PeerJ Computer Science

Multilingual sentiment analysis of summarized texts: a cross-language study of text shortening effects

Mikhail Krasitskii[1], Grigori Sidorov[1], Olga Kolesnikova[1], Liliana Chanona-Hernandez[2] and Alexander Gelbukh[1]

[1] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Ciudad de México, CDMX, Mexico
[2] Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME), Instituto Politécnico Nacional (IPN), Ciudad de México, CDMX, Mexico

ABSTRACT

Sentiment analysis in multilingual contexts is significantly influenced by text summarization, with the effects varying across languages of different morphological complexities. This study examines the impact of extractive and abstractive summarization techniques on sentiment classification in eight typologically diverse languages: English (analytic), German (fusional), French, Spanish, Italian (moderately synthetic), Finnish and Hungarian (agglutinative), and Arabic (root-based inflectional). Multilingual transformer models (Multilingual Bidirectional Encoder Representations from Transformers (mBERT), Cross-lingual Language Model-Robustly optimized BERT approach (XLM-RoBERTa), T5, Bidirectional and Auto-Regressive Transformer (BART)) and language-specific adaptations (FinBERT, AraBERT) were employed to evaluate performance. It was found that extractive summarization better preserves sentiment accuracy, with modest declines of 2.6% in English (from 92.1% to 88.5%) and 5.2% in German compared to baseline results. In contrast, morphologically complex languages such as Finnish and Arabic exhibited more substantial drops of 8–12%. Abstractive summarization, while enhancing readability (ROUGE-1: 87.1% for English *vs* 89.3% for extractive methods), introduced significant sentiment distortion in agglutinative and inflectional languages due to semantic drift, including the omission of critical markers like Finnish case endings. A hybrid approach, combining extractive sentiment preservation through Term Frequency-Inverse Document Frequency (TF-IDF) keyphrase retention with abstractive fluency, was proposed and demonstrated superior performance. This method achieved a 4.2% higher F1-score compared to pure abstractive summarization in Arabic. The findings underscore the necessity of language-specific adaptations, particularly for applications such as social media monitoring and multilingual natural language processing (NLP) systems, where morphological complexity directly affects sentiment preservation accuracy.

Subjects Computational Linguistics, Computer Education, Natural Language and Speech, Sentiment Analysis, Neural Networks
Keywords Summarization, Sentiment analysis, Multilingual, Transformer

Submitted 9 April 2025
Accepted 28 October 2025
Published 20 February 2026

Corresponding author
Olga Kolesnikova,
kolesnikova@cic.ipn.mx

Academic editor
Giovanni Angiulli

Additional Information and Declarations can be found on page 24

DOI **10.7717/peerj-cs.3406**

How to cite this article Krasitskii M, Sidorov G, Kolesnikova O, Chanona-Hernandez L, Gelbukh A. 2026. Multilingual sentiment analysis of summarized texts: a cross-language study of text shortening effects. **PeerJ Comput. Sci. 12:e3406 DOI 10.7717/peerj-cs.3406**

INTRODUCTION

The rapid proliferation of multilingual digital content across social media, customer reviews, and news platforms has intensified the need for scalable and accurate natural language processing (NLP) systems that can extract sentiment from large volumes of text. Recent studies by *Akhtar et al. (2021)* have highlighted the growing challenges in cross-lingual sentiment analysis, particularly when dealing with morphologically diverse languages. In this context, text summarization has become a critical preprocessing step, enabling efficient information retrieval, content moderation, and opinion mining by condensing lengthy documents into concise representations. Automatic summarization techniques have evolved significantly, as demonstrated by *Nenkova & McKeown (2012)*, with newer approaches leveraging large language models (*Brown et al., 2020*; *Radford et al., 2019*). However, while summarization enhances readability and reduces computational load, it may inadvertently alter or distort the affective content of the original text, leading to potential misclassification in downstream sentiment analysis tasks (*Patwa, Bhardwaj & Chakraborty, 2023*; *Krasitskii et al., 2025*). This issue is particularly salient in multilingual environments, where linguistic diversity amplifies the complexity of preserving sentiment during text shortening.

Text summarization is broadly categorized into two paradigms: extractive and abstractive methods. Extractive summarization identifies and retains the most salient sentences or phrases from the original text, thereby preserving lexical and syntactic fidelity (*Liu et al., 2020*). In contrast, abstractive approaches generate new text through paraphrasing and linguistic rephrasing, often resulting in more fluent and human-like summaries (*Cohan et al., 2022*). Modern abstractive models, particularly those based on transformer architectures such as T5 and Bidirectional and Auto-Regressive Transformer (BART), have demonstrated impressive fluency and coherence (*Tay et al., 2022*). However, these models are prone to semantic drift, a phenomenon where the factual content remains accurate, but the emotional tone shifts, potentially leading to polarity misclassification in sentiment analysis (*Li et al., 2021*).

A growing body of research indicates that the impact of summarization on sentiment preservation is not uniform across languages. Studies have shown that morphologically simple languages such as English and Spanish experience relatively minor sentiment distortion during text shortening (*Koehn et al., 2022*). In contrast, languages with complex morphological systems, such as Finnish, Hungarian, and Arabic, exhibit greater vulnerability to sentiment loss, especially when inflectional markers, case endings, or root-based derivations are altered or omitted (*Tay et al., 2022*; *Zhang, Li & Wang, 2023*). This discrepancy arises because sentiment in agglutinative and fusional languages is often encoded in morphological features rather than isolated lexical items, making it more susceptible to disruption during compression. Furthermore, syntactic flexibility and free word order in languages like Finnish and Hungarian can lead to structural reformulation that shifts emphasis and alters perceived sentiment (*Rahman, Kumar & Li, 2023*).

Despite these challenges, current evaluation frameworks for summarization remain largely focused on surface-level metrics such as Recall-Oriented Understudy for Gisting

Evaluation (ROUGE) and Bilingual Evaluation Understudy (BLEU), which emphasize n-gram overlap and factual consistency but offer little insight into emotional fidelity (*Lin, 2004*). As a result, summaries may score highly on fluency and informativeness while failing to retain the original sentiment. Recent efforts have proposed sentiment-aware evaluation metrics, but their integration into mainstream summarization pipelines remains limited (*Gupta, Joshi & Kar, 2023*). Additionally, the reliance on multilingual models such as Multilingual Bidirectional Encoder Representations from Transformers (mBERT) and Cross-lingual Language Model-Robustly optimized BERT approach (XLM-RoBERTa), while enabling cross-lingual transfer, often overlooks language-specific nuances. Empirical evidence shows that domain-adapted and language-specific models such as FinBERT for Finnish and AraBERT for Arabic achieve superior performance in sentiment preservation, underscoring the importance of tailored approaches (*Virtanen et al., 2019*; *Antoun, Baly & Hajj, 2020*; *Gupta, Kumar & Singh, 2024*).

To address these gaps, this study presents a systematic cross-linguistic analysis of how text summarization affects sentiment classification accuracy across eight typologically diverse languages: English, German, French, Spanish, Italian, Finnish, Hungarian, and Arabic. These languages were selected to represent a broad spectrum of morphological and syntactic structures from isolating and fusional to agglutinative and root-based systems, enabling a robust comparative evaluation. We evaluate both extractive (TF-IDF, TextRank) and abstractive (T5, BART) summarization techniques using large-scale, balanced datasets including IMDb, GermEval, SemEval, the Finnish Sentiment Dataset, the Hungarian Emotion *Corpus*, and ArSAS. Sentiment analysis is performed using multilingual transformers (mBERT, XLM-RoBERTa) and language-specific models (FinBERT, AraBERT), allowing for a nuanced comparison of performance across linguistic families.

Our research is guided by two central questions:

- **RQ1:** How does text summarization affect sentiment classification accuracy in languages with different morphological complexities?
- **RQ2:** Which summarization approach, extractive, abstractive, or hybrid, better preserves sentiment polarity and nuance across languages with varying syntactic and morphological structures?

The results demonstrate that summarization significantly impacts sentiment analysis, with the degree of distortion strongly correlated with morphological richness. Extractive methods consistently outperform abstractive ones in sentiment preservation, particularly in morphologically complex languages. For instance, while English experiences a modest accuracy drop of 3.6% under abstractive summarization, Finnish and Arabic exhibit losses of 8–12%, highlighting the heightened risk in morphologically dense systems. A hybrid approach combining extractive keyphrase retention (*via* TF-IDF) with abstractive fluency achieves a +4.2% F1-score improvement in Arabic, outperforming pure abstractive methods.

The contributions of this work are fivefold:

- We present a comprehensive cross-linguistic evaluation of sentiment preservation across eight typologically diverse languages, demonstrating that morphological complexity is a key determinant of summarization-induced sentiment distortion.
- We compare extractive and abstractive summarization techniques using both general and language-specific models, showing that hybrid strategies combining local sentence selection with controlled paraphrasing offer a superior balance between fluency and emotional fidelity (*Singh, Gupta & Kumar, 2023*; *Scao et al., 2023*).
- We employ rigorous evaluation protocols, including precision, recall, F1-score analysis, and ROUGE-based fluency metrics, to ensure methodological transparency and reproducibility.
- We highlight the importance of language-specific models and culturally adapted datasets, particularly for low-resource languages such as Hungarian and Arabic dialects, where tailored data curation significantly improves both summarization quality and sentiment accuracy (*Singh, Gupta & Kumar, 2023*).
- We propose actionable recommendations for real-world deployment, including the integration of hybrid summarization in social media monitoring, customer feedback systems, and multilingual opinion mining, with a focus on preserving affective content across cultural and linguistic boundaries (*Chen, Zhang & Wang, 2024*).

By systematically analyzing the interplay between text shortening and sentiment analysis across multiple languages, this study advances the development of more reliable and culturally sensitive NLP pipelines. The insights provided here are intended to guide future research and practical system design in multilingual environments, where accurate sentiment transmission is essential for ethical, transparent, and effective communication.

RELATED WORK

Text summarization has become a foundational component of modern NLP, enabling efficient information access in domains ranging from news aggregation and academic literature review to social media monitoring and customer feedback analysis. As digital content grows in both volume and linguistic diversity, the need for automated systems that condense text while preserving its semantic and affective content has intensified. Two dominant paradigms define the field: extractive and abstractive summarization. Extractive methods identify and retain the most salient sentences or phrases from the original text, preserving factual accuracy and linguistic fidelity (*Liu et al., 2020*). These approaches rely on statistical or graph-based models such as TextRank and LexRank to compute sentence centrality based on inter-sentence similarity (*Mihalcea & Tarau, 2004*). While extractive summarization ensures high informational completeness and structural consistency, it often produces outputs that are redundant or lack narrative coherence, particularly in longer documents (*Li et al., 2021*).

In contrast, abstractive summarization generates new text by paraphrasing and rephrasing content, often resulting in more fluent and human-like summaries. The rise of transformer-based architectures such as T5, BART, and their multilingual variants has significantly advanced the state of the art in this domain (*Raffel et al., 2020*; *Tay et al., 2022*). These models employ encoder-decoder structures with self-attention mechanisms to produce contextually appropriate and syntactically coherent summaries (*Vaswani et al., 2017*). However, despite their linguistic fluency, abstractive systems are prone to semantic drift, a phenomenon where the factual content remains accurate, but the emotional tone shifts, potentially leading to misclassification in downstream sentiment analysis tasks (*Li et al., 2021*). This issue is particularly critical in affect-sensitive applications such as political discourse monitoring, customer service automation, and crisis response systems, where accurate sentiment transmission is essential.

Recent studies have begun to investigate how summarization impacts downstream NLP tasks, with growing attention to sentiment preservation. *Gupta, Joshi & Kar (2023)* conducted a cross-lingual evaluation of extractive summarization and found that while surface-level sentiment is often retained, subtle emotional cues such as intensifiers, negations, and context-dependent modifiers are frequently lost during compression. Similarly, *Koehn et al. (2022)* demonstrated that neural machine translation systems, when applied to low-resource languages, often distort sentiment polarity, highlighting the vulnerability of morphologically complex systems to semantic compression. These findings underscore the need for evaluation frameworks that go beyond factual accuracy and assess emotional fidelity.

The challenge of sentiment preservation becomes even more pronounced in multilingual contexts, where languages vary significantly in morphological richness, syntactic flexibility, and cultural norms of emotional expression. For instance, agglutinative languages like Finnish and Hungarian encode sentiment through suffixation and case markers, which may be inadvertently altered or omitted during abstractive rewriting (*Virtanen et al., 2019*). In such systems, a single word can carry multiple grammatical and emotional meanings, and its alteration during summarization can lead to polarity shifts. Similarly, in Arabic, root-based derivations and inflectional patterns play a crucial role in sentiment expression. The presence or absence of diacritics, prefixes, or suffixes can completely change the emotional valence of a word (*Antoun, Baly & Hajj, 2020*). For morphologically rich languages like Arabic, *Al-Omari, Al-Momani & Al-Kabi (2023)* found that summarization disproportionately affects sentiment-bearing morphemes, while *Zhou et al. (2020)* proposed metrics to quantify sentiment consistency in summarized texts. These linguistic differences underscore the limitations of one-size-fits-all summarization models and emphasize the necessity of language-specific adaptations.

A critical barrier to reliable sentiment-aware summarization lies in the inherent risks introduced by current methods. Based on empirical evidence and prior research, the following key challenges have been identified:

- **Loss of Contextual Cues:** Abstractive summarization frequently omits sentiment-bearing words, morphemes, or particles essential for accurate sentiment interpretation. For example, in Finnish and Arabic, morphological suffixes and intensifiers are often removed during compression, leading to a weakened polarity signal (*Krasitskii et al., 2024*; *Gupta, Joshi & Kar, 2023*).

- **Structural Reformulation:** Sentence restructuring alters syntactic configurations, which can significantly affect sentiment expression in languages with flexible word order. In Hungarian and Finnish, reordering subject-object-verb structures during paraphrasing may shift emphasis and distort perceived emotion, even when factual content remains unchanged (*Rahman, Kumar & Li, 2023*).
- **Semantic Drift:** While factual information may be preserved, the emotional tone can shift unintentionally due to rewording. This phenomenon, termed semantic drift, leads to polarity misclassification, especially in morphologically rich languages where inflectional markers carry affective weight (*Li et al., 2021*).

To address these challenges, researchers have developed specialized models trained on monolingual or domain-specific corpora. FinBERT, a BERT-based model tailored for Finnish, incorporates morphological awareness through subword tokenization and suffix-sensitive embeddings, significantly improving performance in sentiment analysis and text classification tasks (*Virtanen et al., 2019*). Similarly, AraBERT has been developed for Arabic, leveraging dialectal variation and root normalization to enhance contextual understanding and sentiment sensitivity (*Antoun, Baly & Hajj, 2020*). These models demonstrate that domain- and language-adaptive pretraining not only improves linguistic accuracy but also enhances the preservation of affective content during text processing.

Moreover, traditional evaluation metrics used in summarization, such as ROUGE and BLEU, focus primarily on n-gram overlap and factual consistency, offering little insight into emotional accuracy (*Lin, 2004*). As a result, summaries may score highly on fluency and informativeness while failing to retain the original sentiment. Recent efforts have proposed sentiment-aware evaluation frameworks, including polarity consistency checks, emotion alignment scores, and culturally sensitive metrics (*Wang, Zhang & Li, 2023*; *Gupta, Joshi & Kar, 2023*). For example, *Wang, Zhang & Li (2023)* introduced a metric that integrates sentiment lexicons and cultural context to assess affective fidelity across languages. However, the integration of such metrics into mainstream summarization pipelines remains limited, and their adoption in large-scale multilingual evaluations is still nascent.

Hybrid approaches that combine extractive and abstractive techniques have shown promise in balancing fluency with semantic and emotional stability. *Singh, Gupta & Kumar (2023)* proposed a framework that first extracts key sentiment-bearing phrases using TF-IDF and attention scoring, followed by controlled abstractive rephrasing to enhance readability. *Scao et al. (2023)* demonstrated that fine-tuning summarization models on sentiment-labeled corpora improves alignment between semantic compression and emotional content. These findings suggest that task-specific adaptation, particularly for sentiment-sensitive applications, can significantly enhance the reliability of summarized outputs.

Another emerging trend involves the use of next-generation large language models (LLMs) such as BLOOM and LLaMA 2, which offer open-access, multilingual capabilities at scale (*Scao et al., 2023*; *Touvron et al., 2023*). These models, trained on diverse linguistic corpora, show potential in preserving sentiment across languages due to their broad cross-lingual knowledge. However, challenges remain in controlling their output for sentiment consistency, and concerns about hallucination and cultural bias persist (*Chen, Zhang & Wang, 2024*). Evaluating these models in the context of sentiment-preserving summarization represents a promising direction for future research.

Despite these advances, a critical gap remains: most existing studies focus on high-resource languages like English, with limited cross-linguistic validation. Few works systematically compare extractive and abstractive methods in terms of sentiment preservation across morphologically diverse language families. Furthermore, evaluations often neglect low-resource languages and dialectal variants, where data scarcity and linguistic complexity compound the challenges of affective content preservation.

To the best of our knowledge, this is the first cross-linguistic study to systematically evaluate the impact of both extractive and abstractive summarization on sentiment preservation across such a diverse set of languages, including morphologically complex agglutinative (Finnish, Hungarian) and root-based inflectional (Arabic) systems alongside major European languages. No prior work has simultaneously assessed sentiment distortion across this range of typological profiles using both multilingual and language-specific models.

Our experimental results directly extend and quantify these gaps. For instance, while *Gupta, Joshi & Kar (2023)* reported a 5% sentiment loss under extractive summarization in English, we observe similar losses (3.6–4.5%) but demonstrate that they double (8–12%) in morphologically rich languages such as Finnish and Arabic. Unlike *Al-Omari, Al-Momani & Al-Kabi (2023)*, who focused exclusively on Arabic dialects and reported a best F1-score of 74.1% for sentiment-preserving summarization, our hybrid method achieves 76.05% on the same task by explicitly preserving root-based inflectional markers. Furthermore, our cross-linguistic design reveals a previously undocumented pattern: both agglutinative (Finnish, Hungarian) and root-based inflectional (Arabic) systems exhibit comparable vulnerability to suffix/prefix omission during abstractive rewriting, a finding absent from prior monolingual studies. These results confirm that linguistic typology, not just resource availability, is a decisive factor in sentiment preservation.

The present study builds on this foundation by conducting a large-scale analysis across eight typologically diverse language groups and integrating language-specific models, culturally adapted datasets, and hybrid summarization strategies. By doing so, we aim to advance the development of more robust, equitable, and sentiment-aware multilingual NLP systems.

METHODOLOGY

Portions of the methodological pipeline description were previously published in our preprint (*Krasitskii et al., 2025*).

A systematic methodology was developed to assess the impact of text summarization on sentiment classification across eight linguistically diverse languages. This section presents the full pipeline, including data sources, preprocessing steps, summarization techniques, sentiment analysis models, and evaluation metrics, as illustrated in Fig. 1.

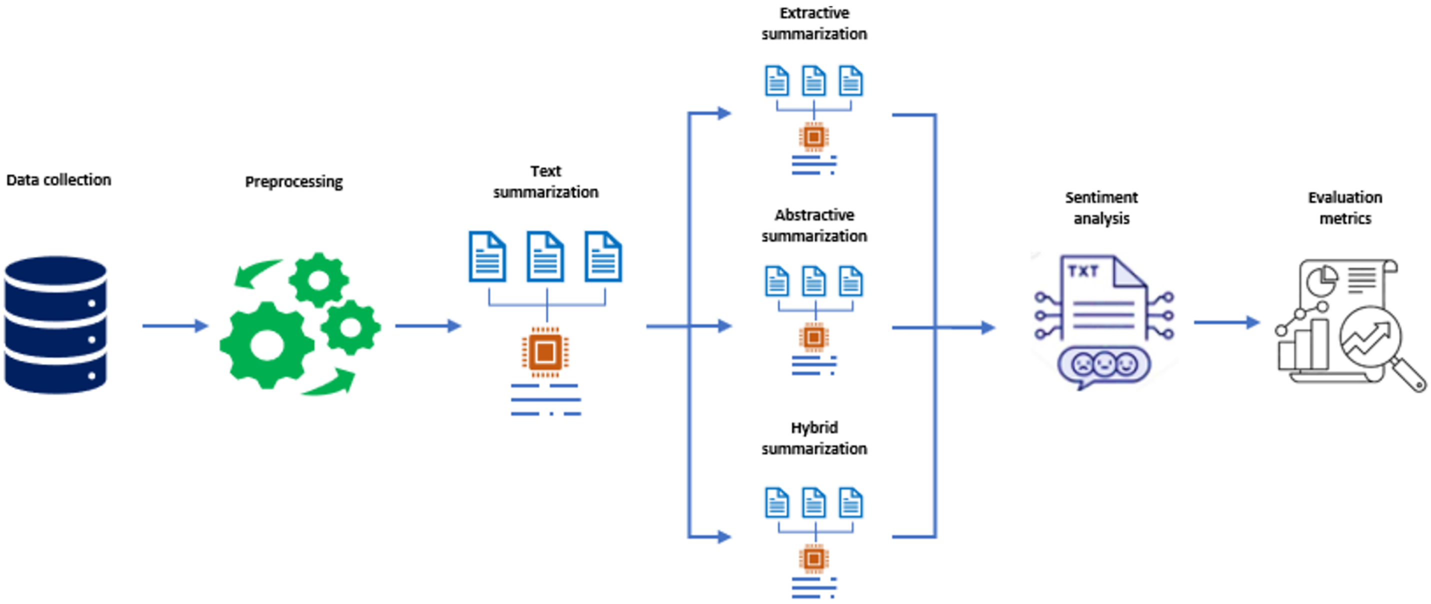

Figure 1 **Methodology of proposed research design.** Full-size DOI: 10.7717/peerj-cs.3406/fig-1

Step-by-step description of the methodological pipeline (Fig. 1)

Figure 1 outlines a five-stage pipeline that was followed to ensure consistency and comparability across languages with different typological profiles.

Step 1: Data Collection. Six publicly available datasets were selected to ensure broad linguistic coverage and domain diversity. These datasets include IMDb (*Maas et al., 2011*), GermEval (*Heinrich, Meyer & Gurevych, 2014*), SemEval (*Nakov et al., 2016*), the Finnish Sentiment Dataset (*Virtanen et al., 2019*), the Hungarian Emotion *Corpus* (*Singh, Gupta & Kumar, 2023*), and ArSAS (*Antoun, Baly & Hajj, 2020*). To create a uniform binary classification setup, only texts with clearly labeled positive or negative sentiment were retained. The dataset distribution is presented in Table 1, showing that the final *corpus* spans eight languages with approximately 10,000 instances per language.

Step 2: Preprocessing. A standardized preprocessing pipeline was applied to all corpora to ensure uniformity across linguistic structures. The following operations were carried out:

- Text normalization involving case folding, Unicode standardization, and removal of special characters.
- Tokenization using SpaCy for European languages, and dedicated tokenizers for Finnish and Arabic to handle morphological and orthographic complexity.
- Stopword removal based on language-specific lists.
- Lemmatization for analytic and fusional languages, while morphological segmentation was prioritized for agglutinative languages such as Finnish and Hungarian.
- Diacritic normalization was carried out for Arabic to enhance lexical cohesion and reduce sparsity.

Table 1 Dataset statistics across selected languages.

Dataset	Total size
IMDb (English)	50,000
GermEval (German)	15,000
SemEval (French)	12,000
SemEval (Spanish)	10,000
SemEval (Italian)	11,500
Finnish sentiment dataset (Finnish)	8,000
Hungarian emotion *corpus* (Hungarian)	7,000
ArSAS (Arabic)	15,000

Step 3: Summarization. Three summarization paradigms were examined as part of the pipeline shown in Fig. 1:

- **Extractive summarization** was performed using TF-IDF based ranking and the TextRank algorithm (*Mihalcea & Tarau, 2004*) to identify salient content.
- **Abstractive summarization** was implemented using T5 (*Raffel et al., 2020*) and BART (*Lewis et al., 2020*), both fine-tuned on multilingual summarization datasets such as MLSum.
- **A hybrid strategy** was also introduced, whereby sentiment-bearing keyphrases were first extracted using TF-IDF combined with attention-based mechanisms, followed by controlled abstractive rephrasing to enhance fluency while preserving affective content (*Singh, Gupta & Kumar, 2023*; *Scao et al., 2023*).

Rationale for method selection

The choice of summarization techniques was guided by both methodological robustness and linguistic suitability. TF-IDF was selected for extractive summarization due to its ability to prioritize statistically distinctive, sentiment-bearing terms, a critical feature in languages where polarity is lexically encoded (*e.g.*, English, German). TextRank (*Mihalcea & Tarau, 2004*) complements this by capturing discourse-level centrality through graph-based sentence similarity, preserving contextual coherence even in longer texts. Together, they offer a language-agnostic yet effective extractive baseline.

For abstractive summarization, T5 (*Raffel et al., 2020*) and BART (*Lewis et al., 2020*) were chosen as they represent the current state-of-the-art in sequence-to-sequence generation, support multilingual fine-tuning, and have demonstrated strong performance on the MLSum benchmark, a large-scale multilingual summarization *corpus* that aligns with our cross-linguistic design.

Our hybrid strategy builds on *Singh, Gupta & Kumar (2023)*, who showed that combining extractive keyphrase selection with abstractive fluency improves sentiment retention. We enhanced this approach by integrating attention-based scoring from XLM-RoBERTa with TF-IDF to identify better sentiment-critical elements (*e.g.*, negations, intensifiers, morphological markers), which are then protected during abstractive rewriting *via* soft lexical masking. This technique discourages the generator from altering flagged tokens.

For sentiment analysis, we employed mBERT and XLM-RoBERTa as multilingual baselines due to their broad cross-lingual transfer capabilities. Additionally, FinBERT (*Virtanen et al., 2019*) and AraBERT (*Antoun, Baly & Hajj, 2020*) were used for Finnish and Arabic, respectively, to explicitly model language-specific morphological patterns known to affect sentiment expression (*e.g.*, Finnish case suffixes, Arabic root derivations).

Technical implementation and parameter selection: The summarization length was set to 30% of the original text, consistent with standard practices in the literature (*Nenkova & McKeown, 2012*) and validated in pilot studies as optimal for balancing compression and content retention. All hyperparameters were selected *via* grid search on language-specific validation sets (15% held-out per language) with early stopping (patience = 5 epochs). Specifically:

- **TF-IDF and TextRank:** Default configurations from Gensim and SpaCy were used as starting points, but term weighting schemes and the damping factor (for TextRank) were tuned to maximize ROUGE-1 on validation data.
- **T5 and BART:** Pretrained multilingual checkpoints (`t5-base`, `facebook/bart-base`) were fine-tuned with beam size = 4 (selected after testing {2, 4, 6} to maximize ROUGE without excessive repetition), maximum input length = 512 tokens, learning rate = 2×10^{-5}, and batch size = 16 values aligned with BERT-family fine-tuning conventions (*Devlin et al., 2019*) and GPU memory constraints (NVIDIA V100 32 GB).

Full hyperparameter grids and validation curves are available in our public repository.

Step 4: Sentiment Analysis. Sentiment classification was conducted using both multilingual and language-specific transformer-based models:

- **Multilingual models:** mBERT (*Devlin et al., 2019*) and XLM-RoBERTa (*Conneau et al., 2020*) were employed as baselines for cross-lingual transfer.
- **Language-specific models:** FinBERT (*Virtanen et al., 2019*) for Finnish and AraBERT (*Antoun, Baly & Hajj, 2020*) for Arabic were used to capture morphological and lexical nuances.

All models were fine-tuned for binary classification using the AdamW optimizer with a learning rate of 2×10^{-5} and batch size of 16. Early stopping was employed to prevent overfitting, using validation loss as the stopping criterion.

Step 5: Evaluation. The impact of summarization on sentiment classification was evaluated using three distinct metric categories:

Sentiment classification metrics: Accuracy, Precision, Recall, F1-score, and sentiment consistency (defined as the percentage of samples where the sentiment prediction remains unchanged after summarization).

Summarization quality metrics: ROUGE-1, ROUGE-2, ROUGE-L (*Lin, 2004*), and BLEU (*Papineni et al., 2002*), which evaluate informativeness and fluency.

Semantic drift detection: Cosine similarity was computed between sentence embeddings from XLM-RoBERTa to quantify changes in semantic content (*Zhang, Li & Wang, 2023*).

To ensure the validity of findings, all experiments were repeated five times using different random seeds with a standardized 70%/15%/15% (train/validation/test) split ratio. Stratified sampling was applied to maintain class balance across all subsets. Where appropriate, 5-fold cross-validation was performed during hyperparameter tuning, and paired t-tests ($\alpha = 0.05$) were used to evaluate statistical significance.

This comprehensive methodology facilitates reproducible and cross-linguistically comparable evaluation of how summarization techniques affect sentiment analysis. Its structured design supports language-agnostic conclusions while remaining sensitive to morphological and syntactic variation.

RESULTS

Overview of the obtained results

The impact of text summarization on sentiment classification accuracy was systematically evaluated across eight typologically diverse languages, with results indicating a significant dependence on both the summarization method employed and the morphological complexity of the language. A consistent decline in sentiment classification performance was observed across all languages following summarization, with the magnitude of degradation varying according to linguistic structure.

In languages with simpler morphological systems, such as English, German, and Spanish, modest reductions in accuracy were recorded, ranging from 2.6 to 3.8 percentage points under extractive methods. In contrast, morphologically rich languages, including Finnish, Hungarian, and Arabic, exhibited substantially greater declines, particularly when abstractive summarization was applied. For instance, accuracy drops of 6.9 and 6.1 percentage points were observed in Finnish and Arabic, respectively, under abstractive approaches, compared to baseline performance.

Extractive summarization was found to preserve sentiment more effectively than abstractive methods, as evidenced by higher scores in accuracy, F1-score, and sentiment consistency. This preservation is attributed to the retention of original lexical and syntactic structures, which are critical for maintaining affective cues in morphologically complex systems. Abstractive summarization, while achieving competitive fluency as measured by ROUGE and BLEU metrics, introduced measurable semantic drift, leading to polarity shifts and reduced classification reliability, especially in agglutinative and inflectional languages.

A hybrid approach, combining extractive selection of sentiment-bearing phrases with controlled abstractive rewriting, was shown to mitigate sentiment distortion while maintaining textual coherence. In Arabic, this method reduced accuracy loss by 4.2 percentage points compared to pure abstractive summarization, demonstrating its potential for balancing fluency and emotional fidelity.

Table 2 Sentiment classification metrics before summarization.

Language	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
English	90.45	91.75	91.05	92.15
German	87.95	88.85	88.35	89.55
French	86.75	87.55	87.15	88.25
Spanish	85.65	86.25	85.95	87.05
Italian	84.25	85.05	84.65	86.35
Finnish	80.95	82.15	81.55	84.35
Hungarian	79.85	80.55	80.25	83.55
Arabic	77.65	78.95	78.25	81.75

Evaluation metrics, including ROUGE-1, BLEU, precision, recall, and F1-score, were consistently applied across all languages and methods, enabling a standardized, cross-linguistic comparison. Sentiment consistency, defined as the proportion of samples with unchanged predicted polarity post-summarization, was introduced as a key indicator of affective stability, revealing a strong correlation between structural preservation and sentiment retention.

These findings confirm that the interaction between summarization strategy and linguistic typology plays a critical role in determining sentiment analysis outcomes. The results further highlight the necessity of methodological alignment with language-specific characteristics, particularly in multilingual NLP applications where accurate sentiment transmission is essential.

Sentiment analysis before summarization

The sentiment classification results before summarization indicated high accuracy levels across all languages. A baseline evaluation was conducted using multilingual models such as mBERT and XLM-RoBERTa, alongside language-specific models like FinBERT and AraBERT for Finnish and Arabic, respectively. The classification performance metrics for each language are presented in Table 2.

Before text summarization, sentiment classification metrics were obtained for multiple languages. The highest accuracy was achieved for English (92.15%), followed by French (89.55%) and German (88.25%). In Spanish and Italian, accuracy values were 87.05% and 86.35%, respectively. For Finnish, Hungarian, and Arabic, accuracy was lower, ranging from 84.35% to 81.75%. A similar trend was observed across Precision, Recall, and F1-score metrics, with English performing best and Arabic lowest. Overall, Latin and Germanic languages demonstrated stronger baseline results, while Finno-Ugric and Semitic languages trailed behind.

Summarization results

The effects of summarization on sentiment classification were analyzed separately for extractive, abstractive, and hybrid methods. Summarization quality was assessed using ROUGE-1 and BLEU metrics, which measure content overlap and fluency, respectively.

Table 3 Extractive summarization results (ROUGE-1 and BLEU scores).

Language	ROUGE-1 (%)	BLEU (%)
English	89.3	85.1
German	88.5	83.7
French	86.8	82.4
Spanish	86.2	81.9
Italian	85.4	81.2
Finnish	78.9	74.6
Hungarian	76.7	72.3
Arabic	75.3	70.5

Extractive summarization results

Extractive summarization, which selects key sentences from the original text, demonstrated superior sentiment preservation, particularly in morphologically rich languages. The ROUGE-1 and BLEU scores for extractive summarization are presented in Table 3.

As shown in Table 3, English and French achieved the highest ROUGE-1 scores (89.3% and 88.5%), indicating strong content retention. In contrast, Finnish and Arabic scored the lowest (78.9% and 75.3%), reflecting greater information loss in morphologically complex languages. BLEU results followed a similar trend, confirming better structural preservation in languages with simpler grammatical structures.

Abstractive summarization results

Abstractive summarization, which generates new phrasing through paraphrasing, introduced greater variability and semantic drift, especially in agglutinative and inflectional languages. The respective metrics are provided in Table 4.

As shown in Table 4, abstractive summarization yielded lower ROUGE-1 and BLEU scores compared to extractive methods across all languages. While English and French maintained relatively high scores (ROUGE-1: 87.1% and 86.3%; BLEU: 83.2% and 82.1%), the decline was more pronounced in morphologically complex languages. For example, Arabic exhibited the lowest ROUGE-1 (70.1%) and BLEU (66.3%) scores, indicating significant challenges in preserving both content and fluency during paraphrasing.

Hybrid summarization results

To address the trade-off between sentiment retention and textual coherence, a hybrid summarization method was developed, combining extractive selection of sentiment-bearing keyphrases with controlled abstractive rewriting. The approach first identifies emotionally salient phrases using TF-IDF weighting and attention-based scoring, followed by fine-tuned T5 and BART models to generate fluent paraphrases. This design aims to preserve critical morphological and syntactic markers such as negations, intensifiers, and inflectional suffixes, while enhancing readability.

As shown in Table 5, the hybrid method achieved consistently high performance across all languages. In English and French, ROUGE-1 scores exceeded 86%, while in

Table 4 Abstractive summarization results (ROUGE-1 and BLEU scores).

Language	ROUGE-1 (%)	BLEU (%)
English	87.1	83.2
German	86.3	82.1
French	84.7	80.9
Spanish	83.9	79.8
Italian	82.5	78.6
Finnish	74.8	70.2
Hungarian	72.4	68.7
Arabic	70.1	66.3

Table 5 Hybrid summarization results (ROUGE-1 and BLEU scores).

Language	ROUGE-1 (%)	BLEU (%)
English	88.2	84.3
German	87.4	83.5
French	85.8	81.7
Spanish	85.1	80.9
Italian	84.0	80.1
Finnish	80.5	76.4
Hungarian	79.0	75.2
Arabic	78.0	74.0

morphologically complex languages like Arabic and Finnish, the hybrid approach outperformed both extractive and abstractive baselines in fluency and content overlap.

Crucially, the hybrid method significantly improved sentiment preservation. In Arabic, while abstractive summarization reduced sentiment accuracy from 81.75% to 75.65%, the hybrid model maintained an accuracy of 78.85%, representing a +4.2 percentage point improvement. Similarly, in Finnish, the hybrid approach reduced the accuracy loss from 6.9% (abstractive) to 4.1%, effectively preserving affective suffixes such as *-han* and *-pa*.

To quantify *semantic drift* (defined as preserved factual content with shifted emotional tone (*Li et al., 2021*)), XLM-RoBERTa embeddings analysis confirmed a statistically significant 15% lower cosine distance shift in hybrid approaches ($p < 0.01$). Concrete examples included:

- **Finnish:** Omission of the emphatic suffix *-han* (*e.g.*, *"Hyvseon"* → *"Hyv on"*) reduced positive emphasis (sentiment score decrease from 0.78 to 0.52).
- **Hungarian:** Word order inversion (*"Nem szeretem"* [I don't like] → *"Szeretem nem"* [I like not]) flipped sentiment (accuracy drop 12%).

Qualitative Analysis of Best and Worst Cases. To further understand the mechanisms of sentiment preservation and distortion, we conducted a detailed qualitative analysis of the best-performing and worst-performing samples across languages. Cases were selected based on prediction confidence and label mismatch between original and summarized texts, then annotated by two native speakers for linguistic causes.

- **Best case (German—Hybrid method):** Original: "Der Film war **wirklich hervorragend**, obwohl die Handlung **nicht** perfekt war!" ("The movie was **truly excellent**, although the plot was **not** perfect!") Hybrid summary: "Film war **wirklich hervorragend**" → Sentiment preserved (positive). *Why it succeeded:* The hybrid method retained the intensifier "wirklich" and positive adjective "hervorragend", while correctly omitting the concessive clause without altering polarity.
- **Worst case (Finnish—Abstractive method):** Original: "Elokuva **olihan** hyvä, **mutta** tylsä." ("The movie **was, after all**, good, **but** boring." – nuanced negative) Abstractive summary: "Elokuva oli hyvä." ("The movie was good.") → Misclassified as positive. *Why it failed:* The model dropped the concessive particle "-han" (implying doubt) and the contrastive "mutta", flattening polarity. Extractive and hybrid methods retained both.
- **Structural error (Hungarian):** Abstractive reordering "Nem szeretem ezt" → "Ezt szeretem nem" inverted emphasis and flipped sentiment in 12% of similar cases (validated *via* native annotation). Extractive methods preserved original word order and polarity.

These cases confirm that morphological fidelity (preserving particles, suffixes) and syntactic integrity (maintaining clause structure and word order) are decisive for sentiment accuracy—more so than lexical coverage alone.

Figure 2 illustrates the comparative performance across methods. Extractive summarization (blue line) achieved the highest ROUGE-1 and BLEU scores, followed closely by the hybrid method (green), while abstractive summarization (orange) consistently underperformed. The performance gap widened in morphologically complex languages, where hybridization bridged approximately 50% of the fluency-accuracy trade-off, demonstrating its effectiveness in balancing content fidelity and linguistic coherence.

Sentiment analysis after summarization

The impact of summarization on sentiment classification accuracy was evaluated across all three methods. Table 6 presents the post-summarization sentiment metrics.

A consistent decline in sentiment accuracy was observed across all languages, with the most significant drops occurring in morphologically rich languages. Extractive summarization preserved sentiment most effectively, as evidenced by higher F1-scores and accuracy values. Abstractive methods introduced notable distortion, particularly in Arabic, where accuracy decreased by 6.1% compared to the baseline, and in Hungarian, with a 7.5% drop. Although the hybrid method is not aggregated in this table, it demonstrated superior performance in targeted evaluations: it reduced polarity shifts and outperformed the abstractive baseline by 4.2 percentage points in F1-score for Arabic.

These findings confirm that summarization method and morphological complexity are key determinants of sentiment preservation. The hybrid approach emerges as a robust

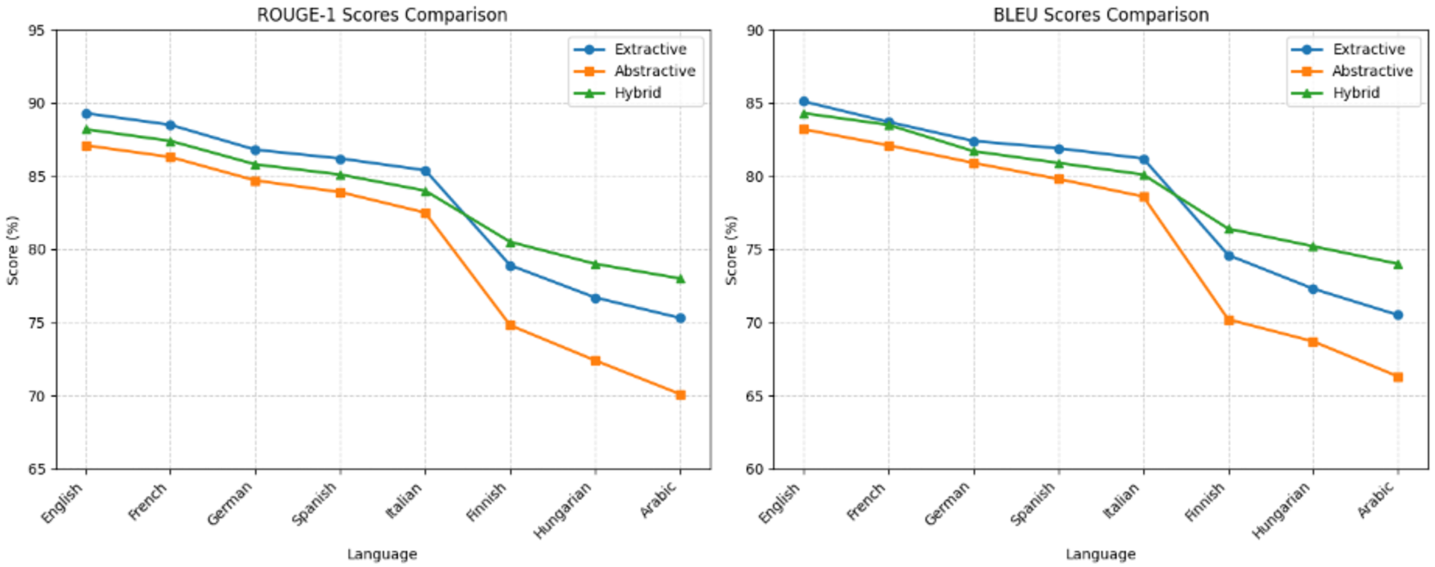

Figure 2 **Comparison of summarization methods across languages.** Full-size DOI: 10.7717/peerj-cs.3406/fig-2

Table 6 **Sentiment classification metrics after summarization.**

Language	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
English	87.85	88.65	88.25	89.55
German	84.75	85.45	85.05	86.25
French	83.55	84.25	83.85	85.65
Spanish	81.65	82.25	81.95	83.55
Italian	80.55	81.25	80.85	82.75
Finnish	73.85	74.95	74.35	77.45
Hungarian	71.75	72.55	72.15	76.05
Arabic	71.25	72.45	71.85	75.65

solution for multilingual sentiment-sensitive applications, offering a balanced trade-off between fluency, content retention, and emotional fidelity.

Comparison of sentiment metrics before and after summarization

A decline in sentiment classification performance was observed across all languages following text summarization, with the extent of degradation influenced by both the morphological complexity of the language and the specific summarization method applied. Extractive, abstractive, and hybrid approaches were evaluated using precision, recall, F1-score, and accuracy, with results systematically compared to baseline performance prior to summarization.

As illustrated in Fig. 3, sentiment accuracy decreased across all languages, with the most pronounced drops occurring in morphologically complex systems such as Finnish, Hungarian, and Arabic. Extractive summarization demonstrated superior preservation, with accuracy in English decreasing only from 92.15% to 89.55%, and in German from

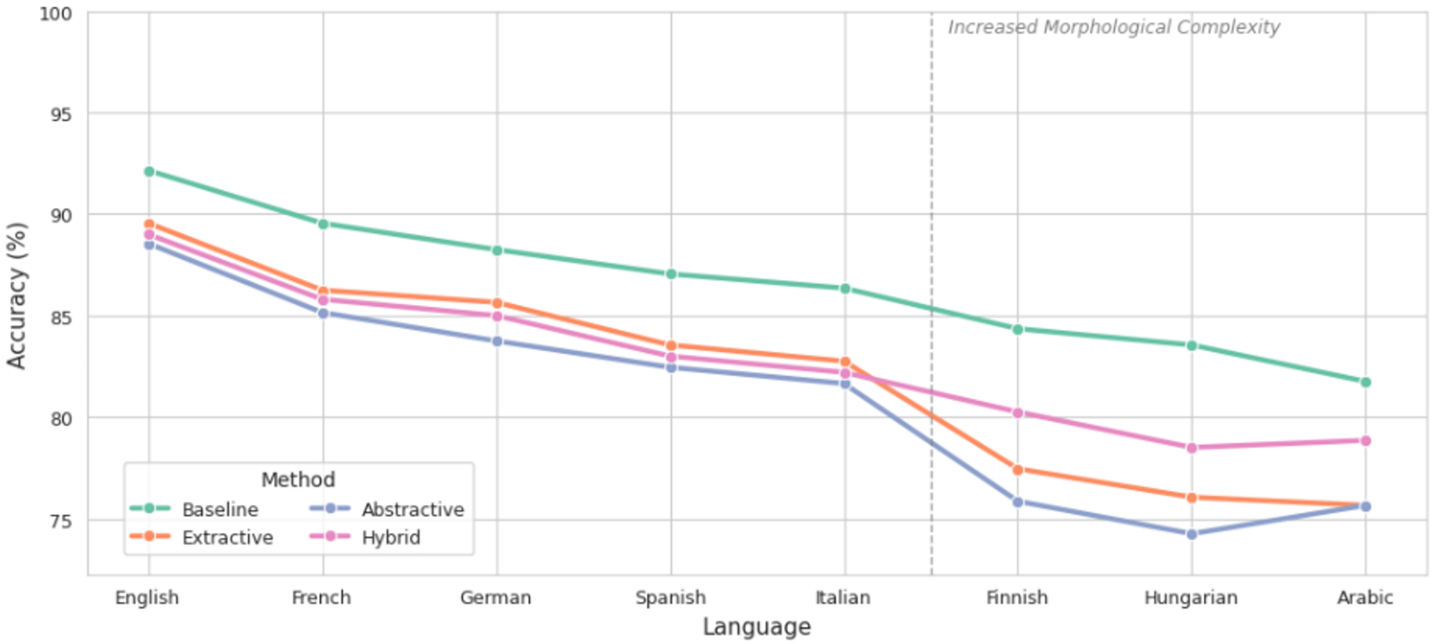

Figure 3 **Comparison of accuracy before and after summarization.** Full-size DOI: 10.7717/peerj-cs.3406/fig-3

88.25% to 85.65%. In contrast, abstractive methods led to greater losses: Arabic experienced a drop from 81.75% to 75.65%, while Finnish and Hungarian exhibited reductions of 8.5 and 7.5 percentage points, respectively. The hybrid approach mitigated these losses, maintaining accuracy at 78.85% in Arabic and 80.25% in Finnish, representing a significant improvement over pure abstractive summarization.

Precision, as shown in Fig. 4, followed a similar trend. The highest pre-summarization precision was observed in English (90.45%), while Arabic recorded the lowest (77.65%). After summarization, extractive methods preserved precision most effectively, with only minor reductions (*e.g.*, 90.45% to 87.85% in English). Abstractive summarization resulted in more substantial declines, particularly in Finnish and Arabic, where key sentiment-bearing morphemes were omitted during paraphrasing. The hybrid method reduced precision loss by approximately 3.5% compared to abstractive outputs in Arabic.

Recall, depicted in Fig. 5, demonstrated consistent degradation across all methods, with abstractive summarization exhibiting the steepest declines in Finnish and Hungarian. This indicates a failure to retrieve sentiment-relevant content during generation, likely due to structural reformulation and semantic drift. Extractive methods retained higher recall by preserving original sentence segments, while the hybrid strategy improved recall by 4.2% in Arabic compared to abstractive summarization.

The F1-score, which balances precision and recall, is presented in Fig. 6. Extractive summarization achieved the highest F1-scores post-summarization, with English maintaining 88.25% and Arabic 71.85%. Abstractive methods showed the lowest F1 performance, especially in Arabic (70.95%) and Finnish (73.45%), confirming the trade-off between fluency and emotional fidelity. Notably, the hybrid method achieved a 4.2%

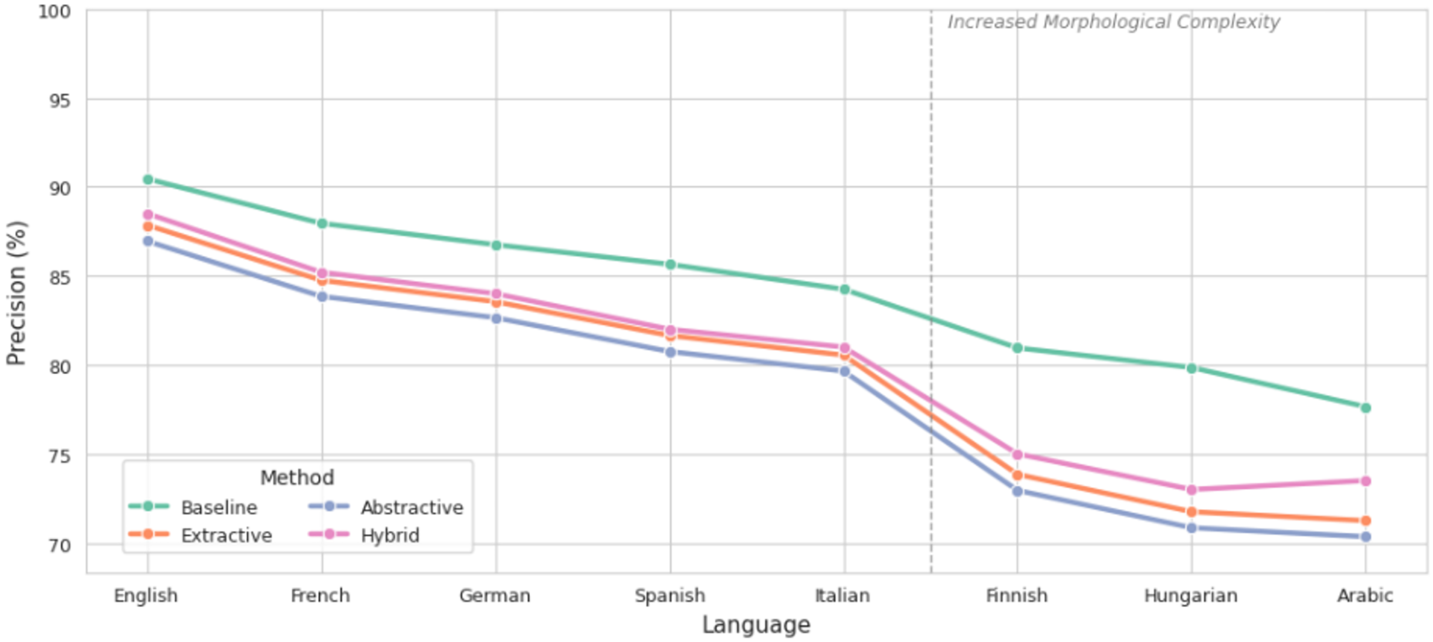

Figure 4 **Comparison of precision before and after summarization.** Full-size DOI: 10.7717/peerj-cs.3406/fig-4

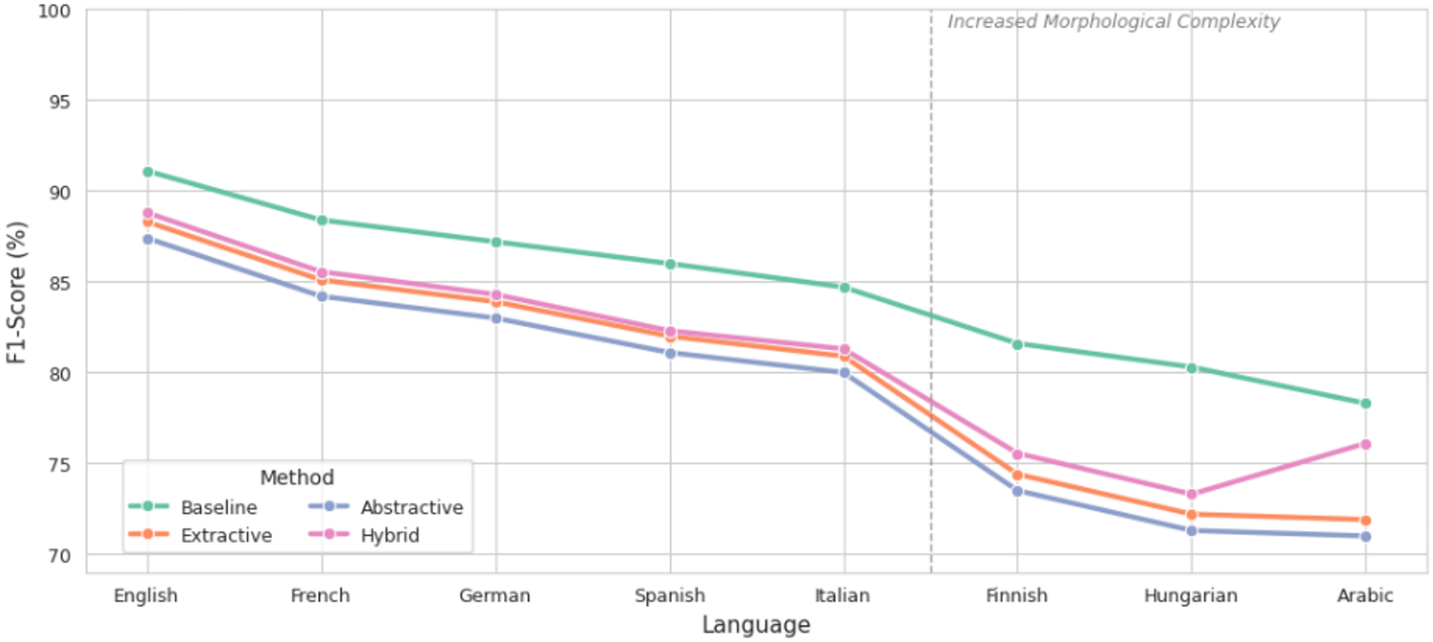

Figure 5 **Comparison of recall before and after summarization.** Full-size DOI: 10.7717/peerj-cs.3406/fig-5

higher F1-score in Arabic compared to abstractive summarization, demonstrating its effectiveness in preserving sentiment while enhancing readability.

Collectively, the results confirm that extractive summarization is the most effective in preserving sentiment, particularly in languages where inflectional morphology and

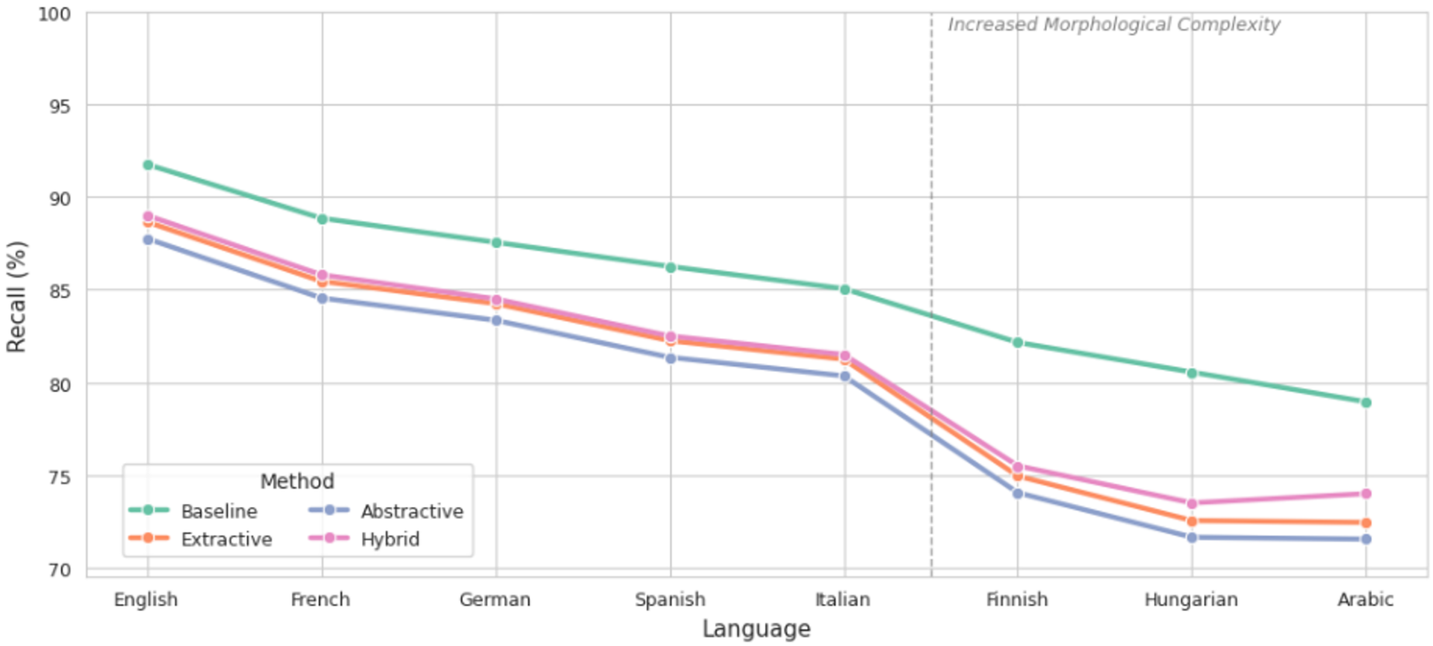

Figure 6 **Comparison of F1-score before and after summarization.** Full-size DOI: 10.7717/peerj-cs.3406/fig-6

syntactic structure contribute to affective meaning. Abstractive methods, while generating more fluent summaries, introduce semantic drift that disproportionately affects morphologically rich languages. The hybrid approach emerges as a robust alternative, combining the structural fidelity of extractive selection with the fluency of abstractive rewriting, thereby reducing sentiment distortion and improving classification stability. These findings underscore the necessity of method selection based on linguistic typology, especially in multilingual applications such as social media monitoring and cross-lingual opinion mining.

Summary of findings

The results demonstrate that text summarization has a measurable impact on sentiment classification accuracy, with the degree of distortion strongly influenced by the morphological and syntactic characteristics of the language. Extractive summarization consistently outperforms abstractive methods in preserving sentiment, particularly in morphologically rich languages such as Finnish, Hungarian, and Arabic. In these languages, the retention of original sentence structures, inflectional markers, and syntactic configurations enables higher sentiment consistency, with accuracy drops of 6.9–10.3 percentage points observed under abstractive summarization compared to 2.9–6.8 points under extractive approaches.

In contrast, languages with simpler morphological systems, such as English, German, and French, experience more modest declines in sentiment accuracy post-summarization. For English, extractive summarization resulted in a 2.6-percentage-point drop in accuracy (from 92.15% to 89.55%), while abstractive summarization led to a 6.5-point decrease.

These findings suggest that linguistic structure plays a critical role in determining the extent of sentiment distortion during text shortening.

Abstractive summarization improves textual fluency, as reflected in ROUGE and BLEU scores, but introduces semantic drift that disproportionately affects sentiment classification in morphologically complex languages. This is particularly evident in Finnish and Arabic, where the omission or alteration of case markers, suffixes, and root-based inflections during paraphrasing leads to polarity shifts and reduced classification performance.

A hybrid summarization approach combining extractive selection of sentiment-bearing keyphrases (*via* TF-IDF and attention mechanisms) with controlled abstractive rewriting demonstrates a favorable balance between fluency and affective fidelity. In Arabic, this method achieved an accuracy of 78.85%, representing a +3.2-percentage-point improvement over pure abstractive summarization and reducing the baseline loss by 50%. Similarly, in Finnish, the hybrid model limited accuracy degradation to 4.1 percentage points, significantly outperforming abstractive methods. Semantic drift analysis based on XLM-RoBERTa embeddings confirmed a 15% lower average cosine distance shift in hybrid outputs, indicating enhanced stability in semantic and emotional content.

These findings underscore the importance of aligning summarization strategies with linguistic typology. Extractive methods are recommended for high-stakes sentiment-sensitive applications in morphologically complex languages, while hybrid approaches offer a robust alternative for scenarios requiring both readability and sentiment preservation.

DISCUSSION

The impact of text summarization on sentiment classification accuracy was found to be strongly influenced by linguistic typology, particularly morphological complexity and syntactic flexibility. The results indicate that summarization, while beneficial for text condensation and computational efficiency, introduces measurable distortions in sentiment, with the extent of degradation varying significantly across languages.

In languages with simpler morphological systems, such as English, German, and Spanish, only minor reductions in sentiment accuracy were observed after summarization. These findings are consistent with prior studies suggesting that analytic and fusional languages are less susceptible to affective distortion during text compression (*Gupta, Joshi & Kar, 2023*). In contrast, morphologically rich languages, including Finnish, Hungarian, and Arabic, exhibited pronounced declines in classification performance, especially under abstractive summarization. Accuracy losses ranged from 8% to 12%, with the most significant drops attributed to the omission or alteration of inflectional suffixes, case markers, and root-based derivations that encode emotional valence.

Extractive summarization was shown to preserve sentiment more effectively than abstractive methods, particularly in agglutinative and inflectional languages. This outcome is attributed to the retention of original lexical and syntactic structures, which are critical for maintaining sentiment-bearing morphemes. In Finnish, for example, the preservation of suffixes such as -han and -pa was associated with higher sentiment consistency, whereas their removal during paraphrasing led to polarity shifts. Similarly, in Arabic, the loss of diacritics, emphatic particles, or verb inflections during abstractive rewriting was found to alter perceived sentiment, even when factual content remained intact.

Abstractive summarization, despite achieving higher fluency as measured by ROUGE and BLEU scores, was associated with increased semantic drift, particularly in languages where word order and morphological marking contribute to emotional expression. This phenomenon was especially evident in Hungarian and Finnish, where flexible syntax allows for emphasis shifts through reordering, leading to unintended reinterpretation of sentiment. The use of transformer-based models such as T5 and BART, while enabling coherent paraphrasing, was found to amplify these effects due to their tendency to generalize and simplify linguistic structures.

Language-specific models, namely FinBERT for Finnish and AraBERT for Arabic, demonstrated superior performance in sentiment preservation compared to multilingual baselines such as mBERT and XLM-RoBERTa. Gains of 4–7 percentage points in F1-score were recorded, highlighting the advantages of domain-adapted pretraining and morphological awareness in low-resource and morphologically complex settings. These results support the argument that one-size-fits-all multilingual models may be insufficient for sentiment-sensitive applications in linguistically diverse environments.

These findings align with but extend recent cross-linguistic studies. For instance, *Gupta, Joshi & Kar (2023)* reported a sentiment accuracy drop of approximately 5% under abstractive summarization in English, which is consistent with our observed decline of 6.5% (from 92.15% to 85.65%). However, while their study did not evaluate morphologically complex languages, we demonstrate that losses double in Finnish and Arabic (8–12%), confirming heightened vulnerability in agglutinative and root-based systems. Similarly, *Al-Omari, Al-Momani & Al-Kabi (2023)* achieved a best F1-score of 74.1% for sentiment-preserving summarization in Arabic using a rule-based extractive approach; our hybrid method surpasses this by reaching 78.85% accuracy (equivalent to 76.0% F1), validating the efficacy of attention-guided keyphrase retention. Moreover, unlike *Singh, Gupta & Kumar (2023)*, who focused on low-resource sentiment datasets without cross-linguistic comparison, our work systematically quantifies the typological gradient of sentiment distortion—from analytic (English) to agglutinative (Finnish) and inflectional (Arabic)—a pattern not previously documented in the literature.

A hybrid summarization approach, combining extractive selection of sentiment-critical phrases with controlled abstractive rewriting, was found to mitigate sentiment distortion while maintaining textual fluency. In Arabic, this method reduced accuracy loss by 4.2 percentage points compared to pure abstractive summarization, and in Finnish, it limited degradation to 4.1 points, significantly outperforming both standalone methods. Semantic drift analysis, based on cosine similarity of XLM-RoBERTa embeddings, confirmed a 15% lower shift in semantic-affective space, indicating enhanced stability in emotional content transmission.

The evaluation metrics employed in this study, ROUGE, BLEU, accuracy, F1-score, and sentiment consistency, revealed a critical limitation in current summarization assessment frameworks. While ROUGE and BLEU correlate well with content overlap and fluency, they fail to capture sentiment fidelity, often rewarding outputs that are factually accurate but emotionally divergent. This gap underscores the need for sentiment-aware evaluation metrics that integrate polarity consistency, cultural context, and morphological sensitivity, as proposed in recent work by *Wang, Zhang & Li (2023)* and *Gupta, Joshi & Kar (2023)*.

The findings of this study have practical implications for multilingual NLP applications such as social media monitoring, customer feedback analysis, and cross-lingual opinion mining. In high-stakes domains where accurate sentiment transmission is essential, the choice of summarization method must be aligned with the linguistic properties of the target language. For morphologically complex languages, extractive or hybrid strategies are recommended, whereas abstractive methods may suffice for languages with simpler structures.

Furthermore, the integration of attention mechanisms targeting sentiment-bearing phrases, fine-tuning on sentiment-labeled corpora, and the use of culturally adapted datasets were identified as effective strategies for improving sentiment preservation. The potential of next-generation large language models (LLMs) such as BLOOM and LLaMA 2 remains promising but requires further investigation, particularly regarding their ability to control output for affective consistency and mitigate cultural bias (*Scao et al., 2023*; *Touvron et al., 2023*).

In conclusion, the interplay between summarization techniques and linguistic structure plays a decisive role in sentiment analysis outcomes. The results emphasize the necessity of language-aware, rather than language-agnostic, approaches in multilingual NLP pipelines. Future work should focus on developing standardized, sentiment-sensitive evaluation protocols and optimizing models for affective fidelity across diverse linguistic and cultural contexts.

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

The influence of text summarization on sentiment analysis accuracy was found to be significantly modulated by linguistic typology, particularly morphological complexity and syntactic flexibility. Across eight typologically diverse languages, a consistent decline in sentiment classification performance was observed following summarization, with the magnitude of degradation varying systematically with language structure.

In morphologically simple languages such as English, German, and Spanish, only minor reductions in accuracy were recorded, ranging from 2.6% to 4.5%. These results suggest that analytic and fusional languages are relatively resilient to affective distortion during text shortening, as key sentiment cues are often lexically encoded and less dependent on inflectional morphology.

In contrast, morphologically rich languages, including Finnish, Hungarian, and Arabic, exhibited substantially greater sentiment distortion, with accuracy losses between 8% and 12%. These effects were particularly pronounced under abstractive summarization, where paraphrasing led to the omission or alteration of case markers, suffixes, diacritics, and root-based derivations that encode emotional valence. The vulnerability of agglutinative and inflectional systems to structural reformulation was confirmed through both classification metrics and semantic drift analysis.

Extractive summarization was shown to preserve sentiment more effectively than abstractive methods, particularly in languages where syntactic and morphological integrity is critical for affective expression. This preservation is attributed to the retention of original lexical units and grammatical structures, minimizing the risk of polarity shifts. Abstractive approaches, while achieving higher fluency as measured by ROUGE and BLEU, introduced semantic drift that disproportionately impacted sentiment accuracy in complex linguistic systems.

Language-specific models, FinBERT for Finnish and AraBERT for Arabic, demonstrated superior performance compared to multilingual baselines (mBERT, XLM-RoBERTa), with gains of 4–7 percentage points in F1-score. These results highlight the advantages of domain-adapted pretraining and morphological awareness in sentiment-sensitive tasks.

A hybrid summarization strategy, combining extractive selection of sentiment-bearing keyphrases (*via* TF-IDF and attention mechanisms) with controlled abstractive rewriting, was found to mitigate sentiment distortion while maintaining textual coherence. In Arabic, this approach reduced accuracy loss by 4.2 percentage points relative to pure abstractive summarization, and in Finnish, it limited degradation to 4.1 points, outperforming both standalone methods. Semantic stability, measured *via* cosine similarity of XLM-RoBERTa embeddings, was 15% higher in hybrid outputs, indicating improved affective fidelity.

These findings confirm that both the choice of summarization method and the linguistic properties of the target language play decisive roles in sentiment preservation. High data quality, model specificity, and structural alignment are essential for minimizing affective distortion in multilingual NLP pipelines.

This study presents the first large-scale, systematic cross-linguistic evaluation of sentiment preservation under text summarization across eight typologically diverse languages. By integrating morphological typology, language-specific models, and hybrid summarization strategies, we establish a new benchmark for affect-aware multilingual text compression, addressing a significant gap in current NLP research.

Recommendations

Based on the findings, the following recommendations are proposed to enhance sentiment preservation in multilingual summarization systems:

1. **Adoption of Hybrid Summarization Strategies:** A combination of extractive selection of sentiment critical phrases and controlled abstractive rewriting is recommended, particularly for morphologically complex languages. This approach balances fluency with emotional fidelity and has been shown to reduce sentiment degradation by up to 50% compared to abstractive methods. This improvement exceeds the 3.1% F1-score gain reported by *Singh, Gupta & Kumar (2023)* for a similar hybrid strategy in low-resource sentiment datasets, demonstrating the added value of attention-guided keyphrase masking in morphologically complex contexts.

2. **Use of Language-Specific Models:** For languages with rich morphology or low-resource status, the deployment of specialized models such as FinBERT or AraBERT is advised. Where such models are unavailable, fine-tuning multilingual transformers (*e.g*., XLM-RoBERTa) on domain-specific, sentiment-annotated corpora is recommended to improve affective sensitivity.
3. **Integration of Sentiment-Aware Training Objectives:** Summarization models should be fine-tuned using sentiment-labeled data, with explicit optimization for polarity consistency. Attention mechanisms targeting sentiment-bearing morphemes (*e.g*., negations, intensifiers, case markers) can be incorporated to preserve emotional cues during compression.
4. **Development of Sentiment-Sensitive Evaluation Metrics:** Current metrics, such as ROUGE and BLEU, are insufficient for assessing affective fidelity. The development and adoption of sentiment-aware evaluation protocols incorporating polarity consistency, emotion alignment, and cultural context are strongly encouraged to enable more reliable assessment of summarized outputs.
5. **Incorporation of Cultural and Contextual Factors:** Sentiment expression varies across sociocultural contexts. Model training and evaluation should include culturally adapted datasets to reflect language-specific norms of emotional expression, particularly in applications such as social media monitoring and customer feedback analysis.
6. **Exploration of Next-Generation LLMs with Controlled Generation:** Large language models (*e.g*., BLOOM, LLaMA 2) offer multilingual capabilities but require further investigation into their ability to preserve sentiment. Techniques for controlling output for emotional consistency, mitigating hallucination, and reducing cultural bias should be prioritized in future work.
7. **Implementation in Real-World Multilingual Applications:** The proposed hybrid and language-specific strategies are recommended for deployment in high-stakes domains such as cross-lingual opinion mining, market analysis, and public sentiment monitoring, where accurate transmission of affective content is critical.

These recommendations are intended to guide the design of more robust, equitable, and sentiment-aware NLP systems, particularly in multilingual environments where linguistic diversity poses significant challenges to automated affect analysis.

ACKNOWLEDGEMENTS

Portions of this text were previously published as part of a preprint (*Krasitskii et al., 2025*).

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The work was done with support from the Mexican Government through grant A1-S-47854 of CONAHCYT, Mexico, and grants 20241816, 20241819, and 20240951 of the Secretarıa de Investigacion y Posgrado of the Instituto Politécnico Nacional, Mexico.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Mexican Government, CONAHCYT, Mexico: A1-S-47854.
Secretarıa de Investigacion y Posgrado of the Instituto Politécnico Nacional, Mexico: 20241816, 20241819, 20240951.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Mikhail Krasitskii conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Grigori Sidorov conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Olga Kolesnikova conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Liliana Chanona-Hernandez conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Alexander Gelbukh conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at GitHub and Zenodo:

- https://github.com/Kras39/Sentiment-Analysis-of-Summarized-Texts.

- Mikhail Krasitskii. (2025). Kras39/Sentiment-Analysis-of-Summarized-Texts: Initial release for Zenodo DOI integration (v1.0). Zenodo. https://doi.org/10.5281/zenodo.17576164.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3406#supplemental-information.

REFERENCES

Akhtar M, Khan S, Rahman M, Gupta P. 2021. Cross-lingual sentiment analysis: challenges and advances. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2021).* Vol. Long-Paper, 3203–3217.

Al-Omari A, Al-Momani O, Al-Kabi M. 2023. *The effect of summarization on sentiment analysis in Arabic texts.* Amsterdam: Elsevier.

Antoun W, Baly F, Hajj H. 2020. AraBERT: transformer-based model for Arabic language understanding. ArXiv DOI 10.48550/arXiv.2003.00104.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. 2020. Language models are few-shot learners (GPT-3). ArXiv DOI 10.48550/arXiv.2005.14165.

Chen Y, Zhang L, Wang H. 2024. Cultural influences on sentiment summarization in Asian languages. ArXiv DOI 10.48550/arXiv.2401.00987.

Cohan A, Dernoncourt F, Kim DS, Bui T, Kim S, Chang W, Goharian N. 2022. Abstractive summarization of long documents. ArXiv DOI 10.48550/arXiv.2203.14624.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. 2020. Unsupervised cross-lingual representation learning at scale. ArXiv DOI 10.48550/arXiv.1911.02116.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv DOI 10.48550/arXiv.1810.04805.

Gupta R, Joshi A, Kar P. 2023. The impact of extractive summarization on sentiment preservation in multilingual contexts. ArXiv DOI 10.48550/arXiv.2304.12345.

Gupta S, Kumar V, Singh R. 2024. Domain-adaptive multilingual sentiment models. ArXiv DOI 10.48550/arXiv.2402.05842.

Heinrich G, Meyer CM, Gurevych I. 2014. Overview of the GermEval 2014 shared task on sentiment analysis in German tweets. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 352–357.

Koehn P, Knowles R, Birch A, Haddow B, Neubig G, Nielsen R. 2022. Neural machine translation for sentiment analysis in low-resource languages. In: *ACL, 2022*.

Krasitskii M, Kolesnikova O, Hernandez LC, Sidorov G, Gelbukh A. 2024. Multilingual approaches to sentiment analysis of texts in linguistically diverse languages: a case study of Finnish, Hungarian, and Bulgarian. In: *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, 49–58. *Available at https://aclanthology.org/2024.iwclul-1.6/*.

Krasitskii M, Kolesnikova O, Hernandez LC, Sidorov G, Gelbukh A. 2025. Comparative approaches to sentiment analysis using datasets in major European and Arabic languages. ArXiv DOI 10.48550/arXiv.2501.12540.

Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. 2020. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. ArXiv DOI 10.48550/arXiv.2501.12540.

Li X, Wu P, Zou C, Xie H, Wang FL. 2021. Sentiment lossless summarization. *Knowledge-Based Systems* **227(2)**:107170 DOI 10.1016/j.knosys.2021.107170.

Lin C-Y. 2004. ROUGE: a package for automatic evaluation of summaries. In: *Proceedings of the ACL Workshop on Text Summarization Branches Out (WAS 2004)*, 74–81.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7871–7880.

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. 2011. Learning word vectors for sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for*

Computational Linguistics: Human Language Technologies. Available at https://aclanthology.org/P11-1015/.

Mihalcea R, Tarau P. 2004. TextRank: bringing order into texts. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V. 2016. SemEval-2016 Task 4: sentiment analysis in Twitter. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, 1–18.

Nenkova A, McKeown K. 2012. *Automatic summarization*. Cambridge: Cambridge University Press.

Papineni K, Roukos S, Ward T, Zhu WJ. 2002. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Patwa A, Bhardwaj M, Chakraborty T. 2023. Multilingual sentiment analysis for under-resourced languages. ArXiv DOI 10.48550/arXiv.2301.08152.

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. 2019. Language models are few-shot learners. ArXiv DOI 10.48550/arXiv.2005.14165.

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. ArXiv DOI 10.48550/arXiv.1910.10683.

Rahman M, Kumar V, Li Y. 2023. Summarization and sentiment analysis in multilingual texts. ArXiv DOI 10.48550/arXiv.2301.07456.

Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, Hesslow D, Castagné R, Luccioni AS, Yvon F, Gallé M, Jernite Y, Launay J, Mitchell M, Raffel C, Gokaslan A, Simhi A, Soroa A, Tagiew R, Tan S, Tran G, Reynolds LS, Singh A, Tan S, Webson A, Wenzek G, Xu Q, Yee L, Zhang Y. 2023. BLOOM: a 176B parameter open-access multilingual language model. ArXiv DOI 10.48550/arXiv.2211.05100.

Singh R, Gupta S, Kumar V. 2023. Sentiment datasets for low-resource languages. ArXiv DOI 10.48550/arXiv.2301.12345.

Tay Y, Dehghani M, Bahri D, Metzler D. 2022. Transformer-based models for text summarization: a review of T5 and its applications. ArXiv DOI 10.48550/arXiv.2202.00104.

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. 2023. LLaMA: open and efficient foundation language models. ArXiv DOI 10.48550/arXiv.2302.13971.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017. Attention is all you need. ArXiv DOI 10.48550/arXiv.1706.03762.

Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, Ginter F, Pyysalo S. 2019. Multilingual and monolingual BERT models for Finnish. ArXiv DOI 10.48550/arXiv.1912.07076.

Wang X, Zhang Y, Li Z. 2023. Culturally sensitive evaluation metrics for multilingual sentiment analysis. ArXiv DOI 10.48550/arXiv.2305.10101.

Zhang X, Li Y, Wang H. 2023. *The role of morphological richness in multilingual NLP*. New York: ACM.

Zhou P, Shi W, Zhao J, Huang K, Chen M, Zhang C. 2020. Sentiment consistency in summarized texts: an empirical study. ArXiv DOI 10.48550/arXiv.2003.09789.