Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00061-7

Original Article

Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals

Kivuyirwa Mystère Kanduki [1, 2], Kalumendo Rodrigue [3], Mpia Héritier Nsenge [4]

1. *Information System, Université de l'Assomption au Congo, Butembo, COD*
2. *Information System, Université Adventiste de Lukanga, Butembo, COD*
3. *Management Information System, Université Adventiste de Lukanga, Butembo, COD*
4. *Computer Engineering, Université de l'Assomption au Congo, Butembo, COD*

Received: March 22, 2026 | Review began: March 24, 2026 | Review ended: April 06, 2026 | Published: April 24, 2026

Abstract

Digital academic platforms increasingly shape how researchers share ideas, communicate results, and sustain collaborations. However, computational studies of scientific discourse remain concentrated on high-resource languages and Western contexts, leaving emerging research communities underexplored. This study investigates sentiment analysis in digital scientific communications produced in the Democratic Republic of the Congo. We collected 26,480 raw texts from Google Scholar, ResearchGate, LinkedIn, X, and public WhatsApp research groups. Because manual annotation was not feasible at this scale, we used a weak-supervision strategy to generate sentiment labels and retained a high-confidence subset of 747 samples for model training and evaluation. This subset should be understood as an initial benchmark rather than a fully representative sample of Congolese scientific discourse. The retained corpus is predominantly French (95%), with limited Swahili and Lingala content. We evaluated term frequency-inverse document frequency with logistic regression, convolutional neural networks, long short-term memory networks, and transformer-based models, while also incorporating a contextual uncertainty signal to improve prediction interpretability. Among the evaluated approaches, BERT (bidirectional encoder representations from transformers) achieved the best overall performance, with a Matthews correlation coefficient of 0.831 and a Macro-F1 score of 0.605, outperforming the classical and neural baselines. These findings indicate that transformer-based models are promising for sentiment analysis in French-dominant Congolese scientific communication, while the moderate Macro-F1 score and limited multilingual coverage highlight the need for larger, more balanced, and externally validated datasets.

Categories: AI applications, Data Mining, Natural Language Processing (NLP)

Keywords: weak supervision, transformers, uncertainty estimation, scientific communication, african nlp

Introduction

Digital technologies have significantly reshaped scientific communication. What was once dominated by one-to-one exchanges, such as correspondence and journal publication, has increasingly shifted to many-to-many interactions via email, forums, and social platforms like WhatsApp, X, LinkedIn, and ResearchGate. In the Democratic Republic of the Congo (DRC), this transformation is particularly visible in highly connected digital spaces [1].

Such digital shifts are closely linked to broader organizational change [2]. They also generate large amounts of unstructured text rich in opinions, emotions, and interactional signals, making text mining and natural language processing (NLP) increasingly important, despite persistent limitations for under-resourced African languages [3].

Despite the growing availability of digital scientific communication, computational studies of this type of discourse remain limited. Most NLP research still focuses on product reviews, general social media content, or political communication. Scientific discourse, however, has its own characteristics, including dense vocabulary, specialized terminology, and often implicit evaluative expressions. In addition, most existing studies rely on English-language corpora from Western research settings, leaving emerging scientific ecosystems comparatively underexplored [4].

This gap is especially evident in the Congolese context, where scientific communication remains understudied, particularly in its multilingual and affective dimensions involving French, Lingala, and Swahili. As a result, little is known about how emotions, attitudes, and evaluative positions shape collaboration and knowledge circulation within the DRC's research community [5]. Although academic communication is often viewed as objective, researchers' exchanges may also convey encouragement, frustration, uncertainty, or irony, all of which can influence scientific visibility and engagement [6]. In this sense, sentiment analysis is useful not only for identifying polarity but also for examining how affective meaning is expressed in specific socio-cultural contexts [7].

To address this gap, this study proposes a weak-supervision framework to build a sentiment dataset and evaluate machine learning models on scientific exchanges produced by Congolese researchers across WhatsApp, X, LinkedIn, ResearchGate, and Google Scholar. The study also takes into account the linguistic and cultural context of the DRC, including sociolinguistic dynamics related to Lingala use in eastern Congo [8]. The study makes four main contributions: it constructs a corpus of Congolese digital scientific communications, develops a weak-supervision framework for sentiment analysis in predominantly French, partially multilingual scientific discourse from the DRC, compares classical and deep learning approaches, and introduces a contextual uncertainty signal to improve the interpretability of predictions. Overall, the study contributes to sentiment analysis in low-resource settings by focusing on an overlooked form of discourse, namely digital scientific communication in the DRC. While the study engages with multilingual dynamics in the DRC, the retained experimental corpus is primarily French-dominant. Accordingly, the present work should be understood as an initial benchmark for Congolese digital scientific communication rather than as a balanced multilingual evaluation across French, Lingala, and Swahili.

Related work

Sentiment analysis has evolved from lexicon-based approaches to machine learning and, more recently, deep learning methods. Early systems relied on predefined dictionaries of positive and negative terms, while later approaches used statistical text representations such as term frequency-inverse document frequency (TF-IDF) combined with classifiers like logistic regression or support vector machines. Although these methods remain useful baselines, they generally treat words as isolated units and therefore capture context only weakly [9]. More broadly, text mining has provided the methodological foundation for extracting patterns, opinions, and trends from large collections of unstructured text, especially in digital communication environments. Typical pipelines include preprocessing steps such as tokenization, normalization, and stop-word removal, followed by modeling techniques for classification, clustering, or dimensionality reduction [10]. These approaches have been widely used to analyze online discourse and user-generated content across platforms [11].

Deep learning has substantially improved sentiment classification by modeling linguistic structure more effectively. Convolutional neural networks (CNN) are useful for capturing local lexical patterns, while long short-term memory (LSTM) networks are better suited to sequential dependencies in text [12]. More recently, transformer-based architectures have become central in NLP because attention mechanisms allow them to model contextual relationships across entire sequences, leading to strong performance in tasks such as text classification and question answering [13]. However, the benefits of these models are not evenly distributed across languages and regions. Many African languages and research communities remain underrepresented because of limited annotated datasets, linguistic diversity, and uneven digital resources [14].

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7

Recent work on low-resource multilingual sentiment analysis has shown that code-switched social media data remain challenging, but transformer-based models can still achieve strong performance when combined with language-specific preprocessing and efficient tuning strategies [15]. In particular, Aliyu et al. [15] reported that AfriBERTa achieved an F1-score of 0.92 on Hausa-English code-switched tweets while maintaining competitive computational efficiency. In the AfriSenti-SemEval setting, the authors evaluated CNN, LSTM, and transformer models on Swahili, Amharic, and Hausa tweets, obtaining macro-F1 scores between 60% and 80%, which illustrates both the promise and the challenges of sentiment modeling in under-resourced contexts [16]. Similar observations were reported by Mabokela et al. [17], who tested multilingual neural language models on African languages and obtained F1 scores above 63% for related languages and up to 77% for Nguni languages. Together, these studies show that advanced models can perform well in African multilingual settings, although most existing work remains centered on social media rather than scholarly communication [17].

Other studies further show how text mining can be adapted to African and socially grounded contexts. The authors combined traditional machine learning and deep learning to study African health-related communication through Semantic Scholar and social media, reaching accuracy levels between 65% and 80% [18]. Liu and Chen [19], using 8,841 mother-child transcripts, combined NLP, lexical analysis, and clustering to identify temporal emotional patterns, highlighting the value of dynamic approaches for extended interactions [19]. In the Congolese context, Rubbers [20] provides a useful empirical backdrop by examining socio-political communication around mining and governance, revealing tensions, power relations, and forms of public expression that resonate with broader communicative practices in the DRC [20]. These issues intersect with structural challenges identified in the Congolese research environment, including isolation and limited resources [21].

Taken together, the literature shows clear progress in sentiment analysis, text mining, and multilingual NLP, but it leaves an important gap [22]. Most prior studies focus on general social media discourse, public debate, or health communication, while digital scientific communication in the DRC remains largely unexplored [8]. This gap justifies the present study, which investigates sentiment in Congolese scientific communications across platforms such as WhatsApp, LinkedIn, X, ResearchGate, and Google Scholar using both classical and deep learning approaches.

Materials And Methods

Data sources

The raw corpus used in this study was assembled from multiple digital sources relevant to online scientific communication in the Congolese context. These sources included scholarly and professional platforms such as ResearchGate, Google Scholar, LinkedIn, and X (formerly Twitter), as well as selected WhatsApp research groups. By combining these complementary sources, we sought to capture both formal and informal forms of scientific communication, ranging from publication-oriented records to shorter interactional exchanges. More specifically, the collected materials comprised two broad categories. First, we gathered document-oriented scholarly content, including publication titles, abstracts, and related metadata retrieved from academic or web-based scholarly sources. Second, we collected conversational and social texts from platforms that support direct interaction and exchange among researchers and professionals. In total, 26,480 textual records were collected. The source-wise distribution is reported in Table *1*.

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7

Platform	Percentage
ResearchGate	38%
Google Scholar	24%
LinkedIn	17%
X	12%
WhatsApp	9%

TABLE 1: Data source distribution

The main steps involved in building the dataset, from data collection to the creation of the high-confidence labeled corpus, are summarized in Figure *1*.

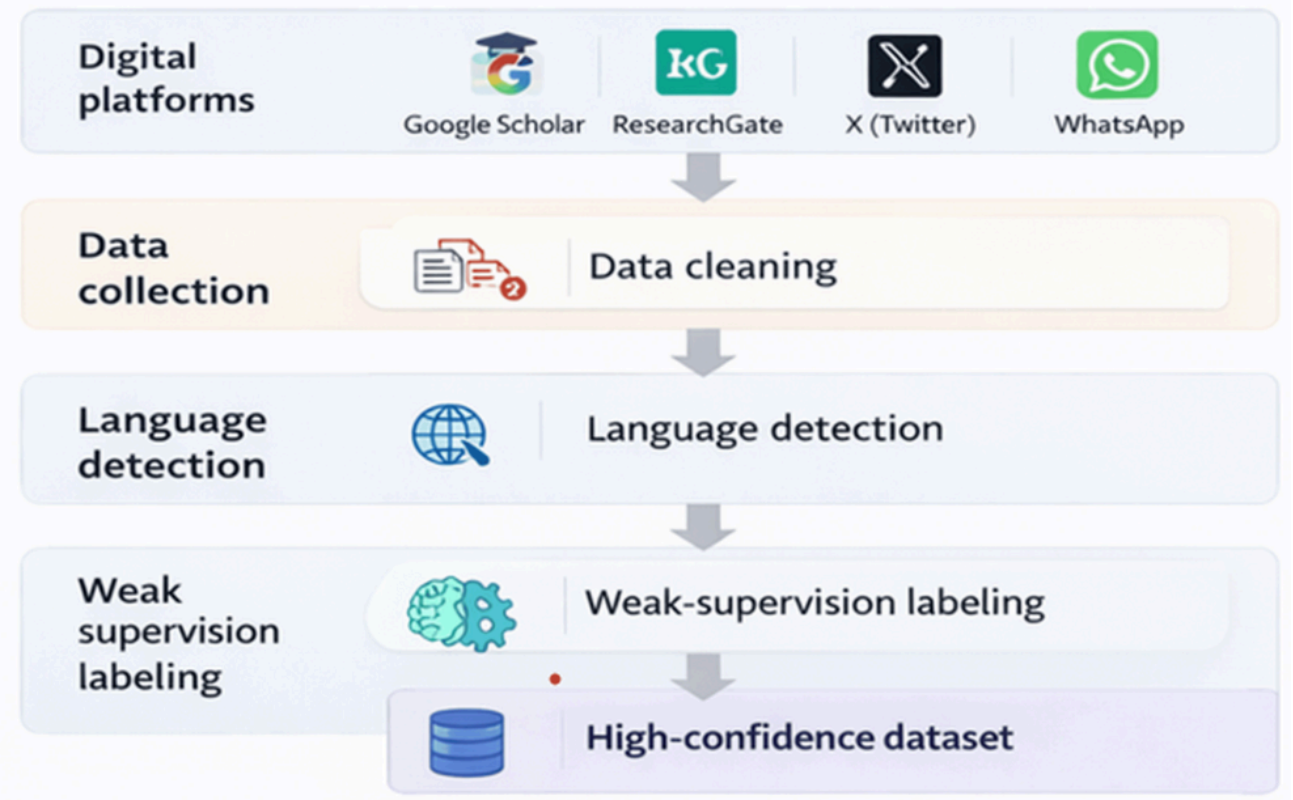

FIGURE 1: Data collection processing pipeline

Data preprocessing

Data preprocessing was carried out through several systematic steps aimed at improving the quality and consistency of the corpus prior to model training. First, duplicate entries were identified and removed to avoid redundancy and any potential bias in the learning process. Second, irrelevant texts, those that did not constitute scientific communications or contained insufficient linguistic information, were removed. The remaining data was then cleaned by normalizing whitespace to ensure consistent formatting throughout the dataset. Finally, titles and abstracts from the same record were merged into a single text sequence.

Formally, each document d_i is converted into a sequence of tokens $(t_{i1}, t_{i2}, \ldots, t_{im})$ and then projected into a latent vector space of dimension k:

$$v_i j \in R^k$$

These embeddings, designed to capture semantic and syntactic patterns, feed into a classification function parameterized by θ:

$$\hat{y}_i = f_\theta(v_i)$$

The predicted polarity $\hat{y}_i$ (negative, neutral, positive) is optimized using the multi-class cross-entropy loss function:

$$L = -\sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic})$$

This mathematical formalization, combined with neural network architectures (CNN, LSTM, BERT [bidirectional encoder representations from transformers]), enables the accurate modeling of sentiments expressed in a multilingual, contextually rich scientific corpus.

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7

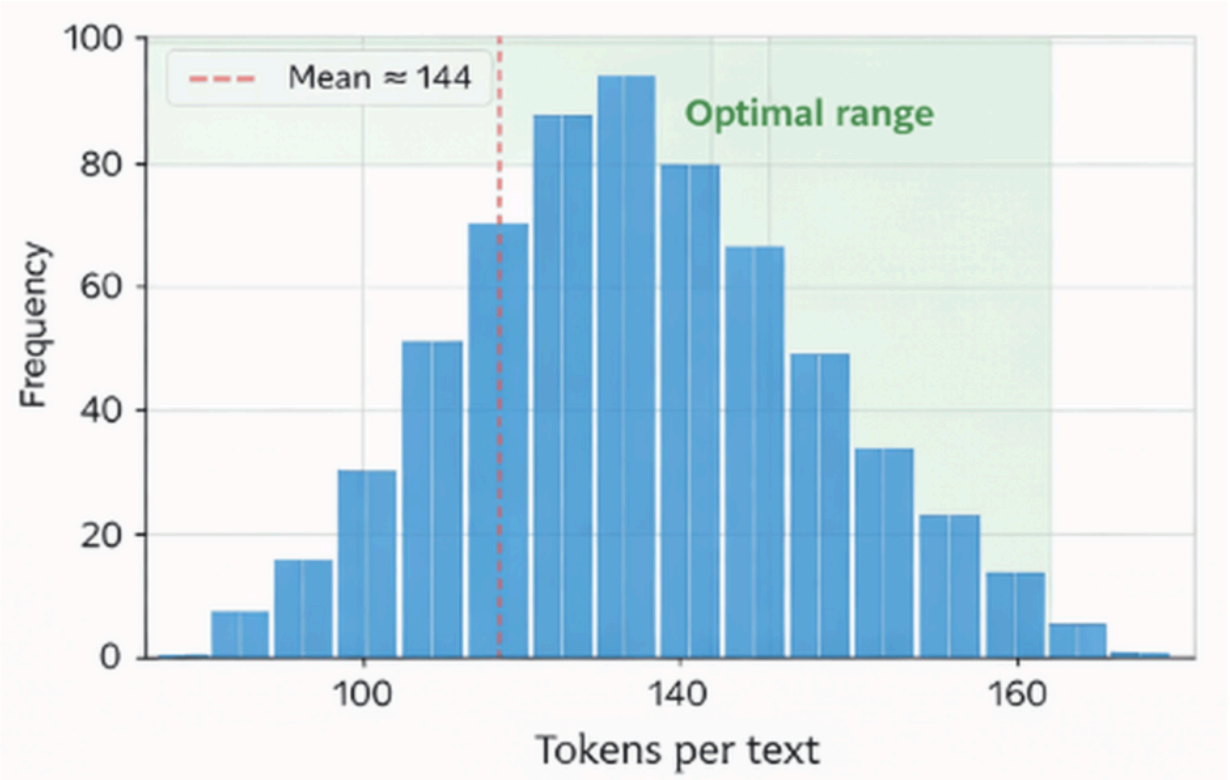

FIGURE 2: Distribution of text lengths in the corpus

The histogram in Figure *2* shows the distribution of token counts across the dataset. The average document length is approximately 144 tokens, and most texts fall within the optimal range for transformer models. Table *2* summarizes the main corpus statistics, including the raw and filtered dataset sizes, average text length, and dominant language.

Feature	Value
Raw dataset	26,480
Filtered dataset	747
Average text length	150 tokens
Main language	French

TABLE 2: Corpus statistics

The linguistic distribution of the collected texts is illustrated in Figure *3*.

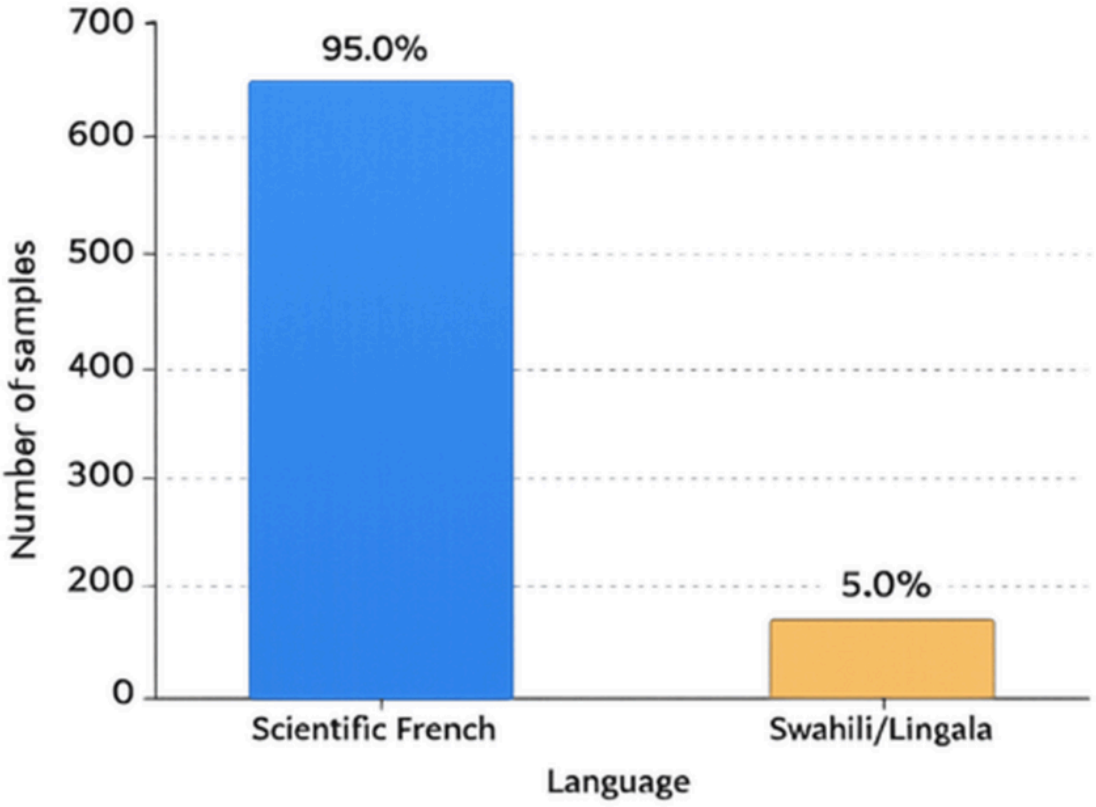

FIGURE 3: Linguistic distribution of the dataset

French dominates the corpus (95%), while Lingala and Swahili account for approximately 5% of the texts.

Weak-supervision framework

Manual annotation of multilingual scientific texts is costly. Therefore, a weak-supervision strategy was implemented. Signals used include (i) sentiment lexicons, (ii) punctuation intensity, (iii) emoji polarity, and (iv) language detection [23]. The sentiment score is computed as follows:

$$score = w_1 S_1 + w_2 S_2 + w_3 S_3 + w_4 S_4$$

Because weak labels can be noisy, we applied a confidence threshold of 0.60 and retained only high-confidence instances for supervised training and evaluation. This choice prioritizes label reliability over corpus size. However, it also reduces coverage and may exclude ambiguous or linguistically mixed cases; therefore, the final 747-sample dataset should not be interpreted as fully representative of the broader 26,480-text raw collection. Only samples with a confidence score of at least 0.60 were retained [24]. As shown in Table *3*, the final weakly labeled dataset contained 481 positive samples (64.4%) and 266 negative samples (35.6%).

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7

Sentiment	Samples	Percentage
Positive	481	64.4%
Negative	266	35.6%

TABLE 3: Sentiment distribution

Contextual uncertainty signal

To improve the interpretability of sentiment predictions, we introduce a contextual uncertainty indicator. For each text, multiple semantic representations are generated using contextual embeddings. Sentiment predictions are computed for each representation.

Let p_i denote the predicted probability of the positive class for representation *i*. The uncertainty score is defined as

$$U = -\sum_i p_i \log(p_i)$$

Higher values indicate greater disagreement between predictions [25]. The diagram in Figure *4* illustrates prediction variability across sampled embeddings.

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7

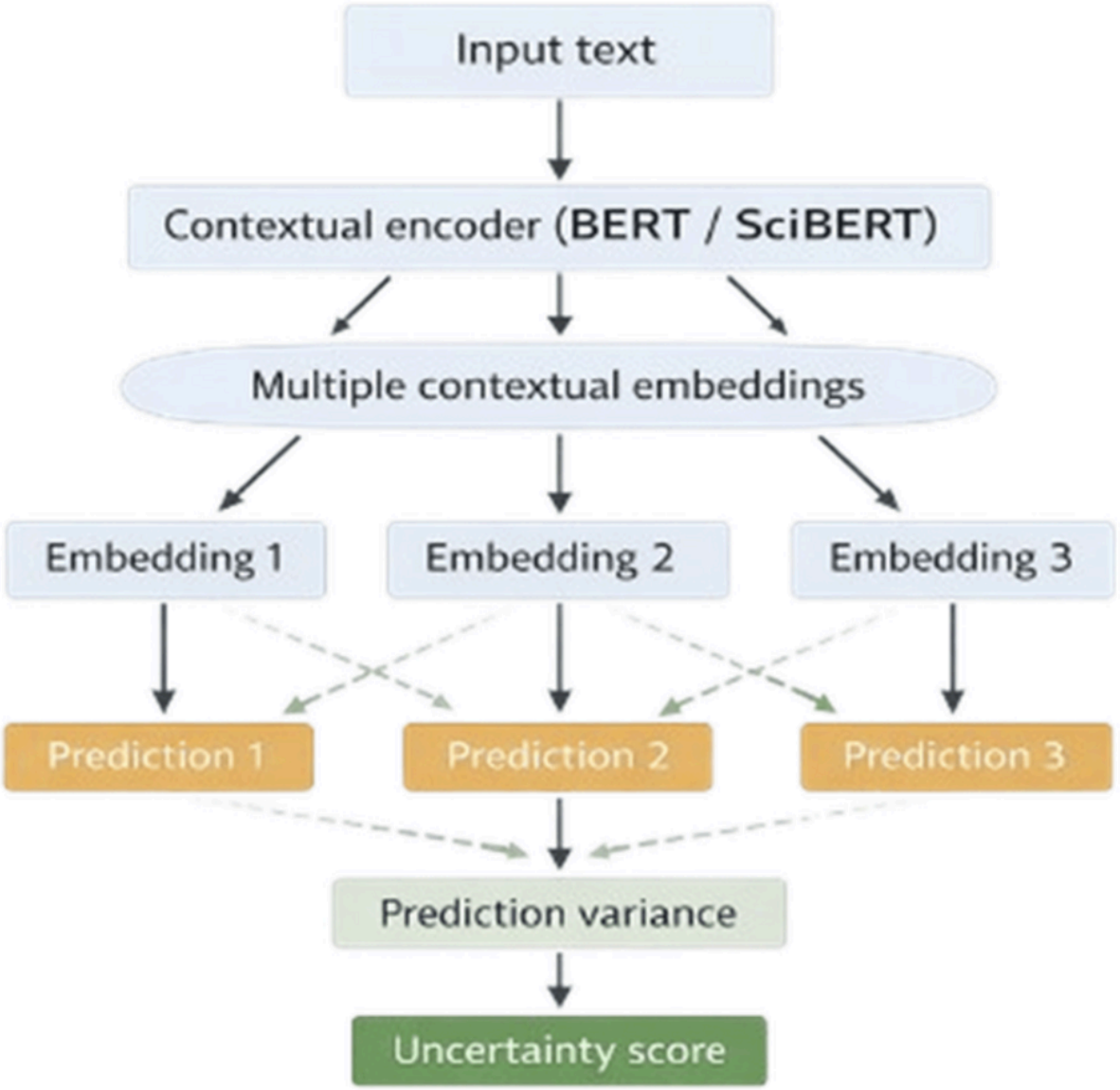

FIGURE 4: Contextual uncertainty estimation

Model architectures

To assess the effectiveness of different modeling strategies for sentiment classification, four representative approaches were evaluated in this study. First, a traditional baseline combining TF-IDF features with logistic regression was implemented, providing a reference for classical machine learning methods based on sparse textual representations. In addition, two deep learning architectures were tested: CNN, which are effective at capturing local lexical patterns, and LSTM, designed to model sequential dependencies across word sequences. Finally, transformer-based models, namely BERT [26] and SciBERT, were evaluated due to their ability to generate contextualized representations of text. Unlike earlier architectures, transformers rely on self-attention mechanisms that allow the model to capture relationships between all tokens in a sentence simultaneously, making them particularly suitable for complex and context-rich texts such as scientific communications.

Results And Discussion

The performance of the evaluated models is summarized in Table *4*. Among the tested approaches, the transformer-based model BERT achieves the best results, with an accuracy of 0.73, a macro-F1 score of 0.60, and the highest Matthews correlation coefficient (MCC) value of 0.83. This confirms the advantage of contextual transformer architectures for sentiment analysis in scientific texts.

A visual comparison of the model performance is presented in Figure *5*(a), which compares the MCC [27] scores of the different models. The results show that BERT clearly outperforms CNN, LSTM, and the TF-IDF baseline. The classification behavior of the best-performing model is further illustrated in Figure *5*(b) through the confusion matrix, which highlights how the model correctly predicts most positive and negative instances.

Model	Accuracy	Macro-F1	MCC
TF-IDF + LR	0.62	0.48	0.56
CNN	0.69	0.55	0.75
LSTM	0.65	0.50	0.58
BERT	0.73	0.60	0.83

TABLE 4: Model performance

TF-IDF, Term Frequency–Inverse Document Frequency; LR, Logistic Regression; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; BERT, Bidirectional Encoder Representations from Transformers; MCC, Matthews Correlation Coefficient

Although BERT achieved the highest MCC and outperformed the other baselines, the Macro-F1 score of 0.60 indicates moderate class-balanced performance rather than near-solved sentiment classification. The results should therefore be interpreted as evidence of relative model superiority within this dataset, not as definitive evidence of robust generalization across all forms of Congolese scientific communication. The dataset-scaling analysis further suggests that the current performance is partly constrained by data availability, as Macro-F1 increased from 0.41 to 0.59 and then 0.60 as the training set grew from 269 to 626 and 747 samples, respectively.

Figure *5* complements these numerical results by visualizing the comparative MCC scores across models and the confusion matrix of the best-performing model, thereby highlighting both the relative advantage of BERT and the remaining classification errors.

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7

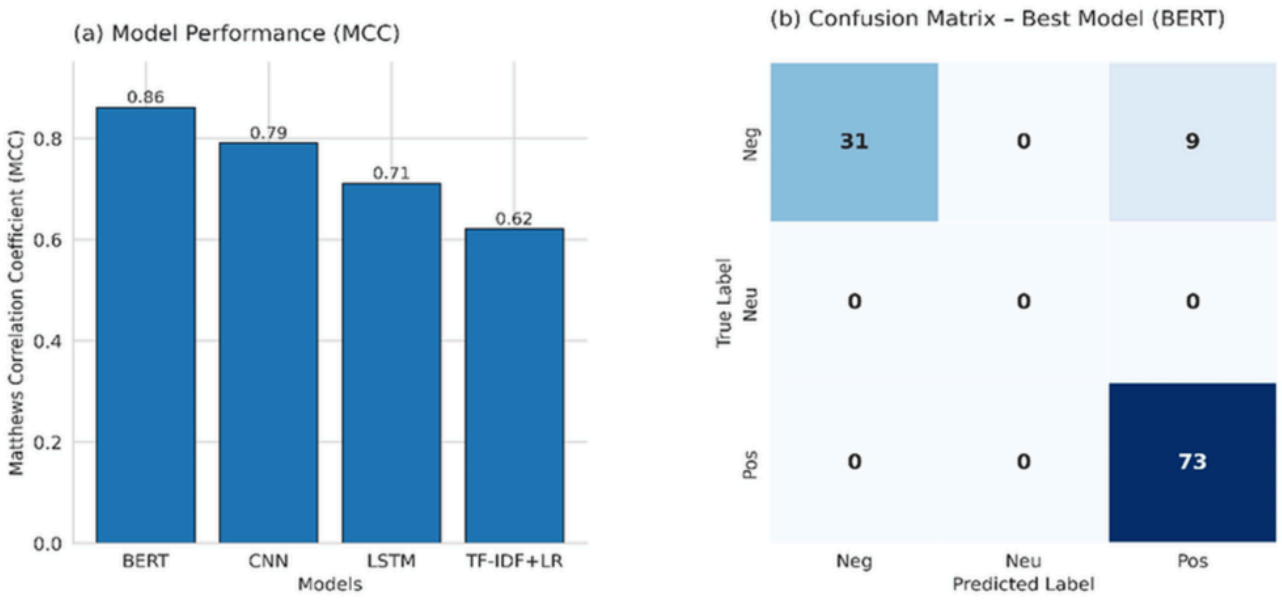

FIGURE 5: Performance evaluation of the sentiment classification models. (a) Comparison of MCC scores across the evaluated models. (b) Confusion matrix of the best-performing model (BERT)

TF-IDF, Term Frequency–Inverse Document Frequency; LR, Logistic Regression; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; BERT, Bidirectional Encoder Representations from Transformers; MCC, Matthews Correlation Coefficient

Dataset scaling analysis

To evaluate the influence of dataset size on model performance, experiments were conducted with increasing dataset sizes. As shown in Table *5*, performance improved consistently with larger datasets, with Macro-F1 increasing from 0.41 for 269 samples to 0.59 for 626 samples and 0.60 for 747 samples.

Dataset size	Macro-F1
269	0.41
626	0.59
747	0.60

TABLE 5: Dataset scaling analysis based on Macro-F1 performance

The table reports the evolution of Macro-F1 as the number of training samples increases.

Visually, it looks like this (Figure *6*):

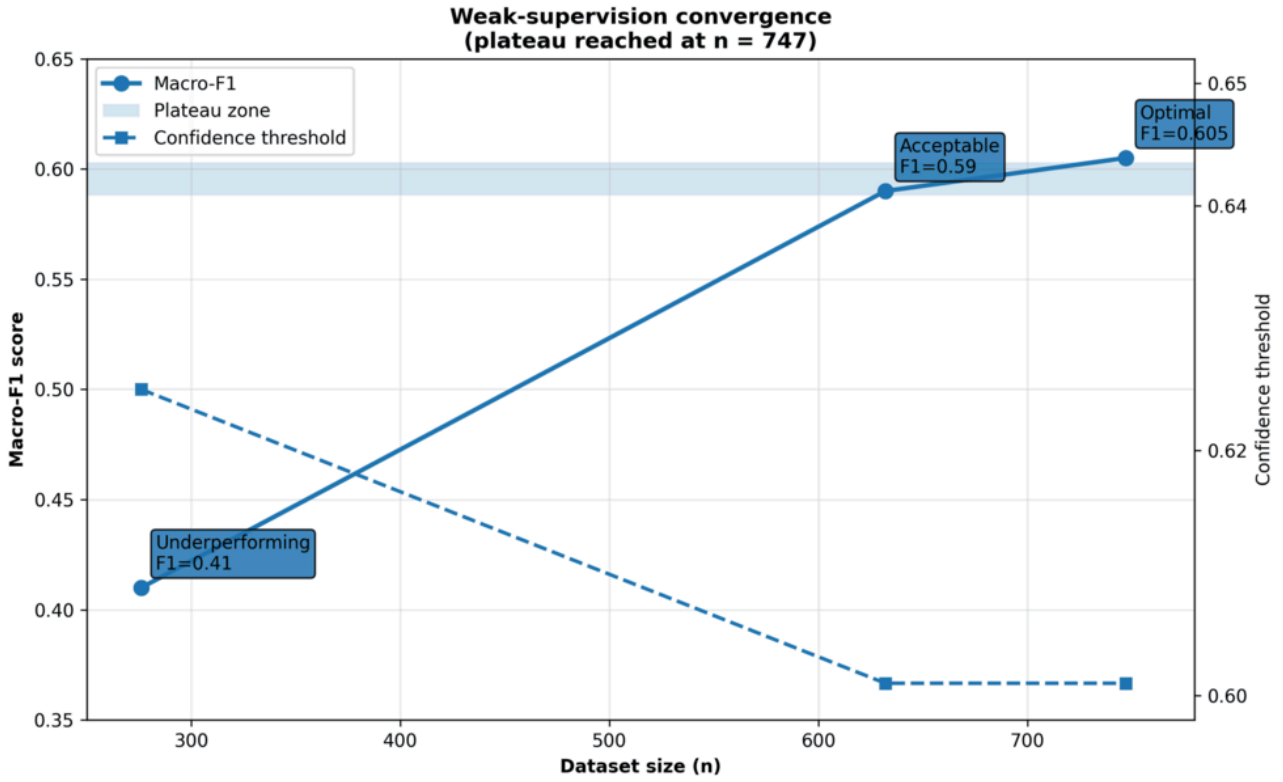

FIGURE 6: Graph showing the scaling of the datasets

Deployment results

The results confirm the superiority of the BERT and SciBERT models (MCC = 0.831), as anticipated in the previous sections on architecture evaluation. This performance can be attributed to their ability to model long-range contextual dependencies, which are crucial for Congolese scientific texts rich in semantic nuances. The predominance of positive polarities (64.4%), consistent with international academic patterns, thus validates the research objectives.

An illustrative example is the following title: Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach with Contextual Uncertainty Signals. The CNN excels at detecting local lexical patterns (e.g., the n-gram "limitation" classified as negative) but fails to capture the overall semantics where "overall robustness" dominates positively. In contrast, BERT, through its multi-head attention mechanism, hierarchically incorporates context and predicts a positive sentiment, accurately reflecting expert human judgment. For practical application, a Streamlit web interface has been deployed, enabling the input of Congolese text, real-time predictions (dominant sentiment, scores by class), and visualizations of overall performance.

Figure *7* presents the web-based interface of the deployed application for real-time sentiment analysis. In this scenario, the user provides a scientific text as input and launches the analysis by selecting the "Analyze" button.

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7

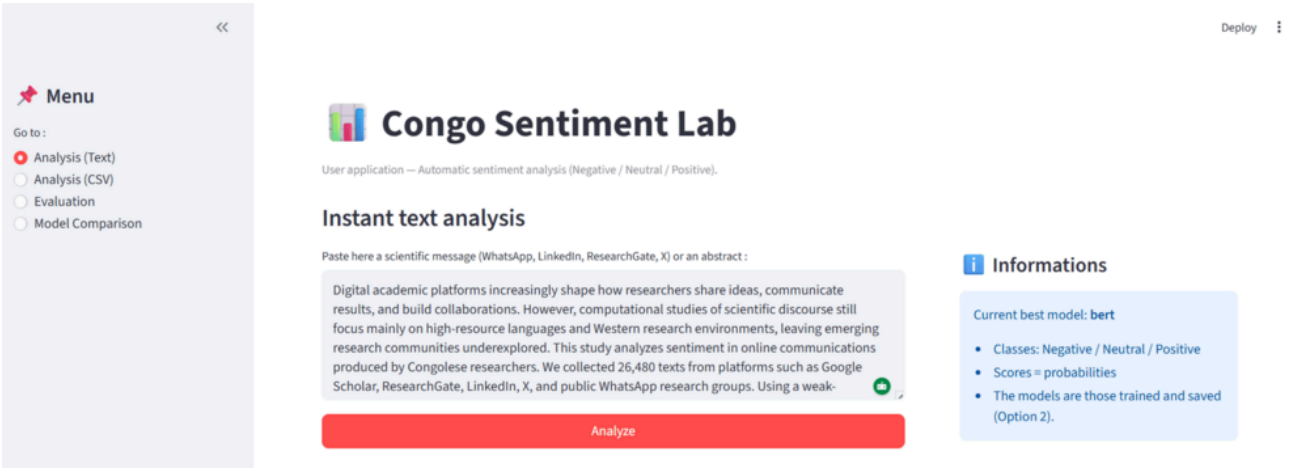

FIGURE 7: Web-based interface of the deployed sentiment analysis system

The resulting prediction generated by the system is displayed in Figure *8*.

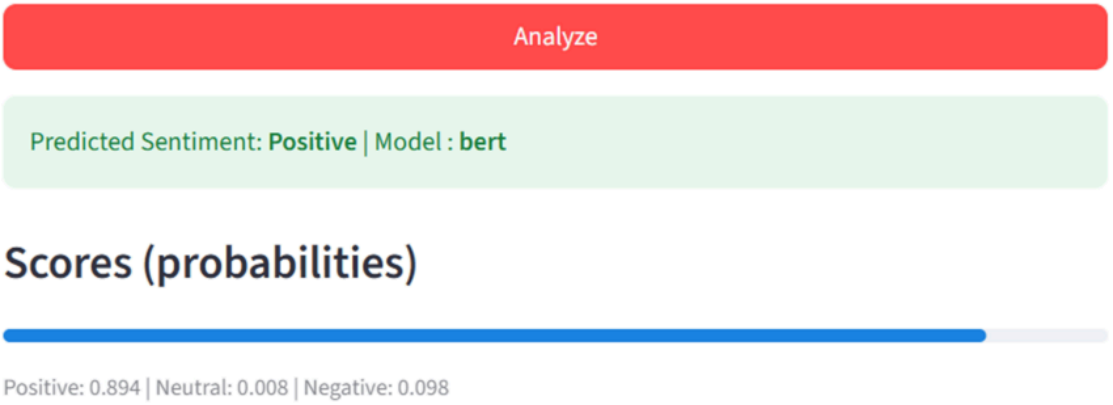

FIGURE 8: Real-time prediction output generated by the deployed BERT model

BERT, Bidirectional Encoder Representations from Transformers

For the selected example, the model assigns the text to the positive class with a confidence score of 89.4%, whereas the neutral and negative classes are associated with very low scores (0.8% and 9.8%, respectively). These results illustrate the operational relevance of the proposed system in a real-time deployment setting.

Discussion

The results demonstrate the effectiveness of transformer models for sentiment analysis of scientific discourse. Compared with classical approaches, transformer models capture contextual relationships more effectively and detect subtle evaluative expressions common in academic texts. The proposed contextual uncertainty signal further improves interpretability by identifying predictions that remain unstable across contextual variations.

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7

Limitations

This study should be interpreted as a proof-of-concept benchmark rather than a definitive multilingual sentiment resource for the DRC. Although the raw corpus comprised 26,480 texts, model training relied on a 747-sample high-confidence subset obtained through weak supervision. This design improves label reliability but may reduce representativeness by excluding ambiguous, mixed, or lower-confidence cases. In addition, the retained corpus is overwhelmingly French-dominant (95%), with only limited coverage of Swahili and Lingala. Accordingly, the present findings mainly reflect French-language or French-dominant scientific communication rather than balanced multilingual discourse. Finally, although the best model achieved a strong MCC, the Macro-F1 remained moderate, indicating that nuanced sentiment classification in this domain remains challenging. Future work should expand the corpus, incorporate more manually validated multilingual data, and evaluate external generalization across datasets and domains.

Conclusions

This paper addressed the underexplored problem of sentiment analysis in digital scientific communications from the DRC. Drawing on texts collected from multiple academic and social platforms, the study proposed a weak-supervision framework to generate sentiment labels in a low-resource setting shaped by partial multilingualism but dominated by French in the retained high-confidence subset. The comparative evaluation showed that transformer-based models, particularly BERT, achieved the strongest performance, highlighting the value of contextual representations for capturing subtle evaluative cues in scientific discourse. In addition, the integration of a contextual uncertainty signal improved prediction interpretability by identifying cases of instability across contextual representations. The deployment experiment further showed that the proposed approach can support real-time sentiment analysis through an operational web interface. Overall, the present study provides encouraging evidence that weak supervision can support initial sentiment modeling for Congolese digital scientific communication, especially in French-dominant settings. However, this work should be viewed as an initial benchmark rather than a definitive multilingual solution, since the restricted size and linguistic composition of the retained high-confidence subset limit broad generalization claims. Future research should therefore focus on larger, more balanced, and externally validated datasets, as well as on stronger comparisons with multilingual and domain-adapted transformer models.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Mpia Héritier Nsenge, Kivuyirwa Mystère Kanduki , Kalumendo Rodrigue

Acquisition, analysis, or interpretation of data: Mpia Héritier Nsenge, Kivuyirwa Mystère Kanduki , Kalumendo Rodrigue

Drafting of the manuscript: Mpia Héritier Nsenge, Kivuyirwa Mystère Kanduki , Kalumendo Rodrigue

Critical review of the manuscript for important intellectual content: Mpia Héritier Nsenge, Kivuyirwa Mystère Kanduki , Kalumendo Rodrigue

Supervision: Kivuyirwa Mystère Kanduki

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial**

relationships: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request. The code, preprocessing scripts, weak-labeling pipeline, and reproducibility materials supporting this study are publicly available in a GitHub repository at https://github.com/staniher/congolesesentimentanalysis

References

1. Williams A, Tkach BK: Access and dissemination of information and emerging media convergence in the Democratic Republic of Congo. Information, Communication & Society. 2022, 25:1383-1399. 10.1080/1369118x.2020.1864004
2. Phogat P, Rab S, Wan M: Science communication in the digital age: Trends, gaps, and interdisciplinary opportunities. Information Services and Use. 2025, 45:148-163. 10.1177/18758789251342896
3. Hoemann K, Lee Y, Dussault È, Devylder S, Ungar LH, Geeraerts D, Mesquita B: The construction of emotional meaning in language. Communications Psychology. 2025, 3:99. 10.1038/s44271-025-00255-0
4. Nandwani P, Verma R: A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining. 2021, 11:81. 10.1007/s13278-021-00776-6
5. Chen S, Mao J, Li G, Ma C, Cao Y: Uncovering sentiment and retweet patterns of disaster-related tweets from a spatiotemporal perspective - A case study of Hurricane Harvey. Telematics and Informatics. 2020, 47:101326. 10.1016/j.tele.2019.101326
6. Pärli R, Byamungu M, Fischer M, et al.: The reality in the DRC is just not the reality in Rwanda" - How context factors affect transdisciplinary research projects. Research Policy. 2024, 53:105035. 10.1016/j.respol.2024.105035
7. Zhang J: Sentiment and language: a socio-semiotic analysis. Philosophy Journal. 2024, 3:118-127. 10.23977/phij.2024.030120
8. Büscher K, D'hondt S, Meeuwis M: Recruiting a nonlocal language for performing local identity: indexical appropriations of Lingala in the Congolese border town Goma. Language in Society. 2013, 42:527-556. 10.1017/s0047404513000651
9. Tan KL, Lee CP, Lim KM: A survey of sentiment analysis: approaches, datasets, and future research. Applied Sciences. 2023, 13:4550. 10.3390/app13074550
10. Hassani A, Iranmanesh A, Mansouri N: Text mining using nonnegative matrix factorization and latent semantic analysis. Neural Computing and Applications. 2021, 33:13745-13766. 10.1007/s00521-021-06014-6
11. Aleqabie HJ, Sfoq MS, Albeer RA, Abd EH: A review of textmining techniques: trends, and applications in various domains. Iraqi Journal For Computer Science and Mathematics. 2024, 5:9. 10.52866/ijcsm.2024.05.01.009
12. Madhurika B, Malleswari DN: Deep learning based SentiNet architecture with hyperparameter optimization for sentiment analysis of customer reviews. Scientific Reports. 2025, 15:35525. 10.1038/s41598-025-19532-3
13. Gordeev M, Aleksandr Z, Bakulin M, Latyshev A, Kozlov D, Yao Y, Anastasia V: Hypercomplex Transformer: novel attention mechanism. Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. 2025, 1845-1851. 10.18653/v1/2025.findings-ijcnlp.115
14. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S: Towards interpretable mental health analysis with large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 6056-6077. 10.18653/v1/2023.emnlp-main.370
15. Aliyu Y, Sarlan A, Danyaro KU, Abd Rahman AS, Muazu AA, Abubakar MY: Deep learning techniques for sentiment analysis in code-switched Hausa-English tweets. International Journal of Information Management Data Insights. 2025, 5:100330. 10.1016/j.jjimei.2025.100330
16. Muhammad SH, Abdulmumin I, Yimam SM, et al.: SemEval-2023 Task 12: sentiment analysis for African languages (AfriSenti-SemEval. Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). 2023, 2319-2337. 10.18653/v1/2023.semeval-1.315

17. Mabokela KR, Primus M, Celik T: Explainable pre-trained language models for sentiment analysis in low-resourced languages. Big Data and Cognitive Computing. 2024, 8:160. 10.3390/bdcc8110160

18. Villanueva-Miranda I, Xie Y, Xiao G: Sentiment analysis in public health: a systematic review of the current state, challenges, and future directions. Frontiers in Public Health. 2025, 13:1609749. 10.3389/fpubh.2025.1609749

19. Liu C, Chen C: Text mining and sentiment analysis: a new lens to explore the emotion dynamics of mother-child interactions. Social Development. 2024, 33:e12733. 10.1111/sode.12733

20. Rubbers B: Mining and the scalar transformations of the state in the Democratic Republic of Congo. The Journal of Modern African Studies. 2023, 61:73-90. 10.1017/s0022278x22000386

21. Horwood C, Mapumulo S, Haskins L, et al.: A North-South-South partnership in higher education to develop health research capacity in the Democratic Republic of the Congo: the challenge of finding a common language. Health Research Policy and Systems. 2021, 19:79. 10.1186/s12961-021-00728-8

22. Merayo N, Ayuso-Lanchares A, González-Sanguino C: Machine learning and natural language processing to assess the emotional impact of influencers' mental health content on Instagram. PeerJ Computer Science. 2024, 10:e2251. 10.7717/peerj-cs.2251

23. Deriu J, Lucchi A, De Luca V, et al.: Leveraging large amounts of weakly supervised data for multi-language sentiment classification. WWW '17: Proceedings of the 26th International Conference on World Wide Web. 2017, 1045-1052. 10.1145/3038912.3052611

24. Jakha H, El Houssaini S, El Houssaini M-A, Ajjaj S, Hadir A: Optimizing sentiment analysis in multilingual balanced datasets: a new comparative approach to enhancing feature extraction performance with ML and DL classifiers. Applied System Innovation. 2025, 8:104. 10.3390/asi8040104

25. Spanos D, Passalis N, Tefas A: Leveraging subclass learning for improving uncertainty estimation in deep learning. Neurocomputing. 2025, 651:130954. 10.1016/j.neucom.2025.130954

26. Mpia HN, Mburu LW, Mwendia SN: CoBERT: A Contextual BERT model for recommending employability profiles of information technology students in unstable developing countries. Engineering Applications of Artificial Intelligence. 2023, 125:106728. 10.1016/j.engappai.2023.106728

27. Mpia HN, Kavalami MM, Lusenge GK, Ushindi KP, Kyalengekania DDK, Ciswaka OM: Lightweight deep learning system for infant-cry recognition with real-time notification in resource-constrained environments. Machine Learning with Applications. 2025, 22:100791. 10.1016/j.mlwa.2025.100791

How to cite this article:
Mystère Kanduki K, Rodrigue K, Héritier Nsenge M (April 24, 2026) Sentiment Analysis in Digital Scientific Communications in the Democratic Republic of the Congo: A Weak-Supervision Approach With Contextual Uncertainty Signals. Cureus J Comput Sci 3 : es44389-026-00061-7. DOI https://doi.org/10.7759/s44389-026-00061-7