Cross-Lingual Approaches for Text Difficulty Classification in Non-English Languages

Huili Dai
Zhengzhou Science and Technology University, Zhengzhou 450064, China
E-mail: daihuili603@126.com

Keywords: Difficulty classification, natural language processing, cross-lingual models, machine translation, machine learning

Received: April 20, 2025

Text Difficulty Classification (TDC) significantly contributes to educational technology, language learning, and information access by automatically predicting the difficulty of texts. While English has been favored with extensive resources and mature readability tools, non-English languages suffer from sparse annotated corpora, linguistic diversity, and insufficient computational resources. Cross-lingual methods, including shared knowledge transfer from high-resource to low-resource languages, have become increasingly popular in recent years. It covers the latest advances in cross-lingual TDC in pipelines of machine translation-based, multi-lingual pre-trained transformers (e.g., mBERT, XLM-R), adversarial and meta-learning-based approaches, and knowledge distillation techniques. It also summarizes available datasets and benchmarks, critiques current practices, and highlights morphological complexity, translation artifacts, and domain mismatch. It identifies significant trends and future promise by critically examining state-of-the-art methods for multimodal TDC, hybrid architectures, and active learning in low-resource languages. The results highlight the significance of non-discriminatory, resource-lean systems for measuring fair-minded text difficulty in many languages.

Povzetek: Pregledni članek povzema navzkrižno-jezikovne pristope k razvrščanju težavnosti besedil, zbere nabore podatkov in kritično izpostavi ter trende k večmodalnim, hibridnim in aktivnim metodam v jezikovno skromnih okoljih.

1 Introduction

Text Difficulty Classification (TDC) is the basis of next-generation learning technology, language learning software, and Natural Language Processing (NLP) solutions [1]. Accurate identification of textual complexity enables adaptive learning platforms to recommend material commensurate with the student's proficiency, ensuring interest and comprehension [2]. TDC facilitates teachers and learners in language pedagogy by simplifying material choice and automatically scoring exercises [3]. Beyond pedagogy, TDC enables access by adapting information for diverse users, such as non-native language users or those with reading impairments. The explosive nature of NLP-based solutions, such as summarizing and suggesting content, has expanded the requirement for resilient TDC solutions [4]. As technology continues to serve as a middleman for knowledge, automating the determination of reading levels has become crucial to delivering fair and efficient learning experiences worldwide [5].

Despite these advances, most advances in TDC have centered on English, for which vast annotated corpora, readability algorithms, and pre-trained models have spurred the development of practical systems. Non-English languages are underrepresented by sparse labeled datasets, linguistic variability, and limited computational resources [6]. Linguistically elaborate languages, such as Finnish, agglutinative languages, like Turkish, and languages with non-Latin scripts, are especially problematic for feature extraction and modeling. Cultural and educational variations similarly affect text richness, making it challenging to use English-centric tools directly [7]. These resource disparities have resulted in unequal access to advanced readability technologies, denying many language populations sufficient tools to support adaptive learning or content personalization. Overcoming these gaps is critical to bringing TDC's benefits to a broader, multi-lingual population.

Cross-lingual techniques offer a solution to fill these resource gaps. They can construct TDC systems with limited labeled data by leveraging knowledge transfer from high-resource languages like English to under-resourced languages [8]. Machine translation-based pipelines, transformer-based multi-lingual models (e.g., XLM-R, mBERT), and adversarial learning allow English-trained models to generalize effectively to unseen languages. For example, multi-lingual embeddings project universal semantic structures into common spaces for encoding, while meta-learning algorithms rapidly learn to adapt to new languages under sparse supervision [9]. In addition to reducing the use of large sets of annotated data, they enable potential lifts to TDC for underrepresented languages and more comprehensive, scalable solutions.

The present study encompasses cross-lingual TDC techniques from both computational and linguistic perspectives. From a computational point of view, we critique algorithmic techniques such as zero-shot transfer, multi-lingual pre-training, and knowledge distillation. Theoretically, we examine language structure differences, morphology, syntax, and orthography concerning system design and efficacy. Cultural background and educational structure are also addressed in terms of their impacts on measuring text difficulty. By synthesizing findings from numerous recent research studies, the study identifies best practices, highlights long-standing challenges, and suggests ways to overcome existing obstacles. The dual computational and linguistic approach ensures a subject-wide understanding of the field and its potential to influence global education and NLP applications.

The rest of the paper is organized as follows. Section 2 provides a background on TDC and its evolution from early readability formulae to state-of-the-art neural techniques. Section 3 overviews cross-lingual methods, including translation-based pipelines, multi-lingual pre-trained models, adversarial, and alignment-based techniques. Section 4 presents an overview of state-of-the-art datasets, benchmarks, and evaluation protocols, and discusses notable challenges such as morphologically challenging texts and translation-based artifacts. Section 5 summarizes the main findings and proposes resource-aware and inclusive TDC solutions.

2 Background and foundations

It is desirable to understand early TDC development to contextualize the development of cross-lingual techniques. Early TDC research relied upon manually crafted readability formulae and superficial linguistic characteristics to gain quick, superficial insights into reading ease. As computational linguistics and NLP advanced, techniques have become increasingly underpinned by machine learning and, more recently, transformer-based architectures to expose high-level syntactic and semantic structures. This section provides an overview of the salient concepts, outdated techniques, and technological milestones that underpin today's TDC systems, laying the groundwork for descriptions of cross-lingual methodologies.

2.1 Definition of text difficulty classification

TDC is a basic NLP task that automatically determines the complexity level suitable for a particular audience or purpose [10]. Although typical formulae for readability focus primarily on superficial aspects, such as sentence length or word count, the latest TDC methods utilize syntactic, semantic, and discourse-level features to provide a more fine-grained analysis. Advanced Transformer language models, combined with multi-lingual embeddings and feature engineering techniques, now enable TDC systems to account for the subtleties of prose difficulty in different languages and contexts.

Table 1: Applications of text difficulty classification

Application area	Description	Example use cases
Adaptive learning	Matches reading materials to a learner's proficiency to optimize engagement and comprehension.	Personalized e-learning platforms, digital textbooks.
Second-language learning	Aligns texts to frameworks like CEFR and supports graded reading exercises.	Language learning apps, automated test generation.
Text simplification	Identifies and modifies complex passages for easier reading.	Simplified news articles, accessibility for dyslexia.
Content accessibility	Adapts information for non-native speakers or audiences with varied literacy levels.	Healthcare instructions, government resources.
Information retrieval	Enhances search and recommendation systems by filtering or ranking texts based on difficulty.	News curation, personalized search engines.
Cross-lingual education	Aligns text difficulty across languages for bilingual or multi-lingual contexts.	Comparative linguistics and bilingual education tools.

As indicated in Table 1, TDC has broad applications in education, access, and information dissemination. Adaptive learning platforms ensure that the learners are given sufficiently challenging material. TDC aligns texts to schemes, such as the CEFR of second language learning, and allows automatic scoring. In sentence condensation for simplified texts, sentence structure must be simplified to facilitate reading for children, non-native readers, or learners with learning disabilities. TDC is also used to adapt news, healthcare, and legal texts to make them easily readable by various user groups. Finally, for cross-lingual cases, TDC enables comparative analysis of languages and matching of bilingual resources.

2.2 Traditional approaches

Traditional techniques of TDC primarily concern formulae of readability based on features. These evaluate superficial characteristics such as mean sentence length, word length, and number of syllables to estimate the hardness of documents. User-friendly measures, such as Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, and SMOG, are presented in Table 2. Because they require minimal computation and are easy to use, they have become cornerstones for teaching and publishing. Although they strongly correlate to simple English reading difficulty and produce quick, interpretable output, they neglect advanced language features such as syntax, semantics, and discourse structure. Thus, they tend to overlook subtle changes in difficulty, especially in texts that contain challenging vocabulary, specialized terms, and non-standard structures.

Feature-based formulae for readability are significantly constrained when applied to non-English languages [11]. Formulas can be specifically adjusted to morpho-syntactic English characteristics and thus may be

Table 2: Common feature-based readability formulas and their limitations

Formula	Key features used	Strengths	Limitations for non-English texts
Flesch reading ease	Sentence length and syllable count	Simple, interpretable, and widely adopted	Ineffective for agglutinative or logographic languages; ignores syntax/semantics.
Flesch-Kincaid grade	Sentence length and syllable count	Maps scores to U.S. grade levels	English-specific calibration; unsuitable for tonal or morphologically rich languages.
Gunning fog index	Sentence length and complex word frequency	Identifies overly complex sentences	Overestimates difficulty in inflection-heavy languages.
SMOG index	Polysyllabic word frequency and sentence count	Useful for shorter texts	Relies on syllable-based measures, not applicable to all languages.
Coleman-Liau index	Character and word counts	Avoids syllable counting and easy to compute	Sensitive to alphabet differences; fails for non-alphabetic scripts.

less reliable in languages with different structures. For example, agglutinative languages like Turkish or Finnish pack strong grammaticality into bound words; hence, word length per se or word-based syllable count are unreliable measures of difficulty. Similarly, tonal languages (e.g., Mandarin Chinese) or logographic script languages do not enjoy direct syllable-oriented metric analogs. Moreover, the cultural and educational background in which difficulty is intuitive implies that English-based thresholds can conflict with the complexities of texts in other languages. Due to these challenges, various solutions or language-specific variations of standard formulae are necessary to estimate accurately.

The weaknesses of standard formulae have motivated the development of higher-level computational TDC models. More advanced methods simulate linguistic features, such as part-of-speech tendencies, syntactic tree depth, and semantic similarity, to gain a more fine-grained understanding of text difficulty. Cross-lingual and multi-lingual methods, in particular, aim to overcome English-biased formula structure biases by leveraging common embeddings or translation-based pipelines. While traditional metrics are fine for primary analysis or comparison to baselines, they fall short of the current diversified linguistic climate. Identifying these weaknesses has inspired research into language-agnostic, resilient, and context-sensitive methods for classifying text difficulty.

2.3 Advances in NLP and multi-lingual models

Recent advances in NLP revolutionized TDC through the advent of pre-trained transformer architectures. Pretrained transformer-based models such as BERT, RoBERTa, and their multi-lingual counterparts (mBERT, XLM-RoBERTa), listed in Table 3, exploit self-attention mechanisms and large-scale unsupervised learning to learn rich contextual representations of texts. Those models encode superficial features and syntax-based, semantic, and discourse-level structures that classical formulas cannot address [12]. Their ability to represent words and sentences in high-dimensional word spaces enables more accurate estimations of text difficulty. For cross-lingual analysis, multi-lingual transformer-based models trained concurrently in dozens of languages create common embedded spaces, transferring knowledge from high-resource to low-resource languages. That has opened novel opportunities for zero-shot and few-shot TDC without requiring large label sets for every target language.

Multi-lingual word and sentence embeddings, such as MUSE, LASER, and LaBSE, can potentially bridge linguistic gaps [13]. These embeddings place semantically similar words/phrases from varying languages into the same vector space, allowing for a comparative study of text difficulty across languages. English-based TDC, for instance, can be employed to classify texts from German or Spanish by relying on common representations of embeddings. Machine translation pipelines, in contrast, can inject translation artifacts. However, multi-lingual embeddings preserve language-specific information while maintaining language comparability. Their application is particularly beneficial in low-resource settings where annotated datasets are unavailable. By utilizing these embeddings, scholars can conceptualize language-agnostic large-scale TDC systems that can be applied to different linguistic populations.

Combined pre-trained transformers and multi-lingual embeddings have significantly enhanced the efficiency and transferability of TDC systems [14]. Backward-compatible methods, such as knowledge distillation and

Table 3: Key multi-lingual models and embedding techniques for cross-lingual TDC

Model/embedding	Type	Languages supported	Key strengths	Example use cases
mBERT	Transformer	100+	Shared embeddings and zero-shot transfer	Cross-lingual TDC, classification tasks
XLM-RoBERTa	Transformer	100+	High performance across low-resource languages	TDC for underrepresented languages
LASER	Sentence embedding	90+	Language-agnostic sentence vectors	Multi-lingual text similarity, TDC
LaBSE	Sentence embedding	100+	Strong alignment for sentence-level tasks	Cross-lingual semantic analysis and TDC
MUSE	Word embedding	30+	Aligns monolingual embeddings into a shared space	Low-resource TDC, bilingual lexicon induction

adapter layers, encourage model adaptability at lower computational costs without compromising precision. Fine-tuning multi-lingual transformers over relatively small labeled sets of the target language has achieved competitive results, even in low-resource environments. It has shifted the paradigm from language-specific, handcrafted features to universal, context-sensitive representations that enable global education, accessibility, and information seeking. As shown in Table 3, state-of-the-art multi-lingual models and embedding methods provide a robust foundation for cross-lingual text difficulty analysis.

3 Cross-lingual approaches

Cross-lingual techniques have emerged as the preferred method of translating TDC from English to resource-scarce languages [15]. Transferring knowledge from high-resource to low-resource languages that are poorly annotated balances the disparity in linguistic resources and enables scalable, multi-lingual solutions. Rather than learning new models for each language, cross-lingual techniques use common representations, machine translation, or alignment protocols to bridge linguistic divides.

Figure 1 provides an at-a-glance overview of the principal categories of cross-lingual techniques, translation-based pipelines, multi-lingual pre-trained models, adversarial and meta-learning techniques, and knowledge distillation, highlighting their relationships and complementary roles. This section addresses how these techniques support TDC in non-English languages.

3.1 Machine translation-based methods

Machine translation-based methods are among the earliest and lightest-weight techniques for non-English TDC support. The general workflow, shown next to its strengths and weaknesses in Table 4, is to translate a source language sentence into English using a machine translation system, classify the output English translation using a pre-trained English TDC model, and project the output back to the original language [16]. The resulting pipeline capitalizes on the maturity and availability of English readability tools and large sets of labeled English corpora. By bypassing the requirement of maintaining large annotated corpora for each target language, it can be employed for fast cross-lingual TDC support, especially when resources for one language are limited.

The advantages of machine translation-based methods include their generic nature and broad applicability. High-performance translation systems, such as Google Translate, DeepL, and MarianMT, can be utilized to translate hundreds of languages; therefore, TDC can be deployed for underrepresented languages without the need to develop new classifiers from scratch. It is also desirable for domains of high translation quality, such as formal or informational texts. Additionally, machine translation pipelines are typically the baseline method for evaluating more advanced cross-lingual techniques and constitute a real-world standard for model quality. For example, those evaluating TDC in German or Spanish can assess their multi-lingual transformer models against a translation-based baseline to estimate the value of cross-lingual pre-training.

Despite its advantages, this method has notable limitations. By adding noise, altering sentence structure, or replacing domain-specific words, translations can be deceptive in terms of the perceived complexity of the original. Cultural references, colloquialisms, and morphology may be underrepresented or overrepresented during translation, resulting in less accurate predictions. Translations are also subject to considerable variability in language pairs—high-resource pairs, such as English–Spanish, translate successfully, whereas, by and large, poor translations are obtained from English–Zulu, a low-resource pair. Translating significant texts is often computationally too costly even for non-realtime educational use cases at scale. These weaknesses

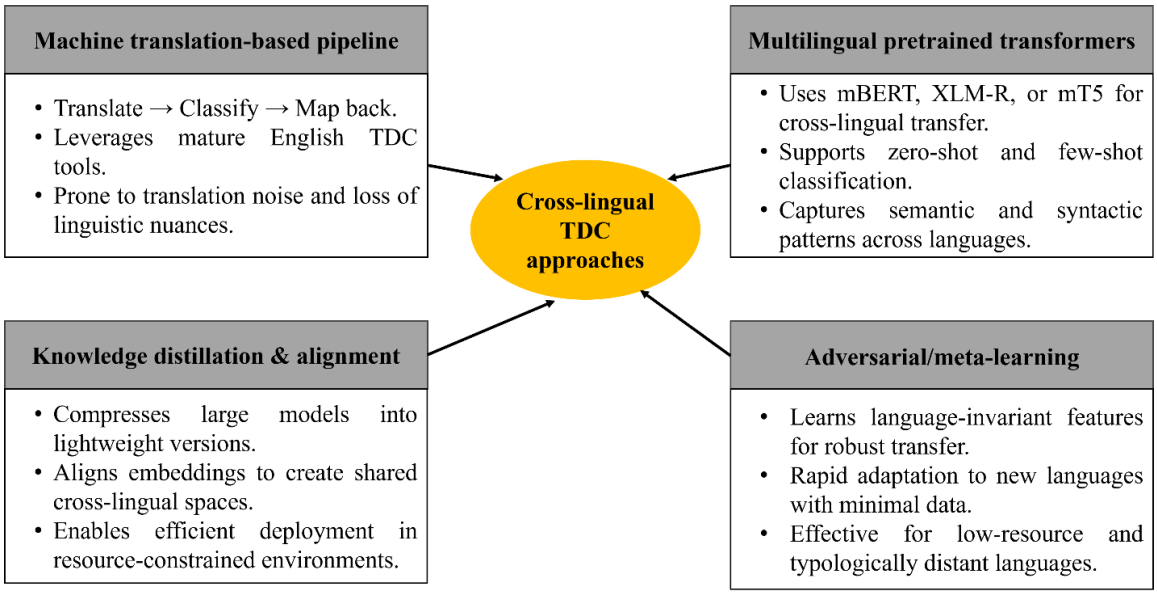

Figure 1: Overview diagram of cross-lingual TDC approaches

Table 4: Pros and cons of machine translation-based TDC

Aspect	Advantages	Limitations	Example use cases
Workflow	Simple pipeline; leverages existing English tools	Dependent on translation quality, it may distort the meaning	Quick deployment for new languages
Resource requirements	No need for target-language labeled data	Computationally expensive for large corpora	Baseline benchmarking for research
Performance	Effective for high-resource language pairs	Poor for low-resource pairs; loses cultural and linguistic nuances	Educational tools for well-supported languages

Table 5: Multi-lingual transformers for cross-lingual TDC

Model	Supported languages	Key strengths	Limitations	Example use cases
mBERT	100+	Zero-shot transfer; robust and widely adopted	Struggles with distant language pairs	Cross-lingual TDC, multi-lingual text classification
XLM-RoBERTa	100+	High performance for low-resource languages	Computationally expensive	Educational content recommendation, semantic similarity
mT5	100+	Supports classification and generation tasks	Resource-intensive training and fine-tuning	Text simplification, adaptive learning pipelines

necessitate the implementation of an alternative to translation-based methods, such as multi-lingual pre-trained models or adversarial learning.

3.2 Multi-lingual pretrained models

Pretrained multi-lingual models, such as mBERT, XLM-RoBERTa, and mT5, have transformed cross-lingual TDC by offering a single architecture for processing different languages. These models are trained on gigantic multi-lingual corpora and thus obtain shared representations that capture semantic and syntactic knowledge common across languages. As demonstrated in Table 5, their architectures can handle scripts and linguistic variations, allowing for knowledge transfer from high-resource languages (e.g., English) to less-resourced ones. The models do not use machine translation but represent texts from different languages within a common vector space, thereby reducing artifacts and preserving cultural or linguistic features.

The primary benefit of multilingual-pretrained models is that they enable zero- and few-shot learning. The TDC model, fine-tuned for English, can classify Spanish or German texts without additional training or with minimal labeled data. This is particularly desirable for languages with limited resources, for which it is not practical to build TDC-specific datasets. Usagees encompass adaptive learning platforms that suggest multi-lingual learning resources, cross-lingual search engines that sort texts by proficiency, and linguistic comparative research. Additionally, it is often possible to fine-tune these models for minor target language datasets, resulting in near-performances comparable to those of monolingual systems, highlighting their versatility and efficacy.

While influential, multi-lingual pre-training models come with many problems. Their computational cost is vast—training and fine-tuning require extensive hardware support, making it infeasible for small institutions. Their performance may be bad in typologically distinct or under-resourced languages for which little pre-training data exists or is of poor quality. Additionally, these models sometimes exhibit biases inherited from learning corpora, potentially warping the difficulty of prediction across cultural or linguistic domains. Complementary research may incorporate language-specific adapter addition, lifelong learning, or active learning methods to improve efficiency and fairness. Hybrid models that combine multi-lingual models with light-weight language-aware components may improve performance in diverse linguistic environments.

3.3 Adversarial and meta-learning techniques

Adversarial learning has emerged as an optimal approach to enhancing cross-lingual TDC by incorporating language-invariant features into models. Under this approach, the classifier predicts text difficulty, while the discriminator attempts to determine the language of the input [17]. The classifier is trained to deceive the discriminator, thereby causing it to focus on features related to complexity rather than language-related features. As Table 6 summarizes, the strategy reduces the application of superficial features, such as script or morphology, whose characteristics differ significantly from language to language. Adversarial learning is particularly valuable when labeled resources are in short supply, as it enables knowledge to be passed from high-resource to low-resource languages. It has emerged that adversarial-training models outperform regular fine-tuning methods in zero-shot and few-shot scenarios, proving an exemplary technology for multi-lingual learning platforms and adaptive learning systems.

Meta-learning techniques model TDC for each language as another similar task, allowing the model to learn. Model-Agnostic Meta-Learning (MAML) and related methods enable rapid adaptation to unseen languages, even with limited labeled data. For example, after being trained on many high-resource languages, the TDC model can be meta-learned to be updatable to a new low-resource language from only a few examples. That quick adaptability makes meta-learning particularly suitable for under-represented languages/dialects or variable teaching-learning scenarios, where new use domains emerge quickly. Meta-learning can be combined with adversarial techniques for better robustness and

Table 6: Adversarial and meta-learning for cross-lingual TDC

Approach	Key benefit	Limitation	Example use cases
Adversarial	Learns language-invariant representations	Requires careful tuning to avoid instability	Zero-shot TDC for multi-lingual education tools
Meta-learning	Rapidly adapts to new languages	Computationally expensive, task balancing needed	Few-shot TDC for low-resource languages
Hybrid methods	Combines robustness and adaptability	Complexity in implementation	Dynamic educational content recommendation

higher generalization across typologically diverse languages.

While adversarial and meta-learning techniques hold much promise, they are computationally costly and must be carefully tuned for hyperparameters to avoid instability or overfitting. Adversarial settings sometimes yield poor outcomes if the discriminator is overly strong or weak, and meta-learning necessitates balanced task sampling to achieve generalization. While they bring those issues, they also hold the potential to open the door to scaled TDC solutions that are friendly to limited resources. Future research can explore hybrid settings that integrate adversarial training with multi-lingual transformers or integrate curriculum learning to guide meta-learning processes. Light-weight adversarial settings or task-specific meta-learning fine-tuning can also render them possible for institutions with limited computational power.

3.4 Knowledge distillation and alignment

Embedding alignment is one of the basic techniques for achieving cross-lingual TDC without machine translation. It is comprised of casting monolingual word or sentence embeddings into a single semantic space such that texts from different languages can be analyzed comparatively directly. Programs like MUSE and VecMap learn linear transformations or higher-order embeddings between embedding spaces from bilingual dictionaries or parallel texts.

Once aligned, embeddings from low-resource languages can benefit from similarity relationships learned from high-resource languages, enabling classifiers from one language to generalize successfully to different languages. As Table 7 demonstrates, embedding alignment is significantly less computationally intensive than full retraining, while retaining language-specific nuances to a greater extent than translation-based pipelines. It is especially valuable for languages that lack large-scale, pre-trained multi-lingual models.

Distillation involves compressing large, high-performing multi-lingual models into smaller, light-weight models of comparable predictive quality. In distillation, a "teacher" model (e.g., XLM-RoBERTa) provides soft predictions or intermediate representations, and another "student" model is trained to emulate these. The student model can therefore replicate comparable performance at much lower computational cost. For TDC, distillation enables robust classifiers even on devices with limited computational power, such as mobile telephones or classroom tablets. It allows large-scale deployment in resource-scarce educational environments with limited computational infrastructure.

It is possible to augment distillation with embedding alignment to reduce cross-lingual quality and overhead.

Distilling knowledge and aligning embeddings offer significant value in terms of efficiency, scalability, and availability. Both enable us to use cross-lingual TDC systems in real-world tasks, from adaptive learning platforms to multi-lingual search engines. But they are also prone to pitfalls. The alignment quality mainly depends on the availability of appropriate bilingual dictionaries or parallel data, of which many under-resourced languages are scanty. Distilling is also at risk of discarding fine-grained linguistic information, especially when compressing large teacher models into light-weight student models. Future research can be conducted to investigate unsupervised alignment techniques, dynamic distillation mechanisms, and hybrid architectures that combine adapters or prompts with distilled models. It is through inventions of this nature that it can be ensured that resilient, economical cross-lingual TDC solutions are of high precision for a large number of languages.

4 Discussion

The preceding sections lay out the developments and challenges of cross-lingual TDC. The prior analysis combines knowledge from datasets and benchmarks, evaluation procedures, challenges and constraints, and the future, providing an evaluative summary of the current state of the field and its innovation potential.

Table 7: Knowledge distillation and alignment in cross-lingual TDC

Technique	Advantage	Limitation	Example applications
Embedding alignment	Enables direct comparison of texts across languages; light-weight	Requires bilingual dictionaries or parallel corpora	Low-resource TDC, multi-lingual similarity tasks
Knowledge distillation	Compresses large models for efficient deployment	Potential loss of nuanced knowledge	Mobile TDC apps, scalable educational platforms
Combined approaches	Balances accuracy and efficiency	Complexity in implementation	Cross-lingual adaptive learning tools

Table 8: Overview of datasets and benchmarking strategies for cross-lingual TDC

Category	Examples	Strengths	Limitations	Potential improvements
Non-English corpora	Klexikon (German) amd CEFR-Spanish	Tailored to specific languages and audiences	Limited size and narrow domain coverage	Expand to multiple genres and proficiency levels
Multi-lingual benchmarks	CLEAR, XGLUE, and MLQA	Enables fair cross-lingual comparison	Focuses on popular languages	Include underrepresented and low-resource languages
Data augmentation methods	Back-translation and synthetic data	Expands resources for low-resource languages	Risk of noise, cultural bias, or mislabeling	Combine augmentation with expert validation
Crowdsourced labeling	Amazon MTurk and Prolific	Scalable and cost-effective	Inconsistent quality and annotator expertise	Use tiered review systems and clear guidelines

4.1 Datasets and benchmarks

Access to reputable datasets and benchmarks is key to the progress of cross-lingual TDC. Without proper, diversified, and representative corpora, the testing and development of TDC systems can become biased toward a few high-resource languages and domains. Non-English datasets, such as Klexikon (German encyclopedia texts for children) and CEFR-aligned Spanish datasets, described in Table 8, serve as valuable resources for TDC progress in specific languages. However, these resources are of limited scope and availability, typically restricted to particular domains such as educational texts or children's literature. Limited coverage of this kind prevents scaling TDC models to more complex writing varieties, such as news articles, blog posts, or social media.

Multi-lingual benchmarks, such as CLEAR, XGLUE, and MLQA, can be analyzed comparatively across dozens of languages. The benchmarks define standard test conditions that can be utilized to test cross-lingual transfer and model strength. However, they overwhelmingly favor popular languages such as Spanish, German, or French, at the expense of underrepresented languages like Yoruba or Inuktitut. Therefore, model performance over these benchmarks may not generalize to real-world deployment over dozens of distinct linguistic ecosystems.

To fill these gaps, data augmentation techniques such as back-translation, synthetic data generation, and crowdsourced annotation have become increasingly used. Back-translation generates auxiliary training pairs by translating them into a pivoting language (e.g., English) and then back to the target language. In contrast, synthetic generation creates pseudo-annotated examples through large language models. Crowdsourcing is a scaled-up annotation process that can vary in quality if its annotators are uninformed or inconsistent. Quality control procedures and dataset diversity measures will be crucial for developing robust cross-lingual TDC systems.

4.2 Evaluation strategies

Robust evaluation techniques are necessary to evaluate the effectiveness and generalizability of multi-lingual and multi-cultural coverage cross-lingual TDC models. Because of TDC's multi-lingual and multi-cultural coverage, it is impossible to assess these by restricting to standard metrics, as this would not accommodate finer-grained linguistic complexities and audience expectations. Accuracy, F1-score, Kendall's τ, and Spearman's ρ are typically used to measure performance, giving classification and ranking information. These metrics can conceal systematic biases, particularly when comparing languages with high resources to those with low resources. It is necessary to incorporate multiple complementary metrics to better understand the model's behavior across different datasets.

Another critical point is cross-language verification, in which models are verified for multiple target languages rather than a single source-target pair. It reveals how much models generalize across typologically different languages and varying scripts. If a system performs well over English–German and poorly over English–Finnish sets, it may be sensitive to morphological complexity. Cross-language verification is also a fairer benchmark for comparing methodologies, such as machine translation pipelines, multi-lingual transformers, or meta-learning methods.

Qualitative analysis offers further nuance by searching misclassified/borderline instances for cultural/linguistic mismatches. Idiomatic expressions, for example, or culturally based allusions may confuse models even with favorable overall metrics. Blending qualitative insights with best practices for reproducibility, involving sharing hyperparameters, pre-training details, and test scripts, opens up and facilitates progress in the field. Collectively, as shown in Table 9, combining quantitative and qualitative methods, cross-language

Table 9: Key components of evaluation strategies for cross-lingual TDC

Component	Description	Benefit	Applications
Metrics (Accuracy, F1, Kendall's τ, Spearman's ρ)	Quantitative measures of classification and ranking performance.	Enables standardized performance comparison	Evaluating TDC on multi-lingual benchmarks like XGLUE.
Cross-language validation	Testing across multiple target languages to assess generalization.	Identifies strengths and weaknesses across languages	Comparing English-trained models on German, Finnish, etc.
Qualitative analysis	Manual review of misclassifications and edge cases.	Reveals cultural or linguistic biases	Analyzing idiomatic or genre-specific errors.
Reproducibility protocols	Sharing hyperparameters, datasets, and scripts for transparent evaluation.	Ensures fairness and repeatability	Public leaderboards and shared evaluation repositories.

verification, and procedures for reproducibility allows for an all-around evaluation framework that can quantify both performance and fairness.

4.3 Challenges and limitations

Cross-lingual TDC is hindered by numerous long-standing issues that limit scalability, fairness, and accuracy. One of the most considerable challenges is morphological richness, particularly for agglutinative or heavily inflected languages such as Finnish, Turkish, or Hungarian. For languages of this type, grammaticality is indicated by prolific affixation, i.e., by a word that would be an entire phrase in English. English-oriented feature-based or transformer-based models tend to incorrectly analyze tokens of this type, leading to underestimation and overestimation of difficulty. Defeating it means developing language-specific preprocessing or subword tokenization techniques to maintain morphological variations better.

Another notable limitation is the presence of translation artifacts, particularly for machine translation-based methods in low-resource languages. Automated translation programs may simplify advanced structures or miss culture-bound and idiomatic expressions. Distortion of TDC predictions can be biased and provide an incorrect representation of the hardness of the textual origin. Translating into English, for example, a culturally idiomatic text may lose subtleties, making it even more complicated for native users. These artifacts require various solutions, such as multi-lingual pre-trained models or hybrid methods that don't significantly rely on translation.

Variability across domains and genres is also a problem. TDC models trained on formal writings, such as research articles or encyclopedia entries, may not generalize to informal domains, including blog posts, social media postings, or transcribed oral discourses, which have quite different sentence boundaries, word usage, and stylistic routines. As the variability over domains is significant, it is challenging to generalize over domains. Low-resource languages exacerbate all these problems. Most lack annotation collections, large-scale pre-training data, or language resources such as morphological analyzers. As shown in Table 10, all these intertwined aspects complicate model construction and highlight the need for diversified datasets, sound evaluation regimes, and methods attuned to linguistic variability.

4.4 Future directions

The future of cross-lingual TDC lies in developing methods that balance precision, efficacy, and fairness in different languages and conditions. Of interest is multimodal TDC, which combines visual, audio, or contextual cues with written language. Examples include children's picture books or language learning books, which typically incorporate images accompanied by written text. Adding these multimodal cues may complement predictions of difficulty. Additionally, incorporating prosodic or speaking characteristics into audio resources can enhance evaluation on multi-lingual learning platforms. Multimodal methods could be extended to benefit accessibility aids by varying their complexity according to the text and accompanying media.

Another research area is architecture incorporating language-agnostic representations and language-specific adapters/prompters. Such models can leverage higher proficiency by hybridizing universal multi-lingual models that learn idiosyncratic linguistic characteristics. Active and few-shot learning can be similarly employed to rapidly augment coverage for low-resource languages and reduce over-reliance on costly annotation campaigns. Models can be fine-tuned iteratively by leveraging human-in-the-loop mechanisms with minimal label effort. Such procedures are particularly valuable in teaching

Table 10: Key challenges and limitations in cross-lingual TDC

Challenge	Description	Impact on TDC	Potential mitigation strategies
Morphological complexity	Agglutinative or highly inflected languages complicate feature extraction.	Reduced accuracy and misclassification.	Use subword tokenization and morphology-aware models.
Translation artifacts	Biases or errors introduced by machine translation distort difficulty measures.	Over/underestimation of text complexity.	Favor multi-lingual embeddings or hybrid translation-free approaches.
Domain & genre variability	Formal vs. informal texts vary in structure and vocabulary.	Limited cross-domain generalization.	Train on diverse genres and use domain adaptation methods.
Low-resource languages	Lack of labeled data and pre-training corpora for many languages.	Hinders scalability and fairness.	Use few-shot learning, active learning, or crowdsourced labeling.

Table 11: Promising future directions for cross-lingual TDC

Future direction	Potential benefit	Implementation challenge	Applications
Multimodal TDC	Captures visual/audio context to improve accuracy	Requires multimodal datasets and annotation	Children’s book recommendation, accessible e-learning
Hybrid architectures	Balances universality and language-specific adaptation	Increased model complexity	Bilingual education platforms, adaptive tutors
Active/few-shot learning	Quickly expands low-resource datasets	Needs expert involvement for quality assurance	Rapid deployment for underrepresented languages
Fairness and bias mitigation	Ensures equitable performance across languages	Developing culturally sensitive metrics	Inclusive educational and accessibility tools
Standardized evaluation	Improves reproducibility and comparability	Reaching consensus across research communities	Benchmarks for multi-lingual NLP and TDC research

environments where expert intervention ensures the pedagogical accuracy of difficulty ratings.

Inclusiveness and fairness should also guide the development of TDC in the future. Optimized TDCs that are biased in favor of imbalanced data may perform poorly for underrepresented cultures/languages. Researchers, therefore, must construct bias-detection tools, culture-adaptive evaluation metrics, and balanced test sets that better reflect the world's linguistic diversity.

Standard evaluation frameworks, such as those presented in Table 11, must also be established to facilitate the comparability of studies and enable replicability. Such frameworks must be based on cross-domain testing, culture-adaptive metrics, and data provenance reporting. Following these routes, future research can develop transparent, scalable, and fair cross-lingual TDC systems, enhancing global knowledge sharing, accessibility, and educational support.

5 Conclusion

Cross-lingual TDC has emerged as a crucial research area at the intersection of natural language processing, machine learning, and accessibility. Cross-lingual methods have enabled scalable solutions that generalize from English-centric early TDC systems by transferring knowledge from higher-resource to lower-resource languages. Machine translation pipelines, multi-lingual pre-trained transformers, adversarial and meta-learning techniques, and knowledge distillation architectures advanced the field and opened practical routes to empower various linguistic populations.

Despite these advances, many challenges remain. The complexity of morphology, translation-based artifacts, variability of domains, and the availability of limited-quality datasets continue to limit the precision and accuracy of TDC systems. Evaluation measures must be advanced to include culturally sensitive metrics, cross-language usage verification, and qualitative analyses to ensure a resilient and justifiable operating capability. These issues must be addressed to deliver adaptive learning technologies and usable information resources across languages and territories. Future research should incorporate multimodal inputs, hybrid structures, active learning, and fairness-aware techniques. The field can construct resource-aware, inclusive systems to facilitate knowledge access by collaborating with computational linguists, pedagogues, and culture specialists. Cross-lingual TDC can thus transcend linguistic boundaries to empower learners, teachers, and information systems worldwide.

References

[1] M. Nalini, R. K. Dhanraj, B. Balusamy, V. Abirami, and K. Kavya, "Intelligent Assistants Using Natural Language Processing for Hyperautomation," *Hyperautomation for Next-Generation Industries,* pp. 91–126, 2024.

[2] C. Troussas, A. Krouska, and C. Sgouropoulou, "Case Studies of Interactive Machine Learning for Adaptive Learning Technology Systems," in *Human-Computer Interaction and Augmented Intelligence: The Paradigm of Interactive Machine Learning in Educational Software*: Springer, 2025, pp. 347–385.

[3] S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma, "Text-based question difficulty prediction: A systematic review of automatic approaches," *International Journal of Artificial Intelligence in Education,* vol. 34, no. 3, pp. 862–914, 2024.

[4] V. Dogra *et al.*, "A complete process of text classification system using state-of-the-art NLP models," *Computational Intelligence and Neuroscience,* vol. 2022, no. 1, p. 1883698, 2022.

[5] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," *Journal of Network and Computer Applications,* vol. 97, pp. 23–34, 2017, doi: https://doi.org/10.1016/j.jnca.2017.08.006.

[6] S. M. Marier, X. Chen, L. Zhu, and X. Kong, "Grammatical error correction for low-resource languages: a review of challenges, strategies, computational and future directions," *PeerJ Computer Science,* vol. 11, p. e3044, 2025.

[7] K. Nguyen-Viet, "Navigating personal name avoidance in artificial intelligence: challenges, adaptations, and ethical considerations," *Ethics and Information Technology,* vol. 27, no. 3, pp. 1–12, 2025.

[8] F. Pallucchini, L. Malandri, F. Mercorio, and M. Mezzanzanica, "Lost in Alignment: A Survey on Cross-Lingual Alignment Methods for Contextualized Representation," *ACM Computing Surveys,* 2025.

[9] E. Hashmi, S. Y. Yayilgan, and M. Abomhara, "Metalinguist: enhancing hate speech detection with cross-lingual meta-learning," *Complex & Intelligent Systems,* vol. 11, no. 4, p. 179, 2025.

[10] B. Wang, "A Hybrid Fuzzy Logic and Deep Learning Model for Corpus-Based German Language Learning with NLP," *Informatica,* vol. 49, no. 21, 2025.

[11] Y. Ren, W. Fan, and J. Wang, "Intelligent text analysis for effective evaluation of english Language teaching based on deep learning," *Scientific Reports,* vol. 15, no. 1, p. 28949, 2025.

[12] L. Li, "Comparative Performance of Neural Networks and Ensemble Methods for Command Classification in ALEXA Virtual Assistant," *Informatica,* vol. 49, no. 2, 2025.

[13] H. Rizwan Iqbal, M. Sharjeel, J. Shafi, U. Mehmood, S. U. Hassan, and A. A. Raza, "Urdu paraphrased text reuse and plagiarism detection using pre-trained large language models and deep hybrid neural networks," *Multimedia Tools and Applications,* pp. 1–23, 2025.

[14] R. Piperno, L. Bacco, F. Dell'Orletta, M. Merone, and L. Pecchia, "Cross-lingual distillation for domain knowledge transfer with sentence

transformers," *Knowledge-Based Systems,* vol. 311, p. 113079, 2025.

[15] P. Q. Huy, "Cross-Lingual Evidence-Based Strategies for Identifying Fabrications in Neural Translation Systems," *Transactions on Artificial Intelligence, Machine Learning, and Cognitive Systems,* vol. 9, no. 11, pp. 1–10, 2024.

[16] J. Rodríguez-Miret, E. Farré-Maduell, S. Lima-López, L. Vigil, V. Briva-Iglesias, and M. Krallinger, "Exploring the Potential of Neural Machine Translation for Cross-Language Clinical Natural Language Processing (NLP) Resource Generation through Annotation Projection," *Information,* vol. 15, no. 10, p. 585, 2024.

[17] M. B. Bagherabad, E. Rivandi, and M. J. Mehr, "Machine Learning for Analyzing Effects of Various Factors on Business Economic," *Authorea Preprints,* 2025, doi: https://doi.org/10.36227/techrxiv.174429010.09842200/v1.