I.J. Information Technology and Computer Science, **2026, 2, 146-160**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2026.02.09

Lightweight and Explainable Neural Models for Multilingual Movie Script Certification

Pratik N. Kalamkar*
Shri Jagdishprasad Jhabarmal Tibrewala University/Department of Computer Science and Engineering, Jhunjhunu, Rajasthan, India
E-mail: pratik2kcn@gmail.com
ORCID iD: https://orcid.org/0009-0002-5714-1843
*Corresponding author

Prasadu Peddi
Shri Jagdishprasad Jhabarmal Tibrewala University/Department of Computer Science and Engineering, Jhunjhunu, Rajasthan, India
E-mail: peddiprasad37@gmail.com
ORCID iD: https://orcid.org/0000-0001-9717-934X

Yogesh K. Sharma
K L Deemed to be University (KLEF)/Department of Computer Science and Engineering, Guntur, Andra Pradesh, India
E-mail: dr.sharmayogeshkumar@gmail.com
ORCID iD: https://orcid.org/0000-0003-1934-4535

Received: 10 October 2025; Revised: 28 November 2025; Accepted: 08 February 2026; Published: 08 April 2026

Abstract: Automated film certification remains an underexplored regulatory challenge, requiring scalable yet transparent models capable of handling full-length multilingual scripts. This paper presents a unified framework that delivers lightweight, explainable, and calibrated neural classifiers for multilingual movie script certification in English, Hindi, and Marathi. Unlike prior studies that operate on short snippets or monolingual text, our approach models entire scripts through chunk-level transformer encoding, knowledge distillation, and file-level temperature calibration, coupled with explainability-guided rule mapping for interpretable decision refinement. The proposed pipeline systematically integrates six stages—baselines, teacher modeling, distillation, calibration, explainability, and rule enrichment—yielding a compact yet trustworthy system. Experiments show that the distilled students retain over 85% of teacher accuracy while being 3× smaller, and temperature scaling substantially improves reliability (English Expected Calibration Error 0.303→0.086, Brier 0.684→0.540). Faithfulness analysis using deletion Area Under Curve confirms interpretable token attributions (0.157, 0.239, and 0.258 for English, Hindi, and Marathi respectively). Moreover, rule integration improves accuracy (English 0.581→0.587) while offering human-auditable rationales. All models are deployment-feasible, exported to ONNX/TorchScript with 3.5× compression (545 MB→150 MB) and no performance loss. Together, these results establish a reproducible, end-to-end pipeline that works multilingual long-document modeling, calibration, and interpretability for film certification—advancing trustworthy Artificial Intelligence in regulatory Natural Language Processing. To our knowledge, this is the first work to build a unified, multilingual, and explainable pipeline for movie script certification using full-length scripts across MPAA and CBFC regulatory settings.

Index Terms: Natural Language Processing, Deep Learning, Text Processing, Classification, Transfer Learning, Calibration, Explainable Artificial Intelligence (XAI), Movie Scripts.

1. Introduction

Automating film certification has become increasingly relevant in the era of large-scale digital content production. Certification boards such as the Central Board of Film Certification (CBFC) and the Motion Picture Association of America (MPAA) assign age-based categories to guide audiences and enforce media regulations. However, manual certification is slow, subjective, and difficult to scale, especially when full scripts span diverse linguistic and cultural contexts. This motivates the development of reliable natural language processing (NLP) systems that can assist human

reviewers with transparent and consistent recommendations.

Existing research has primarily relied on short text sources such as subtitles, reviews, or scene descriptions [1–5]. While these approaches achieve reasonable accuracy, they do not address core challenges in real-world certification: long scripts often exceed encoder limits; regulatory settings require calibrated probability estimates and interpretable decisions; and no prior work offers a unified multilingual benchmark to evaluate cross-language robustness. As a result, despite progress in transformer-based modeling, there remains no deployable or trustworthy system for multilingual movie script certification. To the best of our knowledge, no prior work has jointly addressed long-document modeling, multilingual variability, calibration, interpretability, and deployment efficiency within a single unified framework.

This work fills these gaps by proposing a lightweight, explainable, and multilingual pipeline for automatic movie script certification. Our framework integrates chunk-level modeling, teacher–student distillation, temperature-based calibration, and explainability-guided rule mapping. Together, these components support accurate, well-calibrated, and auditable predictions for full-length scripts across English, Hindi, and Marathi.

Empirical results demonstrate that the distilled student models retain over 85% of teacher accuracy while being three times smaller. Calibration significantly improves probability reliability (e.g., English ECE 0.303→0.086), and deletion-based faithfulness confirms the quality of token-level attributions. Hybrid neural–rule reasoning further enhances interpretability and yields modest accuracy gains. Finally, ONNX and TorchScript exports with dynamic quantization achieve 3.5× compression with no accuracy loss, enabling practical deployment.

In summary, this study introduces the first multilingual benchmark and deployable pipeline for movie script certification that jointly integrates long-document modeling, distillation, calibration, and rule-based interpretability. By coupling neural explainability with symbolic rules, the proposed framework offers a reproducible and practical foundation for transparent decision-making in media governance.

Contributions

To clearly highlight the novelty of this work, we summarize our main contributions as follows:

- First multilingual benchmark for movie script certification. We compile and standardize full-length scripts across English, Hindi, and Marathi—languages with distinct certification authorities (MPAA and CBFC). Prior work focuses only on monolingual or short-text settings.
- Unified long-document modeling pipeline. We introduce a practical framework that combines chunk-level transformer encoding, file-level aggregation, teacher–student distillation, and post-hoc temperature calibration—an end-to-end design not explored in earlier studies.
- Explainability-guided rule mapping. We propose a novel mechanism that links gradient-based token attributions to symbolic rule lists, enabling human-auditable rationales and bridging neural interpretability with rule-based transparency.
- Comprehensive evaluation of trustworthiness. Beyond accuracy, we assess calibration (ECE, Brier score), deletion-based faithfulness, and cross-method attribution correlations, providing the first reliability analysis for multilingual film certification.
- Efficient and deployable models. Dynamic quantization, ONNX/TorchScript export, and 3.5× compression yield lightweight student models suitable for deployment on commodity hardware—demonstrating practical readiness.

2. Related WorkS

Recent research has developed models such as RNNs, Transformers, and multi-aspect classifiers to predict age ratings (e.g. MPAA, CBFC) directly from movie scripts. These models analyze various content aspects—like violence, sex, profanity, and substance use—to determine severity and suitability for different age groups. Hierarchical and attention-based architectures have shown improved accuracy, with some models outperforming previous state-of-the-art methods by 3–10 F1 score points [6–8]; importantly, hierarchical designs that mirror document structure (words→sentences→document) are a natural fit for long texts [9] and the transformer family (BERT) underpins modern contextual encoders used throughout this work [10].

Surveying the field, we and others have documented a fragmented methodological landscape with reproducibility issues and a lack of standardized benchmarks; the survey therefore advocates more public datasets, benchmarking standards, and multimodal/reproducible frameworks to advance automated movie analytics [11]. In prior work specifically on scripts and rating prediction we showed that lexicon/lexical-baseline methods (including blended lexicon sentiment scoring) are still useful baselines for script-level tasks and that carefully designed hierarchical/ordinal frameworks can improve robustness on full-length scripts [12–14].

Interpretability is a key concern, as rating decisions must be explainable, and that's because film ratings are high-stakes, social decisions that affect people, companies, and legal compliance, so automated ratings must be transparent, contestable and auditable. Multi-aspect classification and visualization strategies have been introduced to provide explanations for model predictions, helping reviewers and stakeholders understand which script elements influenced the rating. This is crucial for transparency and trust in automated systems [6, 7].

Prior NLP work on film and media has mainly investigated subtitles, reviews, and screenplay-level narrative analysis, producing useful datasets and methods for summarization, genre/scene extraction, and user-opinion mining, but these studies typically operate on short or mid-length text (subtitles, plot summaries, reviews, and scene-level script data) rather than full-length scripts and therefore do not directly address certification-style long-document classification [4, 5, 15–25]; script corpora such as ScriptBase exemplify this emphasis on narrative/scene analysis rather than regulatory labels [26].

At the same time, transformer-based advances and long-document architectures have made processing lengthy documents feasible: sparse and local-attention models (e.g., Longformer, BigBird) enable linear or near-linear scaling to thousands of tokens [27–31], and the transformer paradigm itself (BERT) forms the technical foundation for most modern encoders [10]. While multilingual pretrained encoders (mBERT, XLM-R) provide cross-lingual representations that can transfer knowledge between typologically diverse languages; nevertheless, multilingual models remain uneven across resource conditions and are not by themselves a complete solution for domain-specific, long-script tasks [32–35]. Multilingual pretrained encoders (mBERT, XLM-R) provide cross-lingual representations and benchmarks such as XNLI and MLDoc quantify cross-lingual transfer [36–38].

Complementary techniques from model compression and calibration are central when moving from research prototypes to fielded systems: knowledge distillation (and practical instantiations such as DistilBERT) reduces latency and memory while preserving much of the teacher's performance [39–43], and simple post-hoc temperature scaling reliably corrects overconfident probabilities—properties that are essential when predictions inform high-stakes regulatory decisions [44–47].

For interpretability and operational safety, gradient-based attributions (e.g., Integrated Gradients) [48–52] and perturbation-faithfulness tests (deletion/insertion) [48,53–55] are commonly used to produce token-level evidence, though attention weights alone are known to be an imperfect explanation [48,56–59] and should be combined with other attribution signals [48, 60–62]; this motivates hybrid designs that align model attributions with symbolic lexical rules for auditable overrides [60–64].

Existing approaches hence face practical hurdles, full-length scripts exceed standard encoder limits and therefore demand careful chunking and long-document models [14,27]; knowledge-distillation and compression can alter probability distributions and degrade calibration or explanation fidelity [65–67]; and token-level attributions are often unfaithful or unstable when mapped to symbolic rules—a problem amplified across chunks and languages [64, 68, 69]. These technical issues are compounded by multilingual and low-resource gaps and the scarcity of shareable full-script corpora (notably for Indic languages) [13, 35, 38], while community incentives and evaluation practices still favor single-metric accuracy over deployable trustworthiness [70]. Legal/policy constraints further limit dataset sharing and operational adoption [71].

Collectively, these literature provide the tooling (long-document transformers, multilingual encoders, distillation, calibration, explainability, and deployment tooling) but stop short of a unified, reproducible benchmark and pipeline that (i) treats certification as a multilingual long-text classification problem, (ii) couples chunk-level modeling with chunk-level distillation and file-level aggregation, (iii) evaluates calibrated probabilities and attribution faithfulness, and (iv) demonstrates practical export/quantization for real-world deployment—gaps this paper addresses.

3. Dataset

We used a multilingual benchmark dataset for movie script certification across three languages: English, Hindi, and Marathi [13,14]. Each script is labeled with its official certification category according to the Motion Picture Association of America (MPAA) and the Central Board of Film Certification (CBFC) [71, 72]. The dataset consists of full-length scripts rather than subtitles, ensuring that contextual cues such as scene descriptions, violence, and thematic elements are retained [1–3]. This makes the task more challenging but also more representative of the actual decision-making process used by regulatory boards and agencies. Table 1 summarizes the dataset statistics. The English corpus is the largest with 1,142 scripts spanning five classes (G, PG, PG-13, R, NC-17). The distribution is skewed toward R-rated content, highlighting a significant class imbalance. The Hindi subset contains 203 scripts across three categories (U, UA, A), reflecting CBFC certification conventions. Here, the UA class dominates, again introducing imbalance. The Marathi subset consists of 100 scripts, balanced evenly between U and UA, making it a useful low-resource but balanced testbed.

To ensure reproducibility and fair evaluation, we adopt stratified splits into training, validation, and test sets [73–77] at the file level. Stratification preserves class ratios across splits, while low-frequency classes are assigned using random splits when stratification is not feasible [78–81]. This is especially important in the English dataset, where classes such as NC-17 are rare.

Overall, the dataset design reflects three complementary challenges: (i) large-scale, imbalanced, fine-grained classification (English), (ii) medium-scale, moderately imbalanced multilingual classification (Hindi), and (iii) low-resource but balanced classification (Marathi). This diversity enables evaluation of methods under both high and low-resource conditions, motivating the choice of macro-F1 as the primary evaluation metric [70, 82–85].

Table 1. Dataset statistics across languages and certification labels

Language	Total	Labels	Train/Val/Test
English	1142	G:65, PG:153, PG-13:290, R:595, NC-17:39	799 / 171 / 172
Hindi	203	U:52, UA:93, A:58	142 / 30 / 32
Marathi	100	U:50, UA:50	69 / 15 / 16

It is important to clarify that MPAA and CBFC ratings are not merged or aligned into a single unified label set. Each language is evaluated strictly under its native certification scheme: MPAA labels are used only for English scripts, while CBFC labels are used independently for Hindi and Marathi scripts. No cross-standard mapping or pooling of categories is performed. This design preserves the integrity of each regulatory system and prevents label contamination across languages.

Dataset Availability: To support reproducibility, processed dataset, label files, and code used in this study are publicly available in the GitHub repository titled "LightweightExplainableMovieCerti". The raw scripts cannot be released due to copyright restrictions, but all processed resources required to reproduce the experiments are accessible through the repository.

4. Problem Definition and Evaluation Protocol

4.1. Task Definition and Label Taxonomy

We frame movie script certification as a supervised text classification task. Each input is a full movie script d_i written in English, Hindi, or Marathi, and the goal is to predict its official certification label y_i. Because certification schemes differ by language and region, we evaluate models separately on each language under its native label taxonomy:

- English: five classes following MPAA-style ratings, L_{en} = {G, PG, PG-13, R, NC-17}.
- Hindi: three classes following CBFC conventions, L_{hi} = {U, UA, A}.
- Marathi: two classes following CBFC conventions, L_{mr} = {U, UA}.

All results are reported in these native label sets. We emphasize that this design intentionally reflects the heterogeneity of real-world certification unbalance in movie releases: English involves fine-grained distinctions across five classes, Hindi shows mostly three categories in wide spread releases, and Marathi provides a balanced low-resource two-class scenario with almost no releases in A category.

4.2. Formal Task Definition

Given a training set $\mathcal{D} = \{(d_i, y_i)\}_{i=1}^{N}$, where di is the i-th movie script and $y_i \in L_\ell$ is its certification label from the label set L_ℓ for language $\ell \in$ {en, hi,mr}, the objective is to learn a classifier f_θ parameterized by θ such that $f_\theta(di) \approx y_i$. With lightweight and explainable neural models. Since scripts are long (often tens of thousands of tokens), they exceed the maximum input length of transformer encoders. Each script d_i is therefore split into n_i contiguous chunks:

$$d_i \rightarrow \{c_{i1}, c_{i2}, \ldots, c_{ini}\},$$

where c_{ij} denotes the j-th chunk of script i.

For each chunk c_{ij} , the model produces a logit vector $z_{ij} \in R^C$, where $C = |L_\ell|$ is the number of classes in the relevant label set. The file-level logits are obtained by mean pooling across chunks as in (1):

$$\bar{z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} z_{ij}, \quad \hat{y}_i = \arg\max_{k\in\{1,\ldots,C\}} \bar{z}_i[k] \tag{1}$$

where $\bar{z}_i[k]$ is the k-th component of the aggregated logit vector.

4.3. Evaluation Protocol

We evaluate classifiers using standard classification metrics: Accuracy (proportion of correctly classified scripts), and Macro-F1 (unweighted mean of per-class F1 scores).

In addition to accuracy-based metrics, we assess the reliability of predicted probabilities. Two calibration metrics are used:

Brier score. For N test scripts and C classes, let p_{ik} denote the predicted probability that script i belongs to class k, and let $1[y_i = k]$ denote the indicator of the true class. The Brier score is defined as (2)

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}(p_{ik} - \mathbf{1}[y_i = k])^2 \tag{2}$$

Expected Calibration Error (ECE). Predictions are partitioned into M bins according to confidence. For each bin B_m, $\text{acc}(B_m)$ is the empirical accuracy and $\text{conf}(B_m)$ is the mean predicted confidence. The ECE is computed as (3)

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{3}$$

where $|B_m|$ is the number of predictions in bin m.

Lower Brier and ECE values should indicate better calibration of predicted probabilities [65, 86].

5. Methodology

Figure 1 provides an overview of our experimental pipeline, which integrates baseline models, transformer-based teacher models, knowledge distillation, calibration, explainability, and rule-based refinement into a unified framework for multilingual script certification. Each component is described in detail below.

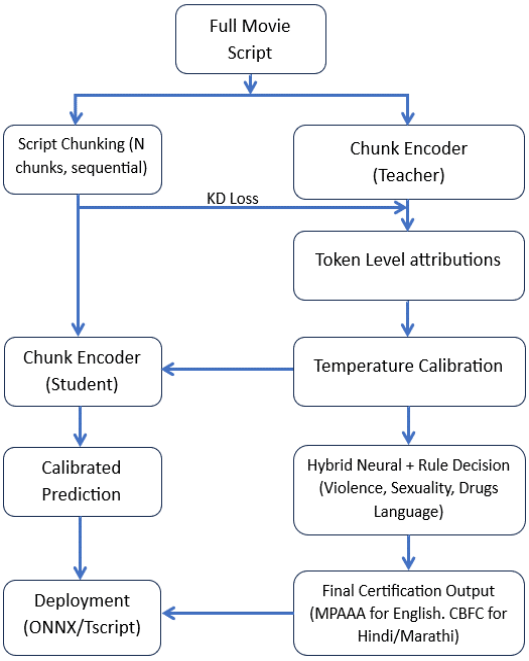


Fig.1. Overview of the experimental pipeline. Scripts are preprocessed, split into chunks, classified by teacher and student models, calibrated, explained, and refined through rule mapping for final predictions

We begin with baseline models to validate dataset quality and establish reference, before progressing to more sophisticated transformer-based approaches.

5.1. Baseline Models

As the first step in our pipeline, we establish classical baselines to validate the dataset and provide lower-bound performance estimates. Scripts are represented using the Term Frequency–Inverse Document Frequency (TF–IDF) scheme. For a term t in document d within a corpus D, the weight is defined as in Equation 4.

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \log \frac{|D|}{|\{d' \in D : t \in d'\}|} \tag{4}$$

where $\text{tf}(t, d)$ is the frequency of term t in document d, $|D|$ is the total number of documents, and $|\{d' \in D : t \in d'\}|$ is the number of documents containing t. In our experiments we used scikit-learn's TfidfVectorizer with ngram range=(1,1), max features=10000, min df=2, max df=0.98, and strip accents="unicode".

Two linear classifiers were evaluated: Logistic Regression (C=1.0, solver="lbfgs", max_iter=2000) and a Linear Support Vector Machine (LinearSVC, C=1.0, max_iter=10000, dual=False). Since linear SVMs do not produce calibrated probabilities, we applied Platt scaling [87] using CalibratedClassifierCV with sigmoid calibration and cv=2 (a pragmatic choice that reduces runtime while maintaining stable calibration; pilot runs with cv=3 and cv=5 produced nearly identical metrics).

Performance is reported using Accuracy and Macro-F1, while probability quality is assessed with Expected Calibration Error (ECE, computed over 15 bins using top-label confidence) and a multiclass Brier score. The latter is implemented as the mean of one-vs-all Brier scores across classes, a standard extension of the binary Brier definition to multi-class settings. While these linear models here establish reliable baselines, they are inherently limited in capturing

long-range dependencies and semantic richness in scripts of different languages [88–91]. To address these limitations, we next introduce transformer-based teacher models capable of handling richer contextual information.

5.2. Teacher Model with Chunking

Building on the baseline results, we employ transformer-based teacher models. However, since scripts are often tens of thousands of tokens long—far exceeding the input limits of standard encoders—we adopt a chunking / hierarchical strategy that mirrors hierarchical document representations [9]: each script is split into overlapping segments of fixed length, processed independently (or with long-document encoders such as Longformer/BigBird) [27, 31], and re-aggregated into file-level predictions. The file-level logits are obtained by mean pooling across chunks as in (1). While mean aggregation is the primary strategy, we also evaluated alternatives such as max pooling and majority voting. Teacher models are based on DistilBERT and XLM-R, fine-tuned using the AdamW optimizer with a linear learning-rate schedule and warmup, gradient clipping, and early stopping based on validation Macro-F1. Chunk-level logits from the teacher model are stored for later use in knowledge distillation.

5.3. Knowledge Distillation

Teacher models deliver strong performance, but their computational cost makes them impractical for large-scale deployment. To address this, we distill knowledge from the teacher into a smaller student model. Distillation aligns the student's output distribution with both the teacher's soft predictions and the ground-truth labels. We minimize the combined loss in (5) – (6) with softened distributions (7).

$$\mathcal{L} = \alpha\, \mathcal{L}_{\mathcal{KD}} + (1 - \alpha)\, \mathcal{L}_{\mathcal{CE}} \tag{5}$$

where L_{CE} is the cross-entropy loss with the true labels, and L_{KD} is the Kullback–Leibler (KL) divergence between softened teacher and student distributions:

$$\mathcal{L}_{\mathcal{KD}} = T^2\, D_{KL}(P_T \,||\, P_S) \tag{6}$$

Here $z_T \in \mathrm{R}^C$ and $z_S \in \mathrm{R}^C$ are the teacher and student logits respectively, with C the number of classes. The softened probability distributions are

$$P_T = \mathrm{softmax}\left(\frac{z_T}{T}\right), \quad P_S = \mathrm{softmax}\left(\frac{z_S}{T}\right) \tag{7}$$

where $T > 1$ is the temperature parameter that smooths the distributions. The scaling factor T^2 balances gradient magnitudes between the distillation and cross-entropy terms.

This procedure yields a compact student model with significantly reduced inference cost and memory footprint, while retaining performance close to that of the high-capacity teacher. Distilled students form the basis for subsequent calibration, explainability, and rule-mapping stages in our pipeline.

5.4. Temperature Calibration

Deep neural classifiers are often miscalibrated, producing overconfident probabilities that do not reflect true correctness likelihoods [65]. To address this, we apply temperature scaling to the logits of the distilled student model. A single scalar parameter $T > 0$ is learned on the validation set by minimizing the Negative Log-Likelihood (NLL). Optimization is performed over $\log T$ using gradient-based methods (L-BFGS or Adam).

The resulting calibrated probabilities preserve the model's classification accuracy but better align predicted confidence with observed accuracy. We evaluate calibration improvements using NLL, Brier score, and Expected Calibration Error (ECE) as defined in Section Evaluation Protocol.

This post-processing step is crucial for regulatory tasks where decisions may depend on probability thresholds [67, 92–94]. By applying temperature calibration, the compact student model achieves trustworthy probability estimates while maintaining efficiency. Beyond probability reliability, however, it is equally important to provide interpretability—especially in sensitive censorship-related contexts.

5.5. Explainability

To this end, we apply two complementary explainability methods on the calibrated student model, ensuring that interpretability analyses reflect the final predictions used in downstream refinement.

- Attention-based scores: Following [95], we extract attention weights from each transformer layer and head. Although attention has often been used as a proxy for explanation, its faithfulness as an interpretability mechanism remains debated [64]. Given L layers and H attention heads, attention-based importance is computed via (8) and aggregated as (9).

$$\text{MeanAttention}_{s,t} = \frac{1}{L \cdot H} \sum_{l=1}^{L} \sum_{h=1}^{H} \text{Attention}_{l,h,s,t} \tag{8}$$

$$\text{Imp}_{t}^{\text{attn}} = \sum_{s} \text{MeanAttention}_{s,t} \tag{9}$$

Token-level importance scores are then computed as the sum of incoming attention weights to each token.

- Gradient-based scores: For a token embedding $e_i \in \mathrm{R}^d$ and the gradient of the target logit with respect to that embedding $\partial L_{\text{target}}/\partial \text{e}_\text{i}$, we compute gradient-based scores using Grad*Input as in (10) [68].

$$\text{Grad} * \text{Input}_\text{i} = \sum_{k=1}^{d} e_{i,k} \frac{\partial L_{\text{target}}}{\partial e_{i,k}} \tag{10}$$

where d is the embedding dimension and $e_{i,k}$ is the k-th component of the embedding. This score approximates Integrated Gradients and captures how perturbations of the embedding affect the model's decision.

Chunk-level importance scores are aggregated across the full script, and faithfulness is quantified via deletion tests: iteratively removing top-ranked tokens should reduce the predicted probability for the chosen class [53]. The retained probability curve is summarized using the deletion AUC, where a lower AUC indicates stronger faithfulness [69, 96].

Outputs include (i) quantitative scores such as deletion AUC per file and per language, and (ii) qualitative artifacts such as HTML highlights of sensitive tokens, which illustrate how the model grounds its certification decisions. These complementary analyses provide both human-facing interpretability and quantitative reliability checks. Importantly, the token-level evidence derived from explainability also enables us to integrate symbolic lexical rules into the pipeline.

5.6. Rule Mapping

While neural models capture patterns from data, film certification often relies on explicit lexical rules such as detecting references to violence, sexual content, or drug use. To incorporate this domain knowledge, we implement a hybrid rule-mapping system that leverages explainability outputs as triggers. Each rule r produces a candidate label y_{rule} and an associated confidence score conf$r$ $(d_i) \in [0, 1]$, computed as the fraction of top tokens in script d_i that belong to the lexical set V_r:

$$\text{conf}_\text{r}(d_i) = \frac{1}{K} \sum_{t \in \text{TopK}(d_i)} \mathbf{1}\{t \in \mathcal{V}_r\} \tag{11}$$

The final prediction combines the student model output with the rule-based override as follows:

$$\hat{y}_{final}(d_i) = \begin{cases} y_{rule,} & if\ conf_r(d_i) \geq \tau, \\ \hat{y}_{model}(d_i), & otherwise, \end{cases} \tag{12}$$

where yˆmodel is the student model's prediction and τ is a threshold tuned on the validation set.

This integration allows the system to override the neural prediction in cases where strong rule-based evidence exists. Evaluation metrics include accuracy, macro-F1, precision, recall, and confusion matrices, comparing model-only, rule-only, and hybrid predictions. By combining machine learning with interpretable rules, the system produces decisions that are both more accurate and more transparent in sensitive regulatory contexts. Building on this foundation, we next enhance the rule-based component by directly coupling it with explainability outputs and systematically tuning thresholds for rule application.

Rule Construction. The lexical rules used in the hybrid stage were built using a three-step procedure applied separately for English, Hindi, and Marathi. First, we compiled seed terms corresponding to age-relevant categories (e.g., violence, sexual content, profanity, substance use) by referencing MPAA/CBFC rating guidelines and established domain terminology. Second, we expanded these seeds using corpus statistics: tf–idf scores were computed on the training scripts to surface additional unigrams and multi-word expressions that co-occurred frequently with the seed terms. Third, the combined candidates were filtered using frequency and ambiguity heuristics and then manually screened to retain only unambiguous lexical indicators. Each rule is activated only when a matched lexical item also receives high attribution from the explainability module (attention or gradient saliency), ensuring that symbolic overrides are applied only when supported by both lexical cues and model-derived importance signals.

5.7. Explainability-Guided Rule Enrichment and Threshold Tuning

We enrich rule mapping predictions with token-level evidence and perform automated threshold tuning for rule activation. For each file, the top tokens identified by explainability methods are attached to the model predictions, producing qualitative evidence that links model behavior with lexical triggers. Each enriched record contains both the predicted label and an explainability snippet (e.g., the top 5 most important tokens).

Building on this enrichment, we introduce a token-fraction proxy to guide rule application. For each rule category (violence, sex, drugs), we compute the fraction of top tokens that match a predefined lexical list. A rule is considered triggered if this fraction exceeds a threshold. To calibrate these thresholds, we perform a grid search over candidate values on held-out validation data, selecting the configuration that maximizes accuracy. The final prediction is determined by combining the student model's output with rule overrides according to the tuned thresholds.

This step provides two benefits. First, it validates whether incorporating rule-specific keyword presence in explainability tokens can improve predictive accuracy. Second, it produces interpretable enriched outputs, allowing us to trace both the model's reasoning and the circumstances under which symbolic rules are applied. These enriched predictions serve as the foundation for the final hybrid system and for qualitative case studies. Having established the full pipeline, we now focus on preparing the models for practical deployment in multilingual regulatory settings.

5.8. Deployment and Quantization

We export the calibrated student models—augmented with explainability and rule-based refinement—to TorchScript and ONNX formats for interoperability, and apply dynamic quantization to reduce storage footprint.

6. Results & Discussions

Results for the baselines, transformer teachers, distilled students, and our final hybrid system on each language are reported in Table 2. We do not discuss calibration metrics for teacher models, since they serve only as logit generators for distillation, and calibration metrics are not defined for rule mapping, which does not operate on probabilistic outputs [66,97].

Table 2. Final results across the pipeline, per language. Metrics are Accuracy, Macro-Precision, Macro-Recall, Macro-F1 (primary), Expected Calibration Error (ECE, lower is better), and Brier score (lower is better). Best value per language/metric is bold

Lang	Model	Acc	Prec	Rec	Macro-F1	ECE ↓	Brier ↓
English	Baseline (SVC)	0.576	0.75	0.46	0.476	0.048	0.114
	Teacher	0.628	0.659	0.586	**0.591**	-	-
	Student	0.581	0.628	0.492	0.506	0.303	0.684
	+ Calibration	0.581	0.628	0.492	0.506	**0.086**	**0.540**
	+ Rules	**0.587**	0.644	0.492	0.507	-	-
Hindi	Baseline (SVC)	0.871	0.85	0.83	0.841	0.135	0.075
	Teacher	0.871	0.842	0.842	**0.843**	-	-
	Student	0.806	0.768	0.759	0.757	0.143	0.239
	+ Calibration	0.806	0.768	0.759	0.757	0.139	0.239
	+ Rules	0.806	0.768	0.759	0.757	-	-
Marathi	Baseline (SVC)	0.500	0.500	0.500	0.467	0.109	0.233
	Teacher	0.563	0.603	0.563	0.515	-	-
	Student	0.563	0.573	0.563	**0.547**	0.050	0.438
	+ Calibration	0.563	0.573	0.563	0.547	**0.049**	**0.437**
	+ Rules	0.563	0.573	0.563	0.547	-	-

Baselines and Teacher Models. Baselines perform well in Hindi (Macro-F1 0.841) due to strong lexical cues, but collapse in English (0.476) and Marathi (0.467), confirming their brittleness. Teacher models with chunk aggregation improve substantially (English 0.591, Marathi 0.515), validating our long-text design, but remain computationally heavy.

Chunk-aware distillation compresses heavy teachers while retaining accuracy. Unlike prior work on short snippets, we distill at the chunk level and re-aggregate at the file level [29,98,99] . This adaptation improves robustness: the Marathi student surpasses its teacher (0.547 vs. 0.515), and the English student retains > 85% of teacher accuracy while being 3× smaller — confirming that our models are lightweight yet effective.

Temperature scaling was applied at the file level, aligning confidence with full-script predictions rather than noisy chunk outputs. This substantially reduced miscalibration without affecting accuracy. For example, in English, Expected Calibration Error dropped from 0.303 to 0.086 and Brier score from 0.684 to 0.540, while accuracy remained unchanged (Table 2). Calibration required the largest adjustment in English (Table 3) with a learned temperature T = 3.16 reflecting strong overconfidence in the uncalibrated student (NLL 1.64 → 0.99). Hindi (T = 0.98) and Marathi (T = 0.89) required minimal correction, and their NLL remained essentially unchanged (0.36 → 0.36 for Hindi, 0.691 → 0.690 for Marathi). This pattern is consistent with the calibration metrics: English showed the largest benefit, while Hindi and Marathi were already close to well-calibrated. Taken together, these results confirm that our lightweight student models can also produce trustworthy probability estimates, a property critical in regulatory contexts where decisions hinge on calibrated confidence thresholds.

Table 3. Temperature values and Negative Log-Likelihood (NLL) before and after calibration

Language	*T*	**NLL ↓**
English	3.16	1.64 → 0.99
Hindi	0.98	0.36 → 0.36
Marathi	0.89	0.691 → 0.690

Token importance was derived from attention and gradient attributions, then validated via deletion tests and cross method agreement, this is from Explainability and Rule Integration step. Faithfulness analysis, refer figure 2, showed that removing high-importance tokens sharply reduced model confidence, with average Deletion AUC of 0.157 (English), 0.239 (Hindi), and 0.258 (Marathi). Interestingly, the correlation between attention-based attributions and gradient-based saliency is near zero for English ($r = -0.008$) and Hindi ($r = -0.001$). A near-zero correlation does not imply complementarity in a positive sense, but rather that the two attribution methods capture largely independent and potentially inconsistent signals, as also reported in prior explainability studies [64]. This behavior is expected because attention weights reflect model-internal distribution mechanisms, whereas gradient scores measure local sensitivity. For Marathi, the correlation is moderately positive ($r = 0.181$), likely due to the smaller evaluation set. Overall, these results indicate that raw attention weights should not be interpreted as faithful explanations on their own, which further motivates our use of deletion-based faithfulness tests and rule-grounding procedures. These empirically grounded signals were operationalized in rule mapping: thresholds were tuned automatically via grid search, producing interpretable rule overrides that corrected sensitive misclassifications (e.g., English accuracy 0.581 → 0.587). Thus, explainability was not only diagnostic but directly contributed to hybrid predictions, making model outputs both interpretable and accountable.

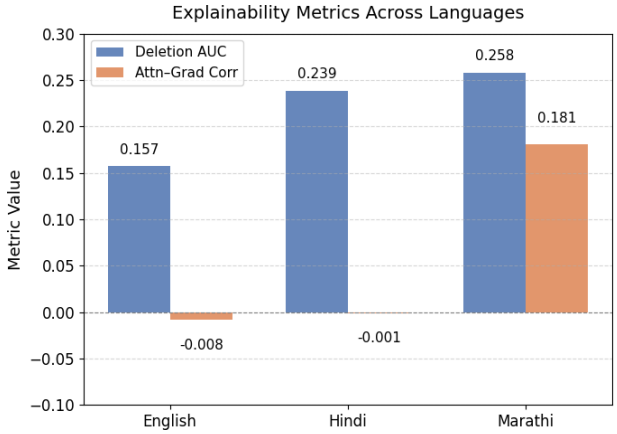
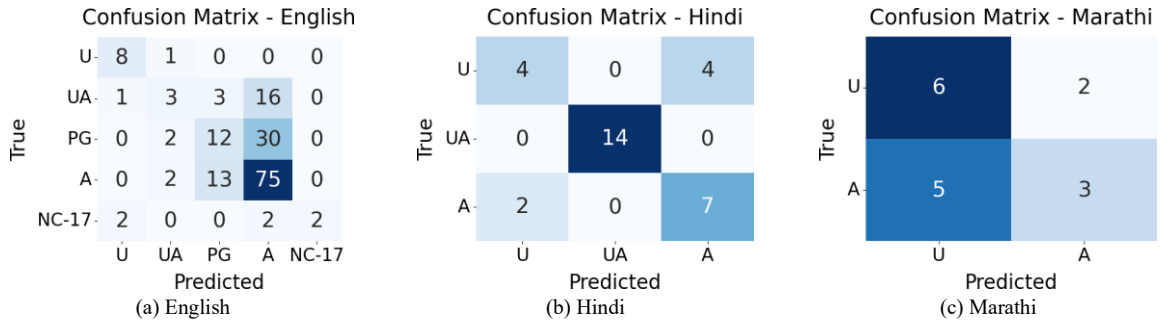


Fig.2. Explainability metrics across languages. Deletion AUC (blue, lower is better) measures faithfulness of token importance, while Attention–Gradient correlation (orange, higher indicates stronger agreement) measures consistency between two attribution methods. English and Hindi correlations are near zero (slightly negative), while Marathi shows moderate positive agreement

Per-Language Insights. English benefits most from calibration (ECE drop 0.303 → 0.086). Hindi already achieves high baseline accuracy, so the value lies in rule-based interpretability. Marathi benefits most from distillation, where the student generalizes better than the teacher despite low resources. These trends show that trustworthiness and explainability behave differently across resource conditions.

Across baselines, distillation, calibration, explainability, rules, and deployment, the results confirm our claim: we deliver lightweight and explainable neural models for multilingual movie script certification. The pipeline achieves efficiency, reliability, and interpretability — a combination rarely reported in multilingual regulatory NLP.

The confusion matrices provide a more granular view of model behavior across languages. For English (Figure 3a), the most frequent errors occur between UA and A, reflecting ambiguity in scripts with moderate violence or thematic intensity. Misclassification into NC-17 is rare, indicating that the model does not over-predict the most restrictive label. Hindi results (Figure 3b) show a recurring confusion between U and A, consistent with cultural differences in annotation guidelines where depictions of mild violence are sometimes inconsistently labeled. Marathi (Figure 3c) exhibits a simpler binary setting, yet the confusion between U and A remains noticeable, highlighting the challenge of low-resource scenarios where domain-specific subtleties are underrepresented in training data. These observations align with the overall quantitative metrics: English benefits from larger and more standardized training material, while Hindi and Marathi suffer from orthographic variation and data scarcity. Importantly, the confusion matrices illustrate that the errors are not random but cluster around adjacent certification categories, suggesting that the model captures broad distinctions but struggles at fine-grained boundaries—precisely where rule-based overrides and human oversight remain essential.



(a) English (b) Hindi (c) Marathi

Fig.3. Confusion matrices for (a) English, (b) Hindi, and (c) Marathi movie datasets

To evaluate practical feasibility, we exported the distilled students across all three languages, refer (Table 4). ONNX conversion and quantization reduced model size by 3.5× (545 MB → 150 MB) with no loss in accuracy. This confirms that our approach yields deployment-feasible, lightweight models suitable for real-world censorship boards.

The near-identical sizes across languages are expected because all models share the same DistilBERT backbone, which dominates parameter count. The language-specific classification head contributes only ≈1–2 MB, making dataset size and label distribution influential for accuracy but not for storage footprint.

These results demonstrate that our distilled students can be exported into deployment-feasible formats without loss in predictive performance, while achieving significant reductions in model size. Although runtime latency was not the primary focus of this stage, the consistent shrinkage across formats highlights the feasibility of deploying lightweight, calibrated, and explainable models in multilingual regulatory environments.

Table 4. Model sizes before and after deployment optimizations. Values are reported in megabytes (MB)

Language	Original	TorchScript	ONNX	Quantized
English	545.2	541.5	185.0	150.8
Hindi	545.2	541.5	184.5	150.7
Marathi	545.2	541.5	184.0	150.5

Taken together, these results show that our approach delivers not only compact and accurate models but also interpretable outputs whose explanatory faithfulness is strongest in high-resource settings and remains meaningful, though noisier, in low-resource contexts. This finding underscores the practical value of explainability for multilingual movie script certification, where trust in automated systems depends on both predictive accuracy and the ability to justify restrictive decisions with human-interpretable evidence.

Statistical Validation. To assess whether the small improvements observed across pipeline stages (Student → Calibrated and Calibrated → Hybrid), we applied McNemar's test to the per-example predictions on the test sets. For Hindi and Marathi, the calibrated and hybrid models produced identical predictions to the student model, resulting in zero discordant pairs; in such cases McNemar's test is undefined and no statistical difference can be inferred. For English, only 3 prediction differences were observed between the calibrated and hybrid models, yielding a McNemar p-value of $p > 0.05$, indicating that the improvement is not statistically significant. This is expected given the very small number of prediction changes. Overall, while the pipeline yields small accuracy improvements, its primary benefits come from calibration quality, explainability, and deployability rather than statistically large performance gains.

7. Conclusions

This work presented the first unified and reproducible framework for multilingual movie script certification, combining teacher–student distillation, temperature calibration, explainability, and symbolic rule integration into a single deployable system. Unlike previous studies focused on short or monolingual inputs, our pipeline addressed the challenge of full-length, heterogeneous movie scripts across English, Hindi, and Marathi—each with distinct label taxonomies and resource conditions. The proposed chunk-aware distillation strategy effectively compressed large transformer teachers while preserving performance: distilled students achieved over 85% of teacher accuracy while being three times smaller, and in Marathi, the student even surpassed the teacher (Macro-F1 0.547 vs. 0.515). File-level temperature calibration significantly improved probability reliability without loss of accuracy, reducing Expected Calibration Error from 0.303 to 0.086 and Brier score from 0.684 to 0.540 in English, confirming that lightweight models can also be trustworthy. Explainability analyses showed high attribution faithfulness (Deletion AUC: 0.157 for English, 0.239 for Hindi, 0.258 for Marathi), and attention–gradient correlations revealed language-specific interpretability patterns. These attributions were operationalized into explainability-guided rule mappings, yielding measurable improvements in sensitive categories (English accuracy 0.581→0.587) and ensuring that model predictions remained both interpretable and auditable.

Beyond modeling, deployment experiments validated that quantized student models retained full accuracy while reducing size by 3.5× (545 MB→150 MB), demonstrating readiness for real-world censorship boards with limited computational resources. Taken together, these results establish that regulatory-grade NLP systems can balance accuracy, calibration, interpretability, and efficiency within a multilingual, long-document setting. The framework not only advances automation for movie certification but also offers a generalizable design for building lightweight, transparent, and reliable AI systems in other high-stakes multilingual domains such as legal and media compliance.

Author Contributions Statement

Pratik N. Kalamkar – Conceptualization and Methodology: Developed the research concept and designed the methodological framework. Data Curation and Software Implementation: Acquired and curated the data, performed dataset preprocessing, and implemented the research model in software. Model Training, Validation, and Performance Evaluation: Led the model training process, validated the results using standard evaluation metrics, and benchmarked performance against existing methods.
Prasadu Peddi – Supervision: Proposed research ideas, Constructed the overall framework, and supervised project execution.
Yogesh K. Sharma – Supervision: Proposed research ideas, Constructed the overall framework, and supervised project execution.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research received no external funding.

Data Availability Statement

The dataset curated and used in this study is publicly available at the following GitHub repository: https://github.com/pratik1986/LightweightExplanableMovieCerti , accessed on 10 January 2026.

Ethical Declarations

This study did not involve human participants or animals; therefore, ethical approval was not required.

Acknowledgments

The authors declare that there are no acknowledgments for this study.

Declaration of Generative AI in Scholarly Writing

The authors used Grammarly, an AI-assisted language editing tool, solely for proofreading and improving grammar, clarity, and readability. No generative AI was used to create, modify, or interpret the scientific content of this manuscript. The authors are fully responsible for the content of the work.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
ONNX – Open Neural Network Exchange
MB – Megabyte
MPAA – Motion Picture Association of America
CBFC – Central Board of Film Certification
XAI – Explainable Artificial Intelligence
NLP – Natural Language Processing
ECE – Expected Calibration Error
BERT – Bidirectional Encoder Representations from Transformers
XLM-R – Cross-lingual Language Model – RoBERTa
XNLI – Cross-lingual Natural Language Inference
G – General Audiences
PG – Parental Guidance Suggested
PG-13 – Parents Strongly Cautioned (Under 13)
R – Restricted
NC-17 – No One 17 and Under Admitted

U – Universal
UA – Universal Adult
A – Adults Only
TF-IDF – Term Frequency–Inverse Document Frequency
KL – Kullback–Leibler Divergence
NLL – Negative Log-Likelihood
L-BFGS – Limited-memory Broyden–Fletcher–Goldfarb–Shanno
AUC – Area Under the Curve
HTML – HyperText Markup Language
SVC – Support Vector Classifier

References

[1] Irena Petrova. Features of the script as text type and its translation when subtitling video content. *Problems of cognitive and functional description of Russian and Bulgarian languages*, 2024.

[2] Ahmad S. Haider and R. Hussein. Modern standard arabic as a means of euphemism: A case study of the msa intralingual subtitling of jinn series. *Journal of Intercultural Communication Research*, 51:628 – 643, 2022.

[3] M. Guillot. The pragmatics of audiovisual translation: Voices from within in film subtitling. *Journal of Pragmatics*, 2020.

[4] Miss. Ghavate Rutuja. Movie review system using nlp and svm. *International Journal for Research in Applied Science and Engineering Technology*, 9:1354–1358, 2021.

[5] V. Bharadi, A. Malji, A. Mestri, Y. Narvekar, and A. Anerao. Movie genre detection from subtitle using nlp: A novel approach to identify movie genre based on the emotions from subtitles. *International Journal of Creative Research Thoughts (IJCRT)*, 11(1):672–675, 2023. Accessed: 2025-10-07.

[6] L. F. Pardo-Sixtos, Adrian Pastor Lopez-Monroy, Mahsa Shafaei, and T. Solorio. Hierarchical attention and transformers for automatic movie rating. *Expert Syst. Appl.*, 209:118164, 2022.

[7] Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and T. Solorio. From none to severe: Predicting severity in movie scripts. *ArXiv*, abs/2109.09276, 2021.

[8] Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar, and T. Solorio. Age suitability rating: Predicting the mpaa rating based on movie dialogues. pages 1327–1335, 2020.

[9] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1480–1489, 2016.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of NAACL-HLT*, 2019.

[11] Pratik N. Kalamkar and Yogesh Kumar Sharma. A review of research trends in movie automation. In *2025 International Conference on Computer Technology Applications (ICCTA)*, pages 154–163, 2025.

[12] Pratik N. Kalamkar and Yogesh Kumar Sharma. Blend of Vader and Textblob for movie script classification. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 875–881, 2024.

[13] Pratik N. Kalamkar and Yogesh Kumar Sharma. Lexical Baselines to Transformers: Hierarchical Ordinal Classification of Indic Movie Full Scripts. In *Proceedings of the 3rd IEEE International Conference on Computational Intelligence and Network Systems (CINS 2025)*, Dubai, United Arab Emirates, 2025. IEEE. In Press.

[14] Pratik N. Kalamkar and Yogesh Kumar Sharma. Hierarchical ordinal framework for automated movie censorship using full-length scripts. October 2025. In Press.

[15] Amr Shahin and A. Krzyz˙ak. Genreous: The movie genre detector. pages 308–318, 2020.

[16] B. A, B.V. Suresh Reddy, A. Nandam, K. Naresh, Sivakumar Depuruu, and M. Sakthivel. Sentimental analysis of movie reviews using nlp techniques. *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 713–720, 2023.

[17] J. Alqurni, M. K. Alsmadi, Hayat Alfagham, Sharaf Alzoubi, Sohayla Ihab, Ahmed Sameh, Diaa Salama Abd Elminaam, and O. I. Khalaf. Streamlining video summarization with nlp: Techniques, implementation, and future direction. *SN Comput. Sci.*, 6, 2025.

[18] Feroza D. Mirajkar, T. A. Mohanaprakash, Kaviya D, and Jasmine Fathima K. Scalable summarization of long-form video transcripts using nlp. *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, pages 672–677, 2025.

[19] Manikanta Grandhi and Sreebha Bhaskaran. A hybrid approach for multimedia summarization. *2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon)*, pages 1–6, 2022.

[20] Ansh Tulsyan, Anshul Bhardwaj, Pranjal Shukla, Jatin Verma, and Tushar Singh. Online platform for movie recommendation. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 2025.

[21] Joshua A Gross, Willie Roberson, and Bay Foley-Cox. Cs 230: Film success prediction using nlp techniques. 2021.

[22] Kyle Jorgensen, Haohong Wang, and Mea Wang. From screenplay to screen: A natural language processing approach to animated film making. *2023 International Conference on Computing, Networking and Communications (ICNC)*, pages 484–490, 2023.

[23] Mar Castillo-Campos, David Becerra-Alonso, and H. Boomgaarden. Automated detection of media bias using artificial intelligence and natural language processing: A systematic review. *Social Science Computer Review*, 2025.

[24] Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. Media framing: A typology and survey of computational approaches across disciplines. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[25] Maxime Cre´pel, Salome´ Do, Jean-Philippe Cointet, Dominique Cardon, and Yannis Bouachera. Mapping ai issues in media through nlp methods. pages 77–91, 2021. Accessed: 2025-10-07.
[26] Philip John Gorinski and Mirella Lapata. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015. ScriptBase corpus associated.
[27] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150, 2020.
[28] Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. Revisiting transformer-based models for long document classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
[29] Renzo Alva Principe, Nicola Chiarini, and Marco Viviani. Long document classification in the transformer era: A survey on challenges, advances, and open issues. *WIREs Data Mining and Knowledge Discovery*, 2025.
[30] Hai Pham, Guoxin Wang, Yijuan Lu, D. Florenˆcio, and Changrong Zhang. Understanding long documents with different position-aware attentions. *ArXiv*, abs/2208.08201, 2022.
[31] Manzil Zaheer, Guru Guruganesh, Karan Dubey, Jason Ainslie, Chris Alberti, Santiago Ontan˜on, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed, and Zhitao Ahmed. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[32] Robert Litschko, Ivan Vulic, Simone Paolo Ponzetto, and Goran Glavavs. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval*, 25:149 – 183, 2021.
[33] Rochelle Choenni and Ekaterina Shutova. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48:635–672, 2022.
[34] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *ArXiv*, abs/1904.09077, 2019.
[35] Alexis Conneau, Kartik Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzma´n, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, 2020.
[36] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4996–5001, 2019.
[37] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2475–2485, 2018.
[38] Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.
[39] Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *ArXiv*, abs/2208.11580, 2022.
[40] Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Improving post training neural quantization: Layer-wise calibration and integer programming. *ArXiv*, abs/2006.10518, 2020.
[41] Bishwash Khanal and Jeffery M. Capone. Evaluating the impact of compression techniques on task-specific performance of large language models. *ArXiv*, abs/2409.11233, 2024.
[42] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
[43] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
[44] Se´rgio A. Balanya, Juan Maron˜as, and Daniel Ramos. Adaptive temperature scaling for robust calibration of deep neural networks. *ArXiv*, abs/2208.00461, 2022.
[45] Qingyang Xi and Brian McFee. Beyond hard decisions: Accounting for uncertainty in deep mir models. 2021. Accessed: 2025-10-07.
[46] I. Vos, Ishaan Bhat, B. Velthuis, Y. Ruigrok, and Hugo J. Kuijf. Calibration techniques for node classification using graph neural networks on medical image data. pages 1211–1224, 2023.
[47] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017. ICML 2017.
[48] Javier Ferrando, Gerard I. Ga´llego, and M. Costa-jussa`. Measuring the mixing of contextual information in the transformer. *ArXiv*, abs/2203.04212, 2022.
[49] Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. Globenc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. *ArXiv*, abs/2205.03286, 2022.
[50] Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. Evaluating human alignment and model faithfulness of llm rationale. *ArXiv*, abs/2407.00219, 2024.
[51] Zhixue Zhao and Boxuan Shan. Reagent: A model-agnostic feature attribution method for generative language models. *ArXiv*, abs/2402.00794, 2024.
[52] M. B. Zafar, Philipp Schmidt, Michele Donini, C. Archambeau, F. Biessmann, Sanjiv Das, and K. Kenthapadi. More than words: Towards better quality interpretations of text classifiers. *ArXiv*, abs/2112.12444, 2021.
[53] Zhixue Zhao and Nikolaos Aletras. Incorporating attribution importance for improving faithfulness metrics. *arXiv preprint arXiv:2305.10496*, pages 4732–4745, 2023.
[54] J. P. Amorim, P. Abreu, Joa˜o A. M. Santos, Marc Cortes, and Victor Vila. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Inf. Process. Manag.*, 60:103225, 2023.
[55] Vangelis Lamprou, Athanasios Kallipolitis, and I. Maglogiannis. On the evaluation of deep learning interpretability methods for medical images under the scope of faithfulness. *Computer methods and programs in biomedicine*, 253:108238, 2024.

[56] Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. A song of (dis)agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing. *arXiv preprint arXiv:2205.04559*, pages 60–78, 2022.
[57] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. *arXiv: Computation and Language*, 2019.
[58] Samira Abnar and W. Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, pages 4190–4197, 2020.
[59] Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607*, pages 149–155, 2020.
[60] Qiang Ding, Lvzhou Luo, Yixuan Cao, and Ping Luo. Attention with dependency parsing augmentation for fine-grained attribution. *ArXiv*, abs/2412.11404, 2024.
[61] Benjamin Cohen-Wang, Yung-Sung Chuang, and Aleksander Madry. Learning to attribute with attention, 2025.
[62] Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. Key phrase aware transformer for abstractive summarization. *Inf. Process. Manag.*, 59:102913, 2022.
[63] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.
[64] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 3543–3556, 2019.
[65] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1321–1330. PMLR, 2017.
[66] Lehan Yang and Jincen Song. Rethinking the knowledge distillation from the perspective of model calibration. *ArXiv*, abs/2111.01684, 2021.
[67] Mari-Liis Allikivi, Joonas Ja¨rve, and Meelis Kull. Cautious calibration in binary classification. *arXiv preprint arXiv:2408.05120*, pages 1503–1510, 2024.
[68] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328. PMLR, 2017.
[69] Leila Arras, Franziska Horn, Gre´goire Montavon, Klaus-Robert Mu¨ller, and Wojciech Samek. Evaluating the faithfulness of explanation methods for neural models. *Information Fusion*, 2022.
[70] Juri Opitz. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12:820–836, 2024.
[71] Central Board of Film Certification. Guidelines — central board of film certification (cbfc), government of india. Accessed: 2025-10-10.
[72] Motion Picture Association. Film ratings — motion picture association. Accessed: 2025-10-10.
[73] Agboeze Jude and Jia Uddin. Explainable software defects classification using smote and machine learning. *Annals of Emerging Technologies in Computing*, 2024.
[74] Vikash R Singh, M. Pencina, A. Einstein, Joanna X. Liang, D. Berman, and P. Slomka. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports*, 11, 2021.
[75] B. Parker, Simon Gu¨nter, and J. Bedo″. Stratification bias in low signal microarray studies. *BMC Bioinformatics*, 8:326 – 326, 2007.
[76] Chansik An, Y. Park, S. Ahn, Kyunghwa Han, Hwiyoung Kim, and Seung-Koo Lee. Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results. *PLoS ONE*, 16, 2021.
[77] Rui Hou, Joseph Y. Lo, Jeffrey R Marks, E. S. Hwang, and Lars J Grimm. Classification performance bias between training and test sets in a limited mammography dataset. *PLOS ONE*, 19, 2024.
[78] Q. Doan, S. Mai, Quang Thang Do, and Duc-Kien Thai. A cluster-based data splitting method for small sample and class imbalance problems in impact damage classification. *Appl. Soft Comput.*, 120:108628, 2022.
[79] Diego Minatel, Angelo Cesar Mendes da Silva, N´ıcolas Roque dos Santos, Mariana Cu´ri, R. Marcacini, and A. Lopes. Data stratification analysis on the propagation of discriminatory effects in binary classification. *Anais do XI Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2023)*, 2023.
[80] Mohd Hafiz Zakaria, Jafreezal Jaafar, and S. J. Abdulkadir. Preliminary investigation of balanced stratified reduction (bsr) for imbalanced datasets. *2023 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pages 55–60, 2023.
[81] Christophe Molina, Lilia Ait-Ouarab, and H. Minoux. Isometric stratified ensembles: A partial and incremental adaptive applicability domain and consensus-based classification strategy for highly imbalanced data sets with application to colloidal aggregation. *Journal of chemical information and modeling*, 2022.
[82] Maria Cristina Hinojosa Lee, Johan Braet, and Johan Springael. Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted f1-scores. *Applied Sciences*, 2024.
[83] Kanae Takahashi, Kouji Yamamoto, A. Kuchiba, and T. Koyama. Confidence interval for micro-averaged f1 and macro-averaged f1 scores. *Applied intelligence (Dordrecht, Netherlands)*, 52:4961 – 4972, 2021.
[84] Tanu Sharma and Kamaldeep Kaur. Benchmarking deep learning methods for aspect level sentiment classification. *Applied Sciences*, 2021.
[85] Haoze Du, Jiahao Xu, Zhiyong Du, Lihui Chen, Shaohui Ma, Dongqing Wei, and Xianfang Wang. Mf-mner: Multimodels fusion for mner in chinese clinical electronic medical records. *Interdisciplinary Sciences, Computational Life Sciences*, 16:489 – 502, 2024.
[86] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.
[87] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Scho¨lkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, 1999.
[88] Jeena Kleenankandy and Abdul Nazeer K A. An enhanced tree-lstm architecture for sentence semantic modeling using typed dependencies. *Inf. Process. Manag.*, 57:102362, 2020.

[89] E. Mercha, H. Benbrahim, and Mohammed Erradi. Heterogeneous text graph for comprehensive multilingual sentiment analysis: capturing short and long-distance semantics. *PeerJ Computer Science*, 10, 2024.
[90] Tomas Brychcin and Miloslav Konop´ık. Latent semantics in language models. *Comput. Speech Lang.*, 33:88–108, 2015.
[91] S. Khudanpur and Jun Wu. Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling. *Comput. Speech Lang.*, 14:355–372, 2000.
[92] Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a calibration metric be both testable and actionable? *arXiv preprint arXiv:2502.19851*, 2025.
[93] Daniel Zeiberg. Learning calibrated classifiers from nonrepresentative data. 2024.
[94] Toby Ord, R. Hillerbrand, and A. Sandberg. Probing the improbable: methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research*, 13:191 – 205, 2008.
[95] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.
[96] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, page 151, 2018.
[97] Ishan Mishra, Riyanshu Jain, Dhruv Viradiya, Divyam Patel, and Deepak Mishra. Knowledge distillation with ensemble calibration. *Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing*, 2023.
[98] Xinyu Liu and Jian Yang. Enhance long text understanding via distilled gist detector from abstractive summarization. *arXiv preprint arXiv:2110.04741*, 2021.
[99] Yan Liu, Yazheng Yang, and Xiaokang Chen. Improving long text understanding with knowledge distilled from summarization model, 2024.

Authors' Profiles

Pratik N. Kalamkar, received his first Bachelor of Engineering degree in Computer Science from the prestigious Savitribai Phule Pune University, India, in the year 2011. He also has a Master of Engineering (Computer Science) degree from Savitribai Phule Pune, India, in 2014. He is currently pursuing his research in the field of Deep Learning and natural language processing as a PhD scholar from Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, India.

Prasadu Peddi currently serves as a Research Supervisor at Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, India. He has successfully guided 13 research scholars to the completion of their doctoral degrees. His scholarly contributions include the publication of 93 research papers in reputed international journals and the presentation of research at 16 national and international conferences.

Yogesh K. Sharma is a Professor and Researcher in the Department of Computer Science and Engineering at K L Deemed to be University (KLEF), Guntur, 522501, India. He also serves as a Research Coordinator, DRC Member, and Professor-cum-Research Faculty. He is an IEEE Senior Member and an ACM Professional Member. He has over 20 years of experience in research and scholarly publications.

How to cite this paper: Pratik N. Kalamkar, Prasadu Peddi, Yogesh K. Sharma, "Lightweight and Explainable Neural Models for Multilingual Movie Script Certification", International Journal of Information Technology and Computer Science(IJITCS), Vol.18, No.2, pp.146-160, 2026. DOI:10.5815/ijitcs.2026.02.09