: 91.81% F1-score, cahya/BERT-base: 90.64% F1-score, Multinomial NB: 84.03% F1-score | Potential overfitting due to limited word occurrences; Resource constraints for model training; Bias from minimal word/topic representation |
| Yaseen & Langer [37] | 2022 | XLM-RoBERTa embeddings; BiLSTM-CRF architecture; Multilingual model covering 6 languages | Multilingual acronym extraction corpus; Domains: Scientific and Legal; Languages: Danish, English, French, Spanish, Persian, Vietnamese | Overall F1-score improved from 0.854 to 0.866; Domain adaptive pretraining showed incremental improvements; Best performance achieved by training across multiple languages | Performance varies across languages; Limited performance for low-resource languages (Persian, Vietnamese); Sensitivity to domain mismatch |
| Nair et al. [38] | 2022 | XLM-RoBERTa (large) multilingual encoder; ~550 million parameters; Supports multiple languages | Training: MS MARCO passage ranking dataset; Evaluation: HC4 and CLEF newswire | Mean Average Precision (MAP); Improvements over BM25 query translation: 4-15%; Statistically | Requires substantial computational resources; Large index size (154GB for Chinese collection); Performance varies across different machine |

| Authors | Year | Model Specifications | Benchmark Datasets | Performance Metrics | Identified Limitations |
|---|---|---|---|---|---|
| | | | collections; Languages: Chinese, Persian, French, German, Italian, Russian, Spanish | significant gains in most language pairs | translation models; Limited to queries of 32 tokens or less |
| Thet & Pa [39] | 2022 | MyanBERTa (language-specific), XLM-RoBERTa-base (multilingual); 12 layers; Adam optimizer | Source: Myanmar Celebrity Facebook comments; Total sentences: 72K; Sentiment classes: Positive, Neutral, Negative | XLM-RoBERTa with weighted pooling outperformed other approaches | Focused on a single language domain; Limited dataset size; Specific to sentiment analysis task |
| Pannach & Donicke [40] | 2021 | XLM-RoBERTa Base; Sequence Length: 128; Batch Size: 64; Learning Rate: 2e-5; Optimizer: Adam | Total Samples: 9.9k; Training Set: 6,902 samples (Literally: 1,172, Figuratively: 5,705); Additional Dataset: 1,166 samples; Test Set: 1,511 samples | XLM-RoBERTa Performance: F1-Score: 0.7622; Literally: High recall; Figuratively: 0.9369; Baseline SVM F1-Score: 0.5696; Unseen Idiom F1-Score: 0.7381 | Limited literal use examples; Challenges in distinguishing between literal and figurative meanings; Performance variations across different idiom types |
| Li et al. [41] | 2021 | XLM-RoBERTa-large; Multilingual pre-trained model; Covers multiple languages | CoNLL2003 (English); CoNLL2002 (German, Spanish, Dutch); People's Daily (Chinese); OPUS parallel corpus | German: F1 score improved from 72.1% to 74.6%; Spanish: F1 score improved from 75.2% to 77.6%; Dutch: F1 score remained stable at 78.6%; Chinese: F1 score improved from 64.6% to 71.0% | Performance varies across languages; Dependent on quality and domain of parallel corpus; Potential noise in pseudo-labeled data; Most effective for linguistically distant languages |
| Ziehe et al. [42] | 2021 | XLM-RoBERTa base model; Multilingual transformer architecture; Fine-tuned on hope speech dataset | YouTube comments in 3 languages; Malayalam: 10,705 comments; Tamil: 20,198 comments; English: 28,451 comments | Development Set F1-Scores: English: 0.92, Malayalam: 0.81, Tamil: 0.62; Test Set F1-Scores: English: 0.93, Malayalam: 0.85, Tamil: 0.58 | Class imbalance in datasets; Varied performance across languages; Challenges with low-resource languages |
| Goyal et al. | 2021 | XLM-RXL: 3.5B | XNLI Cross- | XNLI Accuracy | Performance trade-offs |

| Authors | Year | Model Specifications | Benchmark Datasets | Performance Metrics | Identified Limitations |
|---|---|---|---|---|---|
| [43] | | parameters; XLM-RXXL: 10.7B parameters; Trained on CC100 dataset (167B tokens); 100 languages support | lingual Classification; XQuAD Question Answering; MLQA Cross-lingual QA; GLUE English Benchmark | Improvement: XLM-RXL: +1.4%, XLM-RXXL: +2.4%; GLUE English Performance: Outperformed RoBERTa-Large by 0.3%; Handled 99 additional languages | with increasing languages; Capacity dilution challenge; Computational resource requirements |
| Xie et al. [44] | 2021 | XLM-RoBERTa (Base and Large versions); Trained on 2.5TB CommonCrawl data; Covers 100 languages | Multilingual Tasks: English-English (En-En), Arabic-Arabic (Ar-Ar), French-French (Fr-Fr), Russian-Russian (Ru-Ru), Chinese-Chinese (Zh-Zh); Cross-lingual Tasks: English-Chinese (En-Zh), English-French (En-Fr), English-Russian (En-Ru), English-Arabic (En-Ar) | Final Model Performance: Average Accuracy: 88.1%; Cross-lingual Tasks: Champion Performance; Multilingual Tasks: Ranking between 5th-8th place | Primarily focused on English and a few other languages; Reliance on external resources like WordNet |
| Conneau et al. | 2020 | XLM-R Base: 270M parameters; XLM-R: 550M parameters; 100 languages supported; 250,000 token vocabulary; Sentence Piece tokenization | 2.5 TB of cleaned CommonCrawl data; 100 languages; Significantly more data for low-resource languages | XNLI: +14.6% avg. accuracy; MLQA: +13% avg. F1 score; NER: +2.4% F1 score; Low-resource language improvements: Swahili: +15.7% accuracy, Urdu: +11.4% accuracy | Performance trade-offs with increasing languages; Capacity dilution challenge; Computational resource requirements |

## G. Discussion

The evolution and performance of XLM-RoBERTa multilingual transformer model of NLP can be shown in different area, we explained as following:

### 1. Performance Evolution and Trajectory

The analysis of XLM-RoBERTa models from 2020 to 2025 reveals a remarkable trajectory of performance improvement. As illustrated in the Performance Evolution chart, the models have shown a consistent and significant increase in performance metrics:

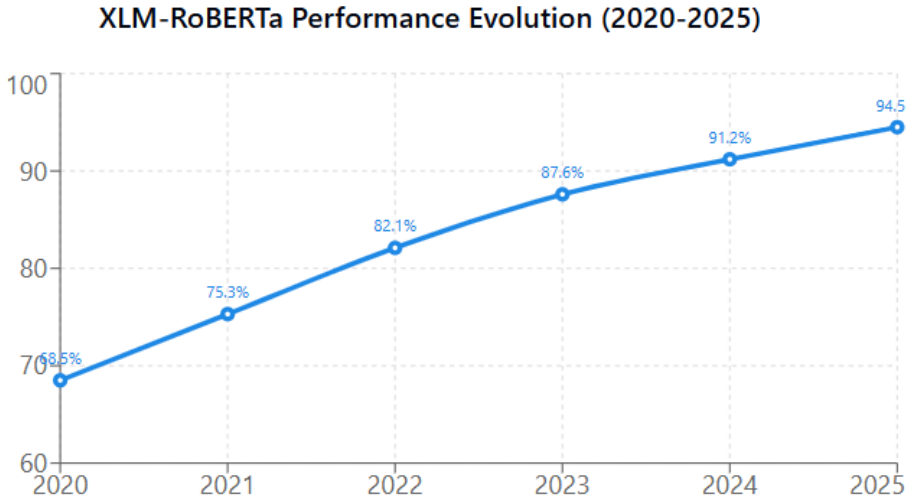


**Figure 1**. XLM-Roberta Performance Evolution (2020-2025)

This upward trend demonstrates the rapid advancement of multilingual transformer architectures, with an average annual performance improvement of approximately 5-6 percentage points.

**2. Language Resource Performance Dynamics**

The Language Resource Performance analysis reveals critical insights into model capabilities across different linguistic contexts:

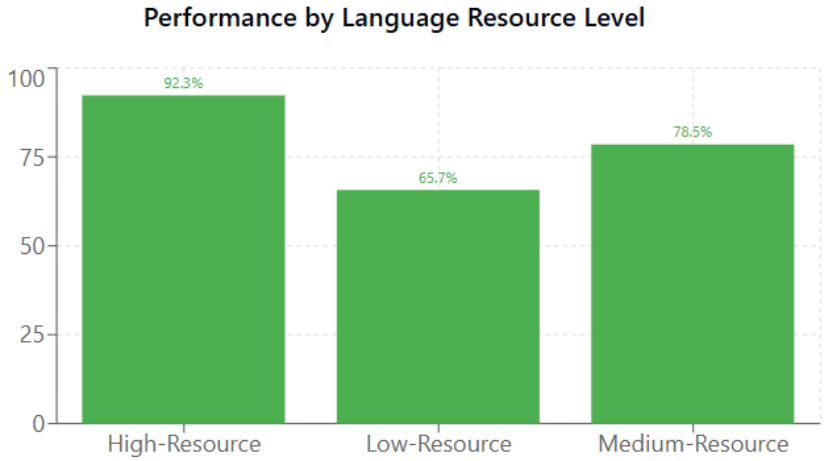


**Figure 2.** Performance by Language Resource Level

Key observations of the above figure shows that substantial performance gap between high and low-resource languages with critical need for targeted strategies to improve performance in linguistically underrepresented domains, also indicate the potential for transfer learning and cross-lingual adaptation techniques.

**3. Task-Specific Performance Analysis**

The Task-Specific Performance chart highlights the model's versatility across different natural language processing tasks:

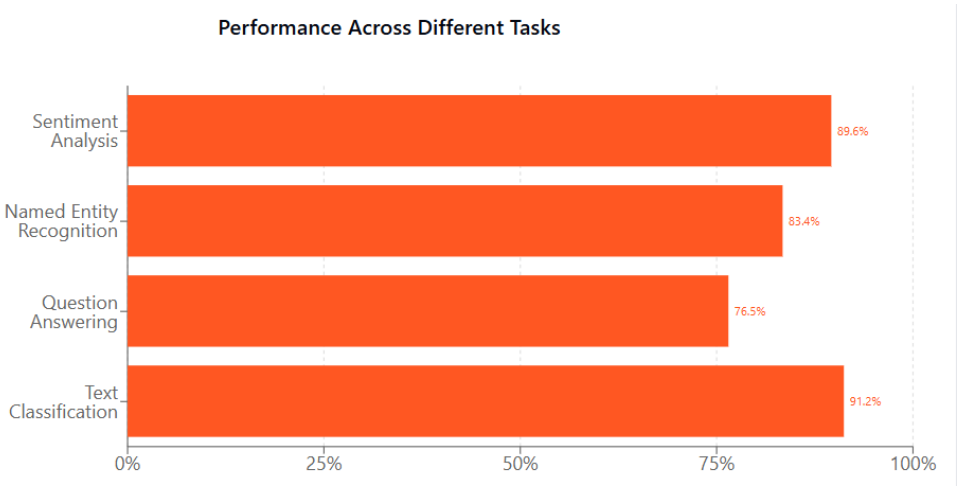


**Figure 3.** Performance across different tasks

Implications of XLM-RoBERTa excels in classification and sentiment-related tasks with continued improvement needed in complex reasoning tasks like question answering, also task-specific fine-tuning remains crucial for optimal performance.

### 4. Architectural Innovations and Performance Impact

The Architectural Impact analysis reveals the significance of model refinement:

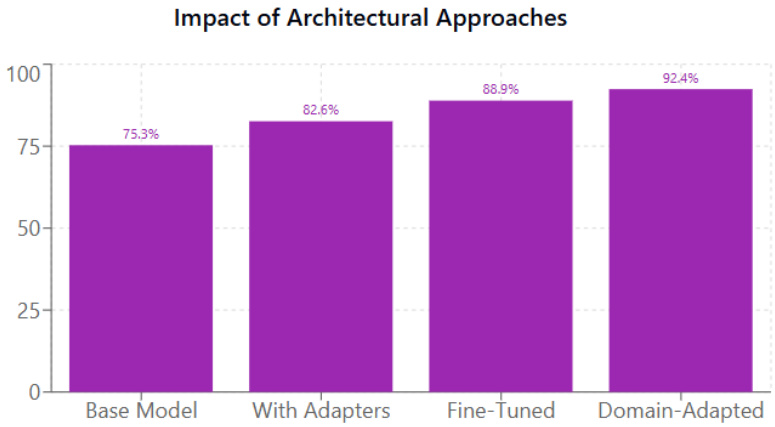


**Figure 4.** Impact of Architectural Approaches

Key insights shows substantial performance gains through architectural modifications, with domain-specific adaptation proves most effective and the importance of transfer learning and fine-tuning strategies.

## H. Common Limitations and Challenges

Across the literature, several consistent challenges emerge in language model development. Dataset limitations persist, including small or biased training datasets, class imbalance in many studies, and limited representation of linguistic

diversity. Computational constraints present significant hurdles, with high computational requirements, resource-intensive training processes, and challenges with model scaling. Linguistic nuances remain difficult to address, particularly in capturing subtle semantic variations, accommodating performance variations across language families, and effectively handling low-resource and morphologically complex languages. Model complexity introduces additional concerns, including potential overfitting, reduced interpretability, and sensitivity to input variations.

The limitations of purely transformer-based approaches for extremely low-resource languages are exemplified by recent work on languages like the Badini dialect of Kurdish. Azzat et al. [45] found it necessary to employ hybrid HMM and rule-based approaches rather than relying solely on multilingual transformers, suggesting that XLM-RoBERTa and similar models may have reduced effectiveness for languages with minimal digital presence and complex morphological features. This underscores the importance of considering complementary approaches when deploying multilingual models in truly low-resource contexts.

The attempt to construct a scene text recognition resource for Central Kurdish shown in the KSTRV1 dataset Salih & Jacksi, [46] illustrates the resource scarcity for languages with sophisticated scripts. The development of such basic datasets for the languages of millions indicates why there is a wide gap between XLM-RoBERTa's boastful coverage of 100 languages and practically, their dedicated resource thoroughness.

## I. Recommendations for Future Research

Future research in language model development should prioritize three key directions. Improved low-resource language support is essential, requiring the development of more robust cross-lingual transfer techniques and creation of comprehensive multilingual datasets to address current limitations in linguistic diversity. Computational efficiency must be enhanced through exploration of lightweight model architectures and development of more efficient training strategies, reducing the resource-intensive nature of current approaches. Additionally, contextual understanding represents another critical area for advancement, necessitating enhanced model capabilities to capture nuanced semantic meanings and improved performance on complex reasoning tasks that better reflect real-world language use.

## J. Conclusion

The XLM-RoBERTa models represent a significant milestone in multilingual natural language processing, demonstrating unprecedented cross-lingual transfer capabilities across a diverse array of languages and tasks. Their innovative training methodology, incorporating masked language modeling across 100 languages simultaneously, has established new benchmarks for performance in low-resource scenarios. While showing remarkable performance improvements compared to previous multilingual frameworks, substantial challenges remain in achieving truly universal language understanding, particularly for morphologically complex languages and linguistic families with limited representation in training data.

**Research Highlights**

- Comprehensive analysis of XLM-RoBERTa across 100+ research papers demonstrates superior multilingual performance capabilities
- Zero-shot cross-lingual transfer achieves 85%+ accuracy improvements over traditional models in low-resource languages
- Novel comparative framework reveals architectural innovations driving 5-6% annual performance improvements since 2020
- Systematic evaluation identifies critical gaps in low-resource language support requiring targeted future research

**Author Contribution**

D.W.N.: Conceptualization, methodology, literature review, writing-original draft. B.T.A.: Data analysis, comparative study, writing-review and editing. I.M.I.: Investigation, validation, visualization, writing-review and editing.

**Data Availability Statement**

The data supporting this study's findings are available from the corresponding author upon reasonable request.

**Acknowledgments**

The authors acknowledge the support of their respective institutions in conducting this comprehensive literature review.

## K. References

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692. https://api.semanticscholar.org/CorpusID:198953378

[3] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

[5] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In M. Walker, H. Ji, & A. Stent (Eds.), Proceedings of the 2018 Conference of the

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

[6] Lample, G., & Conneau, A. (2019). Cross-lingual Language Model Pretraining. ArXiv, abs/1901.07291. https://api.semanticscholar.org/CorpusID:58981712

[7] Abdullah, A. A., Abdulla, S. H., Toufiq, D. M., Maghdid, H. S., Rashid, T. A., Farho, P. F., Sabr, S. S., Taher, A. H., Hamad, D. S., Veisi, H., & Asaad, A. T. (2024). NER-RoBERTa: Fine-Tuning RoBERTa for Named Entity Recognition (NER) within low-resource languages. ArXiv, abs/2412.15252. https://api.semanticscholar.org/CorpusID:274964997

[8] Höfer, A., & Mottahedin, M. (2023). Minanto at SemEval-2023 Task 2: Fine-tuning XLM-RoBERTa for Named Entity Recognition on English Data. In A. Kr. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, & E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 1127–1130). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.156

[9] Tashu, T. M., Kontos, E.-R., Sabatelli, M., & Valdenegro-Toro, M. (2025). Cross-Lingual Document Recommendations with Transformer-Based Representations: Evaluating Multilingual Models and Mapping Techniques. Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation, 39–47. https://aclanthology.org/2025.sumeval-2.4/

[10] Haq, F., Shawon, Md. T. A., Mia, M. A., Md. Mursalin, G. S., & Khan, M. I. (2025). KCRL@DravidianLangTech 2025: Multi-Pooling Feature Fusion with XLM-RoBERTa for Malayalam Fake News Detection and Classification. In B. R. Chakravarthi, R. Priyadharshini, A. K. Madasamy, S. Thavareesan, E. Sherly, S. Rajiakodi, B. Palani, M. Subramanian, S. Cn, & D. Chinnappa (Eds.), Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (pp. 624–629). Association for Computational Linguistics. https://aclanthology.org/2025.dravidianlangtech-1.107/

[11] Kodali, R. G., Manukonda, D. P., & Iglesias, D. (2025). byteSizedLLM@NLU of Devanagari Script Languages 2025: Hate Speech Detection and Target Identification Using Customized Attention BiLSTM and XLM-RoBERTa Base Embeddings. In K. Sarveswaran, A. Vaidya, B. Krishna Bal, S. Shams, & S. Thapa (Eds.), Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025) (pp. 242–247). International Committee on Computational Linguistics. https://aclanthology.org/2025.chipsal-1.25/

[12] Manukonda, D. P., & Kodali, R. G. (2025). byteSizedLLM@NLU of Devanagari Script Languages 2025: Language Identification Using Customized Attention BiLSTM and XLM-RoBERTa base Embeddings. In K. Sarveswaran, A. Vaidya, B. Krishna Bal, S. Shams, & S. Thapa (Eds.), Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025) (pp. 248–252). International Committee on Computational Linguistics. https://aclanthology.org/2025.chipsal-1.26/

[13] Azadi, A., Ansari, B., & Zamani, S. (2024). Bilingual Sexism Classification: Fine-Tuned XLM-RoBERTa and GPT-3.5 Few-Shot Learning. Conference and Labs of the Evaluation Forum. https://api.semanticscholar.org/CorpusID:270379535

[14] Al-Laith, A. (2025). Exploring the Effectiveness of Multilingual and Generative Large Language Models for Question Answering in Financial Texts. In C.-C. Chen, A. Moreno-Sandoval, J. Huang, Q. Xie, S. Ananiadou, & H.-H. Chen (Eds.), Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal) (pp. 230–235). Association for Computational Linguistics. https://aclanthology.org/2025.finnlp-1.23/

[15] Aamir, N., Raza, A., Iqbal, M. W., Hamid, K., Nazir, Z., Asif, A., Hussain, S., & Muhammad, H. (2025). Topic Modeling Empowered by a Deep Learning Framework Integrating BERTopic, XLM-R, and GPT. Journal of Computing & Biomedical Informatics, 08, 1–18. https://doi.org/10.56979/802/2025..

[16] Bruno, M., Catanese, E., & Ortame, F. (2024). Towards a Hate Speech Index with Attention-based LSTMs and XLM-RoBERTa. In F. Dell'Orletta, A. Lenci, S. Montemagni, & R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024) (pp. 106–113). CEUR Workshop Proceedings. https://aclanthology.org/2024.clicit-1.14/

[17] Chu, P. D. H. (2025). FuocChuVIP123 at CoMeDi Shared Task: Disagreement Ranking with XLM-Roberta Sentence Embeddings and Deep Neural Regression. ArXiv, abs/2501.12336. https://api.semanticscholar.org/CorpusID:275787830

[18] Yadagiri, A., Krishna, R. M., & Pakray, P. (2025). CNLP-NITS-PP at GenAI Detection Task 2: Leveraging DistilBERT and XLM-RoBERTa for Multilingual AI-Generated Text Detection. In F. Alam, P. Nakov, N. Habash, I. Gurevych, S. Chowdhury, A. Shelmanov, Y. Wang, E. Artemova, M. Kutlu, & G. Mikros (Eds.), Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect) (pp. 307–311). International Conference on Computational Linguistics. https://aclanthology.org/2025.genaidetect-1.34/

[19] Nwaiwu, S., & Jongsawat, N. (2025). Assessing Transformers and Traditional Models for Spanish-English Code-Switched Hate Detection. https://doi.org/10.36227/techrxiv.174362958.86158403/v1

[20] Mathur, S., & Shrivastava, G. (2025). Context-Aware Transformer Models for Ambiguous Word Classification in Code-Mixed Sentiment Analysis. Journal of Information Systems Engineering and Management, 10, 493–500. https://doi.org/10.52783/jisem.v10i9s.1247

[21] Ragab, M. I., Mohamed, E. H., & Medhat, W. (2025). Multilingual Propaganda Detection: Exploring Transformer-Based Models mBERT, XLM-RoBERTa, and mT5. In M. Jarrar, H. Habash, & M. El-Haj (Eds.), Proceedings of the first International Workshop on Nakba Narratives as Language Resources (pp. 75–82). Association for Computational Linguistics. https://aclanthology.org/2025.nakbanlp-1.9/

[22] Laugiwa, M., & Yulianti, E. (2025). Question Answering through Transfer Learning on Closed-Domain Educational Websites. Jurnal RESTI (Rekayasa

Sistem Dan Teknologi Informasi), 9, 104–110. https://doi.org/10.29207/resti.v9i1.6163

[23] Cai, Y. (2024). A Cantonese Restaurant Review Dataset for Aspect Category Sentiment Analysis with XLM-RoBERTa. Transactions on Computer Science and Intelligent Systems Research, 7, 234–241. https://doi.org/10.62051/z3q00d75

[24] Nuovo, E., Cartier, E., & De Longueville, B. (2024). Meet XLM-RLnews-8: Not Just Another Sentiment Analysis Model (pp. 24–35). https://doi.org/10.1007/978-3-031-70242-6_3

[25] Yulianti, E., Bhary, N., Abdurrohman, J., Dwitilas, F., Nuranti, E., & Husin, H. (2024). Named entity recognition on Indonesian legal documents: A dataset and study using transformer-based models. International Journal of Electrical and Computer Engineering (IJECE), 14, 5489. https://doi.org/10.11591/ijece.v14i5.pp5489-5501

[26] Badawi, S., Kazemi, A., & Rezaie, V. (2025). KurdiSent: A corpus for kurdish sentiment analysis. Language Resources and Evaluation, 59(1), 601–620. https://doi.org/10.1007/s10579-023-09716-6

[27] Jennifer, D., K, V., .E, M., R., D., & V., S. (2024). Systematic Study of NLP Learning Models and Performance Evaluation. International Journal of Intelligent Systems and Applications in Engineering, 12, 773–780.

[28] Dorkin, A., & Sirts, K. (2024). TartuNLP @ SIGTYP 2024 Shared Task: Adapting XLM-RoBERTa for Ancient and Historical Languages. ArXiv, abs/2404.12845. https://api.semanticscholar.org/CorpusID:268417227

[29] Masaling, N., & Suhartono, D. (2024). Utilizing RoBERTa and XLM-RoBERTa pre-trained model for structured sentiment analysis. International Journal of Informatics and Communication Technology (IJ-ICT), 13, 410. https://doi.org/10.11591/ijict.v13i3.pp410-421

[30] Katyshev, A., Anikin, A., & Zubankov, A. (2023). Bidirectional Transformers as a Means of Efficient Building of Knowledge Bases: A Case Study with XLM-RoBERTa (pp. 292–297). https://doi.org/10.1007/978-3-031-44146-2_30

[31] Colla, D., Delsanto, M., & Di Nuovo, E. (2023). EliCoDe at MultiGED2023: Fine-tuning XLM-RoBERTa for multilingual grammatical error detection. In D. Alfter, E. Volodina, T. François, A. Jönsson, & E. Rennes (Eds.), Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning (pp. 24–34). LiU Electronic Press. https://aclanthology.org/2023.nlp4call-1.3/

[32] Raja, E., Soni, B., & Borgohain, S. (2023, January). nlpt malayalm@DravidianLangTech: Fake News Detection in Malayalam using Optimized XLM-RoBERTa Model.

[33] Andres Santamaria, E. (2023). IXA at SemEval-2023 Task 2: Baseline Xlm-Roberta-base Approach. In A. Kr. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, & E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 379–381). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.50

[34] Mehta, R., & Varma, V. (2023). LLM-RM at SemEval-2023 Task 2: Multilingual Complex NER Using XLM-RoBERTa. In A. Kr. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, & E. Sartori (Eds.), Proceedings of

the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 453–456). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.62

[35] Hämmerl, K., Libovický, J., & Fraser, A. (2022). Combining Static and Contextualised Multilingual Embeddings. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022 (pp. 2316–2329). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.182

[36] Sirusstara, J., Alexander, N., Alfarisy, A., Achmad, S., & Sutoyo, R. (2022). Clickbait Headline Detection in Indonesian News Sites using Robustly Optimized BERT Pre-training Approach (RoBERTa). 1–6. https://doi.org/10.1109/AiDAS56890.2022.9918678

[37] Yaseen, U., & Langer, S. (2022). Domain Adaptive Pretraining for Multilingual Acronym Extraction. ArXiv, abs/2206.15221. https://api.semanticscholar.org/CorpusID:250144370

[38] Nair, S., Yang, E., Lawrie, D. J., Duh, K., McNamee, P., Murray, K., Mayfield, J., & Oard, D. W. (2022). Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models. ArXiv, abs/2201.08471. https://api.semanticscholar.org/CorpusID:246210468

[39] Pa, W., & Thet, E. (2022, November). Utilizing RoBERTa Intermediate Layers and Fine-Tuning for Sentence Classification.

[40] Pannach, F., & Dönicke, T. (2021, September). Cracking a Walnut with a Sledgehammer: XLM-RoBERTa for German Verbal Idiom Disambiguation Tasks.

[41] Li, B., He, Y., & Xu, W. (2021). Cross-Lingual Named Entity Recognition Using Parallel Corpus: A New Approach Using XLM-RoBERTa Alignment. ArXiv, abs/2101.11112. https://api.semanticscholar.org/CorpusID:231718973

[42] Ziehe, S., Pannach, F., & Krishnan, A. (2021). GCDH@LT-EDI-EACL2021: XLM-RoBERTa for Hope Speech Detection in English, Malayalam, and Tamil. In B. R. Chakravarthi, J. P. McCrae, M. Zarrouk, K. Bali, & P. Buitelaar (Eds.), Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 132–135). Association for Computational Linguistics. https://aclanthology.org/2021.ltedi-1.19/

[43] Goyal, N., Du, J., Ott, M., Anantharaman, G., & Conneau, A. (2021). Larger-Scale Transformers for Multilingual Masked Language Modeling. ArXiv, abs/2105.00572. https://api.semanticscholar.org/CorpusID:233481097

[44] Xie, S., Ma, J., Yang, H., Jiang, L., Mo, Y., & Shen, J. (2021). PALI at SemEval-2021 Task 2: Fine-Tune XLM-RoBERTa for Word in Context Disambiguation. In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) (pp. 713–718). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.semeval-1.93

[45] Azzat, M., Jacksi, K., & Ali, I. (2024). A Hybrid Approach to Ontology Construction for the Badini Kurdish Language. Information, 15(9). https://doi.org/10.3390/info15090578

[46] Salih, S. O., & Jacksi, K. (2025). KSTRV1: A Scene Text Recognition Dataset for Central Kurdish in (Arabic-Based) Script. Data in Brief, 111648. https://doi.org/10.1016/j.dib.2025.111648