Intelligent Decision Support for Article Translation in Multilingual Newsrooms

Zhan Liu
Media Innovation Lab, Institute of Informatics
University of Applied Sciences and Arts Western Switzerland (HES-SO Valais-Wallis)
Sierre, Switzerland
zhan.liu@hevs.ch

Adrien Bertaud
Institute of Informatics
University of Applied Sciences and Arts Western Switzerland (HES-SO Valais-Wallis)
Sierre, Switzerland
bertaud.adrien@gmail.com

Anne Darbellay
Media Innovation Lab, Institute of Informatics
University of Applied Sciences and Arts Western Switzerland (HES-SO Valais-Wallis)
Sierre, Switzerland
anne.darbellay@hevs.ch

Nicole Glassey Balet
Media Innovation Lab, Institute of Informatics
University of Applied Sciences and Arts Western Switzerland (HES-SO Valais-Wallis)
Sierre, Switzerland
nicole.glassey@hevs.ch

Abstract

We present an intelligent decision support system, powered by AI-driven prediction, designed to assist multilingual newsrooms in selecting articles for cross-regional translation as part of the digital transformation of journalism. The task is to predict whether a German-language article should be translated into French for cross-regional publication. Trained on 15,933 German-language articles from a major Swiss publisher, the system uses a hybrid architecture combining multilingual BERT embeddings with 41 engineered features and incorporates real-time editorial feedback through an active learning loop. It achieves an accuracy of 85.0%. During deployment, F1 improved from 77.5% to 81.2% after four weeks of feedback-driven exemplar refresh. Ablation studies indicate that sentiment polarity, regional relevance, and person-type named entities are the most influential features. The interface highlights key factors, ensuring transparency and consistency with editorial practice. By pairing hybrid NLP with human-in-the-loop prompting, the approach operationalizes intelligent translation triage in a live newsroom while preserving human control over final decisions.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Information systems** → *Data analytics*; Decision support systems.

Keywords

Intelligent systems, Hybrid AI models, Decision support systems, Media prediction, Multilingual newsrooms, Cross-lingual translation, Human-in-the-loop AI

SAC '26, Thessaloniki, Greece

ACM ISBN 979-8-4007-2294-3/2026/03
https://doi.org/10.1145/3748522.3779719

ACM Reference Format:
Zhan Liu, Adrien Bertaud, Anne Darbellay, and Nicole Glassey Balet. 2026. Intelligent Decision Support for Article Translation in Multilingual Newsrooms. In *The 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26), March 23–27, 2026, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3748522.3779719

1 Introduction

In multilingual environments such as Switzerland, news publishers must decide daily which stories are worth translating for different language audiences. These editorial choices, particularly between German and French, are often made under time pressure and rely heavily on intuition. Our interviews with editors at major Swiss newsrooms confirmed this. One senior editor noted, "Sometimes I just know a story will work for our French readers, but I couldn't always tell you why." This challenge exemplifies the need for predictive intelligent systems that can augment human decision-making in digital-era newsrooms. Our work situates this problem within the broader digital transformation of media workflows, where AI is increasingly used not only for automation but for human-centered decision support.

While past research has explored automation in journalism and recommendation systems [6, 12], the specific challenge of cross-regional content selection in multilingual contexts remains largely unaddressed. Traditional systems often optimize for user engagement within a single language [14, 22], not editorial decisions across language boundaries. To address this gap, we developed an AI-powered system that predicts which articles should be translated for cross-regional audiences. By combining natural language processing, large language models, and contextual feature engineering, our system delivers data-driven suggestions that complement editorial expertise. Our approach is hybrid, combining contextual BERT embeddings with structured editorial features.

Previous work has shown the value of predictive analytics in journalism. [4] demonstrated how content targeting increases engagement, while [3] showed how data can support strategic planning in media organizations. Building on these foundations, our work extends predictive models to a multilingual editorial context.

We also address common limitations of AI-based editorial systems. These include lack of transparency, misalignment with editorial criteria, and insufficient support for cross-cultural relevance. Our system extracts interpretable features, integrates feedback through active learning, and emphasizes alignment with real-world editorial practices.

More broadly, this work contributes to emerging efforts at integrating AI responsibly into journalistic processes. The tension between automation and editorial autonomy is central in multilingual newsrooms where translation decisions require both linguistic and contextual awareness. By offering interpretable predictions grounded in editorial reasoning, our model helps bridge the gap between algorithmic efficiency and human judgment. It supports editorial strategies that aim to improve national coherence and mutual understanding across linguistic regions.

By supporting the flow of news across linguistic regions, our system contributes to a more informed public discourse and fosters national cohesion - a clear example of how intelligent systems can serve the social good in a multilingual society.

This study makes four main contributions:

(1) A hybrid translation prediction model combining BERT embeddings with editorially grounded features, validated on 15,933 news articles.
(2) A feature importance analysis showing that source type, person-name mentions, and event characteristics are strong predictors of translation.
(3) An active learning mechanism that integrates real-time editorial feedback, improving F1 from 77.5% to 81.2% in four weeks.
(4) A newsroom deployment with editor-configurable thresholds and a live feedback loop for human-in-the-loop decision support.

While our approach builds on established NLP and recommendation techniques, its novelty lies in operationalizing a hybrid AI system within a live newsroom workflow. The integration of interpretable features, real-time editorial feedback, and configurable decision thresholds creates a practical, human-centered decision support tool not previously demonstrated in multilingual editorial settings.

2 Related Work

2.1 Predictive Models for News Selection

Recent research in predictive analytics has examined how machine learning can enhance editorial workflows. [5] explored the use of machine learning models, including AdaBoost, LPBoost, and Random Forest, to predict online news popularity based on features from the UCI dataset. [25] further advanced performance using Autoencoders with One-Class classifiers, particularly for imbalanced datasets.

Deep learning models have also shown promise in this domain. [21] achieved 96.27% accuracy using artificial neural networks and social media features. [2] used GRUs and Word2Vec embeddings to improve content representation, while [17] proposed a "robot editor" that combined ensemble models with topic modeling to forecast article performance.

Recent advances include the work of [10], who developed a modular framework for multi-aspect neural news recommendation, and [11], who proposed a prompt-based news recommendation system using pre-trained language models. [27] provided a comprehensive survey of personalized neural news recommenders focusing on accuracy and diversity challenges.

Extending this body of work, [15] introduced cross-linguistic enrichment to align semantic representations, a foundation we build upon for multilingual republication prediction.

2.2 Limitations of Traditional Recommendation Approaches

Collaborative filtering remains common in news recommendation [22], but faces limitations such as cold-start problems and poor adaptability to fast-moving news cycles [23]. These history-based systems are poorly suited for editorial judgment.

Content-based methods address some of these issues. For example, [16] detected satire using textual cues without relying on user behavior, highlighting the predictive power of linguistic signals. Yet, most existing systems are limited to single-language scenarios and do not support cross-lingual editorial decisions.

Multilingual content selection poses additional challenges that exceed standard recommendation requirements. [8] explored few-shot news recommendation via cross-lingual transfer, while [9] introduced xMIND, a multilingual dataset derived from machine-translated news articles across 14 diverse languages. [13] investigated multilingual news consumption patterns, revealing significant differences in querying and selection behaviors across language groups.

Complementary work in cross-cultural information retrieval has underscored the importance of tailoring content to linguistic and social preferences [18]. Editors must consider language use in social interaction, cultural salience, and cross-regional relevance. Our system addresses these needs by embedding such considerations directly into the predictive architecture.

2.3 Feature Evaluation and Multilingual Gaps

Prior research has applied ablation and interpretability techniques in related domains. For instance, [20] proposed a model-agnostic interpretability method. Similarly, [24] developed automated feature elimination strategies, and [1] explored CNN-based feature evaluation for parameter optimization in machine learning applications. In the social media context, [7] performed comparable evaluations for engagement forecasting.

Despite these innovations, such techniques remain underutilized in multilingual editorial support. Prior research tends to focus on virality and user preference, rather than editorial alignment across languages. Our work contributes to this understudied area by pairing ablation analysis with editorial heuristics and multilingual feature sets.

Recent research by [19] highlighted several persistent challenges and opportunities in news recommendation systems, emphasizing the need for context-aware features that can capture cultural and linguistic nuances. [11] demonstrated that large language models can significantly improve recommendation quality through better

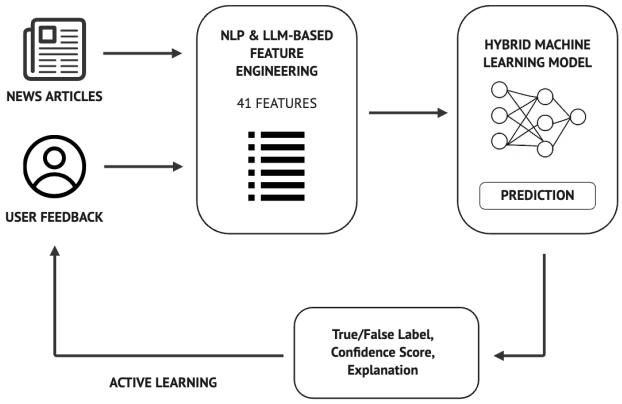

Figure 1: Five-stage predictive workflow for cross-regional article selection in multilingual newsrooms.

understanding of semantic context, particularly for cross-lingual applications.

By integrating structured features with transformer-based representations and editorial feedback, our approach bridges computational capabilities and domain-specific expertise, creating a synergy that is critical for successful newsroom integration. Having identified these research gaps, we now present our predictive workflow, detailing the dataset, feature engineering, and active learning mechanisms that underpin our system.

3 Methodology

Our predictive workflow follows five key stages: (1) article preprocessing, (2) real-time feature extraction using NLP and LLM techniques, (3) prediction using a hybrid model that combines BERT embeddings with structured features, (4) editorial feedback collection, and (5) model refinement through active learning. This cyclical structure ensures continuous alignment between system outputs and evolving editorial judgment. Figure 1 provides an overview of the end-to-end workflow, from preprocessing to editorial feedback. In the next section, Figure 2 zooms in on the core predictive component of this workflow, the hybrid model that integrates contextual BERT embeddings with structured editorial features.

3.1 Dataset and Preprocessing

We curated a dataset of 15,933 German-language news articles from a major Swiss publisher spanning 2023-2024. Each published article was labeled to indicate whether it had been translated into French ("translated") or not ("not translated"). The dataset contains 4,001 translated articles and 11,932 non-translated ones. Each record included the article's headline (avg. 9.3 words), body text (avg. 472 words), an editorial summary (35-50 words), category metadata (e.g., politics, culture), and publication timestamps. For translated articles, we also stored the corresponding French publication URLs for reference.

Preprocessing steps included normalization, tokenization, compound noun splitting (via morphological analysis), part-of-speech tagging, and named entity recognition (NER) using spaCy's German models. We also applied semantic similarity filtering to remove near-duplicates and non-parallel reports (e.g., syndicated duplicates, merged newswire content), ensuring labels correspond to deliberate translation actions rather than syndicated content. These German-specific preprocessing steps ensured that the model's inputs were clean and reliable.

3.2 Real-Time Feature Engineering

We constructed a 41-dimensional feature vector for each article. These features reflect linguistic, structural, and editorial signals that influence translation suitability. Table 1 provides a taxonomy of the engineered features, grouped into seven editorial dimensions (Regional Impact, News Quality, etc.), along with examples. The count in each category indicates the number of individual features it contains.

To provide a clearer picture of the engineered features, we summarize here how the 41 dimensions behave in practice. Regional Impact features show a bimodal distribution: approximately 23% of articles explicitly mention French-speaking regions, while 61% have only German-region mentions. Source-type features indicate that 38% of articles originate from agency feeds, while 62% are original reporting. Person-Named-Entities appear in 72% of translated articles, compared with 41% of non-translated ones, underscoring their predictive strength. Sentiment polarity is predominantly neutral overall, but emotionally charged pieces are more common among translated articles (31% vs. 12%).

3.2.1 Ontology-Based Features. To capture domain-specific semantics (e.g., political entities, public events), we developed a lightweight ontology using structured knowledge sources such as Wikidata [26]. Articles are first preprocessed:

$$D' = \text{Preprocessing}(D) \tag{1}$$

where D is the raw document corpus and D' is the cleaned, lemmatized, and Named Entity Recognition (NER)-tagged version. Tokens t_i from the ontology are matched against article d_j using:

$$\text{FeatureWeight}_i = W_i \cdot \text{match}(t_i, d_j) \tag{2}$$

Here, W_i is a weight reflecting term i's relevance, calculated via a TF-IDF based scheme with domain-specific boosting (higher weights for terms linked to cross-regional interest, e.g., federal referendums). Geographic entities are mapped to Swiss cantons, and political references are linked to government offices, votes, or public initiatives. The resulting feature weights contribute to the structured feature vector used for classification.

3.2.2 LLM-Driven Features. Editorially nuanced signals were extracted using prompt-based queries to large language models (e.g., perplexity). For a predefined set of editorial criteria $C = \{C_1, \ldots, C_k\}$, the LLM returns binary indicators $y_k \in \{0, 1\}$ for each article d_j, according to the function:

$$y_k = \text{LLM}(d_j, C_k) \tag{3}$$

Examples of criteria (each criterion corresponds to one feature per article):

Table 1: Feature taxonomy with descriptions, count, and example features per category

Feature Category	Description	Count	Example Features
Regional News & Impact	Local relevance, community significance, regional mentions	6	Regional names (French-speaking)
News Quality & Type	Article structure, source type, information density	8	Agency vs. original, emotional tone
Politics & Public Affairs	Political entities, voting patterns, civic engagement	7	Federal votes, parliament mentions
Economy & Consumption	Financial indicators, consumer trends, market references	6	Inflation rate, consumer confidence
Lifestyle & Recreation	Cultural content, entertainment topics, leisure activities	5	Food, travel, entertainment
Safety & Crime	Security incidents, risk factors, legal proceedings	4	Criminal cases, army
International Affairs	Cross-border topics, global events, international actors	5	EU politics, Foreign leaders, crises

- Does the article cover international topics with Swiss relevance?
- Is the tone emotionally charged or neutral?
- Is the central figure a person affected by violence, injustice, or discrimination?
- Is this article related to a national or French-speaking cantonal vote?

Each LLM response was post-processed using rule-based heuristics (e.g., majority voting from multiple prompt formulations) to ensure consistency. We also implemented a prompt versioning system for traceability of these features. Notably, using an LLM in this way provides real-time content analysis that adapts as articles arrive, complementing the static ontology-based features with more context-sensitive judgments (e.g., detecting emotional tone or implicit relevance cues).

3.3 Feature Optimization

To ensure efficient model performance and maintain interpretability, we applied several optimization techniques to the structured feature set. These included normalization, exploratory dimensionality reduction using PCA and t-SNE, and Recursive Feature Elimination (RFE). This process reduced redundancy and highlighted the most informative features for classification.

This feature taxonomy supported model refinement by clarifying the contribution of different editorial dimensions. Performance evaluation later demonstrated that certain categories played a disproportionately important role in prediction accuracy, informing both model tuning and interface design.

3.4 Active Learning Feedback Loop

To adapt to changing editorial judgment, we implemented a feedback loop using uncertainty sampling. Articles with prediction scores near the decision threshold are prioritized for review. Editors accept or reject suggestions through a dedicated feedback interface, and this data is reintegrated weekly to fine-tune the model.

The selection threshold is configurable and plays a central role in controlling the volume of daily suggestions. By default, a threshold of 0.6 results in approximately 50 articles being proposed per day from a pool of around 250 published German-language articles. Editors can adjust this trigger based on their editorial capacity or strategic focus. For instance, increasing the threshold to 0.7 results in more conservative, higher-confidence recommendations. While lowering it to 0.5 yields more inclusive suggestions. This flexibility lets each desk balance precision vs. recall according to their needs.

Editors approve or reject each suggestion via the feedback button; we store the article ID, decision, and timestamp. Weekly, we refresh a small exemplar bank, a set of representative example cases drawn from recent editor feedback (per desk/topic). At inference, the system retrieves the five most similar exemplars (based on embedding similarity and uncertainty cues) and injects them into the few-shot prompt that guides feature extraction and explanation. This lightweight loop improved F1 from 77.5% to 81.2% over four weeks (~820 feedback events), with most gains in the first three weeks. The F1-score reported here is computed on the held-out validation split described later in Section 5.1, ensuring consistency across weekly updates. Editors also adjust per-desk thresholds (e.g., higher for politics, lower for culture) to trade precision vs. recall without retraining.

4 Model Design and Implementation

4.1 Hybrid Model Architecture

To establish a semantic baseline, we also trained a BERT-only model that processes both the article headline and body text. This model encodes the unstructured textual content using the multilingual BERT-base-cased encoder, without incorporating any of the handcrafted editorial features. It serves as a deep contextual benchmark to assess the added value of structured signals in the hybrid architecture.

We designed a hybrid classification model that merges contextual embeddings with structured editorial signals. The architecture consists of three main components:

(1) Text Encoding Branch: Article headlines and body text are tokenized and encoded using the multilingual BERT-base-cased model, pretrained on Wikipedia and fine-tuned on our dataset. This generates a 768-dimensional contextual embedding for each article.
(2) Structured Feature Branch: Articles are simultaneously processed through NLP and LLM-based pipelines to extract a 41-dimensional handcrafted feature vector representing editorially relevant dimensions such as sentiment, person named entities, and regional relevance.
(3) Classification Head: The outputs from the two branches are concatenated and passed through a batch-normalized feedforward network with a sigmoid output layer. This component generates the final binary prediction: translate or not.

This architecture captures both latent semantic information and interpretable editorial signals. It enables flexible feature attribution

by identifying which structured variables contributed most to a prediction, an essential requirement for supporting human-in-the-loop decision-making. A schematic representation is shown in Figure 2.

4.2 Training Procedure and Hardware

We trained the hybrid model on the full dataset of 15,933 labeled articles using an 80/20 stratified train-test split. Class imbalance (approximately 25% positive) was addressed by a combination of weighted binary cross-entropy loss and mixed sampling strategies. The model was implemented in PyTorch and trained with the Adam optimizer (Adaptive Moment Estimation). We tuned hyperparameters via grid search and validated performance using 10-fold cross-validation. The grid search covered learning rate, dropout rate, hidden-layer width, batch size, and the number of dense layers in the classification head.

All experiments were run on a GPU-enabled infrastructure. In our case, we used a single NVIDIA Tesla V100 (16 GB memory) for model training. We employed a batch size of 32 and trained for up to 10 epochs, with early stopping on the validation loss to prevent overfitting. Training was manageable on a single GPU, with the hybrid model requiring several hours per epoch and simpler baselines completing within minutes.

4.3 System Implementation, Inference and Deployment

The final model was deployed as a RESTful API within a web-based editorial dashboard. We implemented the following endpoints:

- /predict: Accepts an article's content and metadata, returns a binary prediction (translate vs. not translate) along with a confidence score and a list of top contributing features.
- /feedback: Accepts an article ID and an editor's decision (translated or not), and records this feedback for learning.
- /metrics: Returns performance statistics (accuracy, recent feedback incorporation status, etc.) for monitoring.

Feature extraction is triggered in real time as new articles are ingested into the newsroom's CMS. The backend pipeline performs tokenization, structured feature computation (ontology lookup, sentiment analysis, etc.), and LLM-based queries as needed, then generates a prediction. This end-to-end inference pipeline is optimized to align with editorial review rhythms; in practice, an article's suggestion is available within seconds of publication.

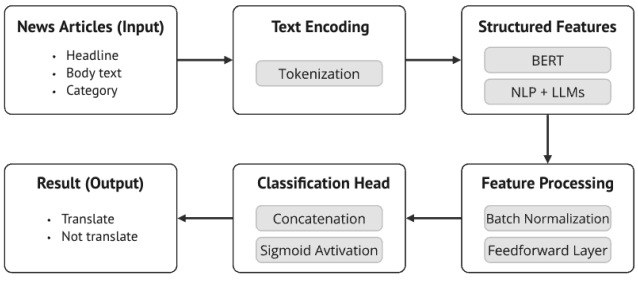

Figure 2: Hybrid architecture combining multilingual BERT with LLM/NLP-based structured features and dense classification layers

During implementation, we took care to integrate with existing editorial tools. Predictions are delivered to editors with justifications such as "high regional relevance" or "strong emotional tone," based on the top weighted features for that article. These explanations are shown in the interface next to each suggested article. Editors can click an article to see a breakdown of why it was recommended, which fosters trust and transparency. The system initially rolled out in a pilot with four editors across national and regional desks. Based on their feedback, we adjusted the confidence score visibility and explanation granularity. For instance, overly detailed rationales were condensed for clarity, and thresholds were raised for politically sensitive content to favor precision over recall (fewer borderline suggestions in that domain).

Full integration followed a staged deployment across editorial roles, with each editor focusing on a distinct content category (politics, regional affairs, culture, etc.). Onboarding sessions allowed each editor to explore the system and configure it to their preferences (e.g., setting a higher threshold if they prefer only very certain recommendations). The platform records both the system's recommendation and the editor's final action (translated or not). This logging enabled us to analyze disagreement patterns and supply "counterfactual" training examples in the active learning loop (e.g., if the model did not recommend an article that an editor chose to translate, that article becomes a valuable positive example to add in training). This real-time human-in-the-loop integration has made the system a practical enhancement to the newsroom workflow. Editors reported that the AI suggestions serve as a useful second opinion, especially under tight deadlines. The combination of integration into their dashboard, adjustable settings, and clear explanations contributed to a high adoption rate and a perception that the tool is augmenting rather than encroaching on their decision autonomy.

5 Experimental Evaluation and Results

5.1 Experimental Setup

We evaluated the performance of our prediction models using standard classification metrics and comparisons against multiple baselines. Models were trained and tested on the same 80/20 stratified split of the dataset, and we report average results across 5 random restarts to ensure robustness. The primary evaluation metrics were Accuracy, Precision, Recall, and F1-score. Accuracy is the overall fraction of correctly classified instances. Precision is the fraction of articles the model predicted "translate" that were actually translated (a measure of suggestion quality), and Recall is the fraction of actual translated articles that the model successfully identified (a measure of coverage). F1-score is the harmonic mean of precision and recall, summarizing the balance between the two.

We also measured inter-annotator agreement to understand the consistency of editorial judgments. Two senior editors independently reviewed a randomly selected subset of 200 German articles and indicated whether they would translate each for French audiences. We found that the editors agreed on 170 out of 200 decisions (85% raw agreement), with a Cohen's κ of 0.68. This substantial

Table 2: Performance comparison of models across evaluation metrics

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	69.2%	58.3%	67.1%	62.4%
Random Forest	71.8%	64.5%	60.9%	62.7%
Gradient Boosting	75.3%	68.2%	64.7%	66.4%
Neural Network	82.4%	75.9%	74.1%	75.0%
BERT-only	61.0%	55.2%	56.3%	55.7%
Hybrid (BERT+Features)	**85.0%**	**78.7%**	**76.3%**	**77.5%**

agreement, while not perfect, quantifies the inherent subjectivity in the task. It sets an approximate upper bound on prediction performance, since even human experts differ in some cases. It also underscores the importance of our configurable threshold: what one editor might consider a "translate" candidate, another might not, so the system must allow tuning to individual or role-specific standards.

5.2 Comparative Results and Baselines

We trained five baseline models for comparison: Logistic Regression, Random Forest, Gradient Boosting Machine, a simple Neural Network (multi-layer perceptron), and the BERT-only text model as shown in Table 2. The logistic, tree-based, and neural network baselines used only the 41 engineered features as input (simulating a scenario without the article text embeddings), whereas the BERT-only baseline used only the article text (headline and body) without our structured features. All models were evaluated on the same test set. The hybrid model that combines BERT embeddings with the 41 features is our proposed approach. We did not include an LLM-only baseline for two reasons. First, prompt-based zero-shot classification with API-based LLMs (e.g. GPT models) is difficult to reproduce, as model behavior varies with prompt phrasing, hidden system settings, and undocumented model updates. Such variability makes it an unstable scientific baseline. Second, translation decisions rely heavily on editorial signals, including regional relevance, source type, and specific actor categories, which are not reliably inferable from text alone and are better captured by structured features. For these reasons, we focus on reproducible local baselines and on evaluating how LLM-derived features complement traditional signals rather than on end-to-end LLM classification.

The feature-only models performed respectably, with Gradient Boosting achieving 75.3% accuracy (F1 66.4%) and the simple Neural Network reaching 82.4% accuracy (F1 75.0%). The BERT-only baseline, in contrast, reached only 61.0% accuracy (F1 55.7%), indicating that the raw multilingual BERT struggled to learn the task from text alone, likely due to class imbalance and the subtlety of the cues. This result is consistent with the nature of the prediction task. Translation decisions often depend on explicit editorial cues, such as whether a region is French-speaking, whether an article contains specific political actors, or whether the source is agency-generated, none of which are reliably encoded in raw text embeddings. Many such signals appear only in metadata or external context and cannot be inferred purely from text embeddings. The multilingual BERT encoder also struggles with highly domain-specific entity types (Swiss cantons, government offices), and class imbalance pushes it toward conservative predictions that favor the majority class. Indeed, the BERT-only model frequently defaulted to predicting "not translate," which explains its low recall. In contrast, the structured feature branch encodes these signals directly. The strong performance of feature-only models thus confirms that translation prediction is less about deep semantic nuance and more about editorial criteria that are difficult for a text-only model to learn. The hybrid model benefits from the complementary strengths of both modalities, combining semantic context from BERT with explicit editorial cues.

Our hybrid model achieved 85.0% accuracy, with 78.7% precision, 76.3% recall, and an F1-score of 77.5%. This was the highest overall performance among all models. Notably, combining BERT embeddings with structured features increased F1 from 55.7% to 77.5% and accuracy from 61% to 85% compared to the BERT-only baseline. The hybrid model also outperformed the best feature-only model (Neural Network) by about 2.6 points in F1 and 2.6 percentage points in accuracy, demonstrating that the textual semantic features and the editorial features provide complementary information.

The model achieved its highest accuracy on political and economic articles reaching, approximately 89% accuracy, likely due to their consistent structure, formal tone, and recurring named entities (e.g., government figures, economic indicators). In contrast, sports and culture articles, characterized by more varied language and local context, showed slightly lower accuracy of approximately 80%, while still outperforming all baselines. We also evaluated performance by the newsroom's content categories (Domestic, Politics, Economy, Culture, Crime/Society, Sports, Digital, International). The hybrid model's precision was highest on Politics and Economy (both over 87% precision), while categories like Culture and Sports achieved higher recall but slightly lower precision. This pattern suggests that editorially subjective or feature-rich topics, such as culture and sports, are more challenging. In these cases, the model tends to capture most true positives (high recall) at the expense of a few false positives (lower precision). In contrast, fact-heavy domains with clearer signals (politics, finance) allow the model to be more precise. These observations informed the selection of category-specific thresholds during deployment: for domains where false positives are more tolerable than missed stories (e.g. culture, where editors prefer to see all possibly interesting pieces), we chose thresholds favoring recall; whereas for domains where a flood of suggestions would overwhelm editors (e.g. politics), we tuned for higher precision.

5.3 Feature Importance and Ablation Study

To identify the most predictive features in our hybrid model, we conducted a feature importance analysis and ablation study. First,

we trained a Random Forest classifier using only the 41 structured features and computed Gini importance scores for each feature. Figure 3 plots each feature's importance. The top features aligned well with editorial intuition: Sentiment polarity was highly predictive (articles with a strong emotional tone were more often translated, especially human-interest pieces), as were counts of person named entities (articles mentioning prominent people had higher translation likelihood) and regional relevance indicators (articles explicitly about French-speaking regions or national stories were more likely to be translated). Additionally, whether an article came from a news agency feed versus original reporting proved important: agency-sourced articles were less likely to be selected for translation, presumably because those stories might already be available in French from the wire service or be less region-specific. These observations were confirmed by our ablation test: removing Person Named Entities features caused a 4.7-point drop in F1, removing Regional Relevance caused a 3.9-point drop, and removing Sentiment Polarity caused a 3.4-point drop (Table 3). Removing all three top feature categories together diminished F1 by about 12 points. This shows that each contributes unique information not fully captured by the others or by the BERT text embeddings. In summary, explicit editorial cues (people, places, tone, source type) greatly enhance the model's decision-making, which aligns with how editors described their own process.

Table 3: Top Feature Categories by F1 Drop in Ablation Study

Feature Category	F1 Drop (Percentage Points)
Person Named Entities	−4.7
Regional Relevance	−3.9
Sentiment Polarity	−3.4

5.4 Active Learning Performance and Editor Alignment

During the four-week newsroom deployment, the active learning loop steadily improved model performance. Specifically, the model's F1-score on validation data rose from 77.5% to 81.2% after incorporating approximately 820 feedback events through weekly exemplar-bank refreshes. Most gains occurred in the first 2-3 weekly updates, after which improvements leveled off, suggesting diminishing returns as the model converged toward the editors' decision boundary. We also noted that the feedback particularly boosted recall in under-represented areas: for example, initially the model missed some niche cultural articles, but after editors explicitly approved a few, the model learned to broaden its criteria. Precision also improved slightly as the model learned to avoid types of articles editors consistently rejected (like very local news or routine sports recaps).

Editor trust and uptake grew over this period. We conducted a brief internal survey at the end of the trial: 80% of participating editors agreed that the recommendations aligned well with their own expectations (with many commenting that the suggestions reinforced ideas they were considering anyway), and 75% expressed interest in expanding the tool to additional language pairs beyond German-French. This positive feedback underscores that transparency, adaptability, and editorial control were key to the system's acceptance.

Finally, as noted above in the Experimental Setup, we quantified the agreement between different editors to contextualize the model's performance. The Cohen's κ of 0.68 among editors indicates that even humans have some disagreements in borderline cases. Our model, operating with an F1 in the high 70s to low 80s, is approaching the consistency of an editorial team member. We find this a promising result: it suggests the model can be seen as a reasonable "second opinion." Cases where the model and an editor disagree often correspond to cases where editors themselves might disagree - these are the very cases where providing a data-driven perspective can be most valuable, as it prompts a closer look. Going forward, we aim to further analyze these disagreement cases to refine both the model and editorial guidelines (for example, clarifying criteria for translating sports human-interest stories, which were a frequent source of differing opinions).

5.5 Error Analysis

To better understand the system's limitations, we examined misclassified articles in the validation set. False positives often involved culturally nuanced stories that editors considered too local or stylistically unsuitable for translation despite strong semantic relevance. False negatives typically included niche human-interest stories that lacked explicit regional or political cues but were still considered valuable by editors. These cases highlight the role of subtle contextual knowledge such as tone, narrative framing, and institutional familiarity, which remain difficult to encode through features alone. Incorporating discourse-level features, richer entity linking, or lightweight summarization may help address these errors in future iterations.

6 Conclusion and Discussion

Our experimental evaluation showed that the hybrid model significantly outperformed baselines, with accuracy reaching 85% and F1 improving to 81.2% after active learning. These results confirm the value of combining BERT embeddings with structured editorial features. Although the gains are incremental rather than disruptive, they are meaningful for newsroom practice and demonstrate the value of tailored hybrid architectures over generic LLM or text-only models. Building on these findings, we now discuss their implications for newsroom practice and future research.

We addressed the challenge of predicting which German-language news articles should be translated for French-speaking audiences in a multilingual newsroom. We proposed a hybrid AI model that integrates BERT-based semantic embeddings with structured editorial features, reinforced by real-time feedback from editors through an active learning loop. Trained on a large, real-world dataset of 15,933 German-language news articles, the system achieved 85.0% accuracy and a 77.5% F1-score, significantly outperforming both traditional baselines and a text-only BERT model. Performance varied by topic: structured domains like politics and economics yielded higher accuracy, while culturally nuanced categories like sports or lifestyle were slightly lower, reflecting the varying complexity of editorial decisions across genres.

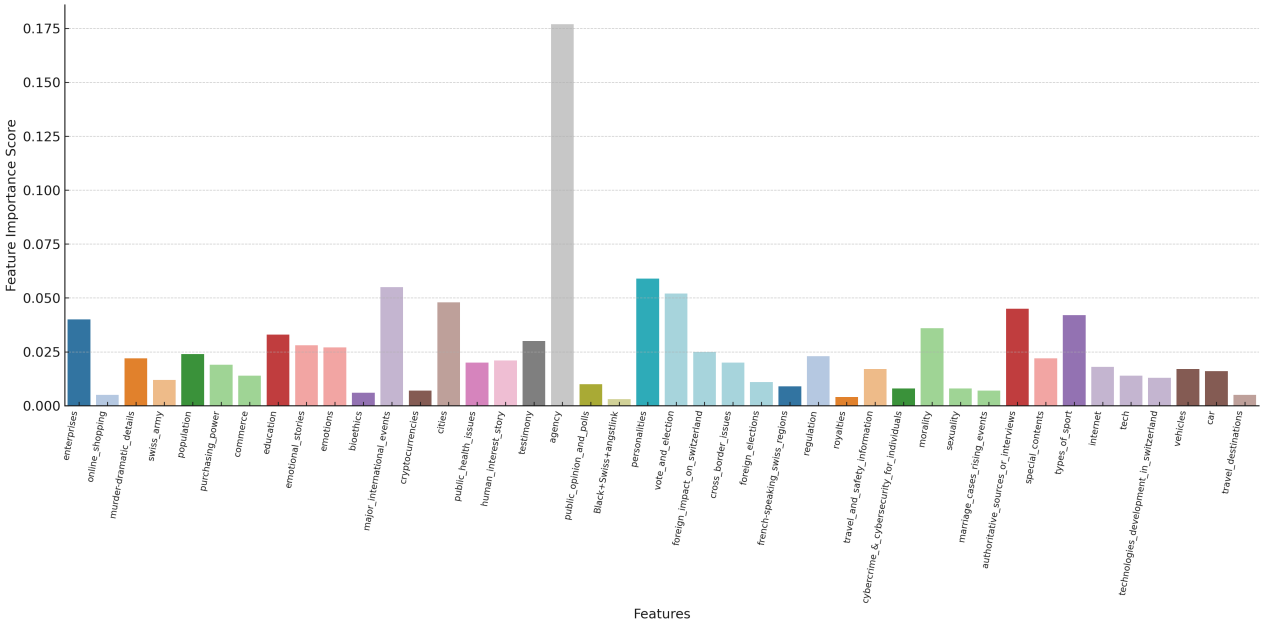

Figure 3: Feature importance scores based on Gini impurity (Random Forest classifier)

Our ablation studies confirmed the importance of sentiment polarity, person-name mentions, and regional relevance as core signals. The integration of editorial feedback via uncertainty sampling led to measurable improvements over four weeks of deployment, enhancing both model alignment with editorial choices and the editors' trust in the system.

From a theoretical standpoint, our findings contribute to the literature on human-AI collaboration in editorial settings under multilingual constraints. The importance of emotionally resonant and context-aware content reinforces that translation decisions involve more than factual news value, as they also encompass cultural salience and perceived audience relevance. Practically, this work demonstrates that hybrid models can augment editorial workflows by providing interpretable predictions aligned with professional heuristics. Rather than replacing editorial decision-making, the system supports it by improving consistency, efficiency, and coverage diversity while remaining responsive to feedback. In contrast to prior "black-box" recommenders, our system's integration of an editor-configurable threshold and live feedback loop is a unique innovation that empowers editors to guide the AI's behavior. Editors could directly influence which stories get flagged by tuning a simple slider and by giving a thumbs-up/down on suggestions, a level of operational control that sets our approach apart from earlier newsroom AI tools.

6.1 Limitations

Our study has five main limitations. First, niche or hyper-local stories are underrepresented, which reduces confidence in rare cases. Second, the target label ("translated" vs. "not translated") reflects editorial preference rather than intrinsic article value. Translation decisions depend on workflow constraints, desk priorities, perceived audience interest, and timing. As shown by our inter-annotator study (κ = 0.68), even senior editors disagree on borderline cases, indicating that the task contains inherent subjectivity. This limits the upper bound of predictive performance and should be considered when interpreting results. Third, the results are specific to the German-French context, and transferring to other language pairs will require light adaptation (language resources, region/entity maps, and threshold tuning). Fourth, we do not explicitly model temporal drift, so rapid shifts (elections, crises) may outpace weekly exemplar refreshes. Fifth, because the data and models come from a single publisher, reproducibility is constrained, although the method is portable and can be rebuilt on public multilingual sources. To support independent replication, we provide the full feature taxonomy, model configuration, and evaluation scripts so that researchers can apply the pipeline to publicly available multilingual corpora. Because both the dataset and the production codebase are under strict licensing agreements with our media partner, they cannot be released. However, we will provide documentation outlining the feature extraction process, prompt templates, and full model configuration to facilitate re-implementation.

6.2 Ethical and Editorial Considerations

Although the system is designed to support editors rather than automate translation decisions, several ethical considerations arise when deploying AI tools in newsrooms. First, editorial autonomy must be preserved: our system provides suggestions rather than decisions, and editors retain full control over final translation choices. Second, AI models risk amplifying existing imbalances in historical data. For example, over-represented regions or topics may be suggested more often unless feedback or threshold adjustments correct these trends. To mitigate this, our deployment included

per-desk threshold tuning and exemplar-bank diversification allow editors to steer the model toward more balanced coverage. Third, transparency is essential for maintaining trust among newsroom staff. By surfacing top contributing features and enabling editors to inspect the rationale behind each suggestion, the system supports contestability and accountability. These considerations illustrate that integrating AI responsibly into editorial workflows requires not only strong model performance but also respect for professional judgment, cultural nuance, and institutional values.

6.3 Future Work

Future work will add multimodal cues (images/layout) to better capture cross-regional salience, and introduce time awareness via drift monitoring and time-weighted exemplar refresh. We will assess cross-lingual transfer by starting zero-shot with multilingual embeddings, then adapting with few-shot prompting from a small target-language exemplar bank, measuring gains versus exemplar size. We also plan editor-centric explainability that highlights rationale spans, and supports desk-level "don't show this" controls that update the exemplar bank. Finally, we aim to replicate the pipeline on public multilingual corpora and release prompt templates, the feature taxonomy, and evaluation scripts.

In summary, our work shows that a hybrid AI system can meaningfully support editorial judgment in multilingual settings. By leveraging both data-driven insights and human expertise, and by building adaptability into the tool, we achieved a workflow integration that editors found valuable. We believe that through responsible and thoughtful expansion, including support for additional languages, incorporation of multimedia content, and maintenance of a human-centered design, such systems can enhance journalistic collaboration across language barriers and provide newsrooms with powerful new capabilities for cross-cultural news exchange.

By situating our work within the broader digital transformation of media, we demonstrate how intelligent decision support systems can enhance newsroom practices today while providing a template for future cross-domain applications in today's digital era, from multilingual publishing to other sectors where human-centered AI decision support is critical. Beyond technical performance, the societal benefit of this approach lies in enabling more equitable access to information across language communities, thereby strengthening democratic dialogue and inclusiveness in today's digital society.

Acknowledgments

This work was supported by the Initiative for Media Innovation (IMI) under grant number 127384. We thank our media partner for providing data access and editorial feedback during the project.

References

[1] Linlin Bie, Xu Wang, and Jari Korhonen. 2019. Optimizing the parameters for post-processing consumer photos via machine learning. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 1504–1509.

[2] Yan Cai and Zhiqiang Zheng. 2022. Prediction of news popularity based on deep neural network. *Scientific programming*, 2022, 1, 8280036.

[3] Hyunyoung Choi and Hal Varian. 2012. Predicting the present with google trends. *Economic record*, 88, 2–9.

[4] Deborah S Chung. 2008. Interactive features of online newspapers: identifying patterns and predicting use of engaged readers. *Journal of Computer-Mediated Communication*, 13, 3, 658–679.

[5] Dhanashree Deshpande. 2017. Prediction & evaluation of online news popularity using machine intelligence. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. IEEE, 1–6.

[6] Nicholas Diakopoulos. 2019. *Automating the news: How algorithms are rewriting the media*. Harvard University Press.

[7] Keyan Ding, Ronggang Wang, and Shiqi Wang. 2019. Social media popularity prediction: a multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2682–2686.

[8] Taicheng Guo, Lu Yu, Basem Shihada, and Xiangliang Zhang. 2023. Few-shot news recommendation via cross-lingual transfer. In *Proceedings of the ACM web conference 2023*, 1130–1140.

[9] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2024. Mind your language: a multilingual dataset for cross-lingual news recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 553–563.

[10] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2023. Train once, use flexibly: a modular framework for multi-aspect neural news recommendation. *arXiv preprint arXiv:2307.16089*.

[11] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. Pbnr: prompt-based news recommender system. *arXiv preprint arXiv:2304.07862*.

[12] Tommy Carl-Gustav Linden. 2017. Algorithms for journalism: the future of news work. *The journal of media innovations*, 4, 1, 60–76.

[13] Chenjun Ling, Ben Steichen, and Silvia Figueira. 2020. Multilingual news–an investigation of consumption, querying, and search result selection behaviors. *International Journal of Human–Computer Interaction*, 36, 6, 516–535.

[14] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, 31–40.

[15] Zhan Liu and Nicole Glassey Balet. 2023. Adaptive semantic matching in a multilingual context. *International journal of semantic computing*, 17, 03, 435–453.

[16] Zhan Liu, Shaban Shabani, Nicole Glassey Balet, and Maria Sokhn. 2019. Detection of satiric news on social media: analysis of the phenomenon with a french dataset. In *2019 28th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 1–6.

[17] Fei Long, Meixia Xu, Yulei Li, Zhihua Wu, and Qiang Ling. 2018. Xiaoa: a robot editor for popularity prediction of online news based on ensemble learning. In *Intelligence Science II: Third IFIP TC 12 International Conference, ICIS 2018, Beijing, China, November 2-5, 2018, Proceedings 2*. Springer, 340–350.

[18] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: a survey. *Computational linguistics*, 42, 3, 537–593.

[19] Shaina Raza and Chen Ding. 2022. News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review*, 1–52.

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

[21] Kawser Irom Rushee, Enan Ajmain, Md Moshavvir Hossain Captain, Md Talha Jubayer, and Wahidul Islam. 2022. Prediction of online news popularity using ann deep learning. In *Proceedings of the 2nd International Conference on Computing Advancements*, 415–419.

[22] KG Saranya and G Sudha Sadasivam. 2017. Personalized news article recommendation with novelty using collaborative filtering based rough set theory. *Mobile Networks and Applications*, 22, 719–729.

[23] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 253–260.

[24] Sina Sheikholeslami, Moritz Meister, Tianze Wang, Amir H Payberah, Vladimir Vlassov, and Jim Dowling. 2021. Autoablation: automated parallel ablation studies for deep learning. In *Proceedings of the 1st Workshop on Machine Learning and Systems*, 55–61.

[25] Min-Jen Tsai and You-Qing Wu. 2022. Predicting online news popularity based on machine learning. *Computers and Electrical Engineering*, 102, 108198.

[26] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57, 10, 78–85.

[27] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized news recommendation: methods and challenges. *ACM Transactions on Information Systems*, 41, 1, 1–50.