International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Context-Aware Hybrid XLM-RoBERTa Framework for Multilingual Cyberbullying Detection in Noisy Social-Media Text

Himanshi Rathore
Department of Artificial Intelligence
Noida Institute of Engineering and Technology
Greater Noida, India
himanshir603@gmail.com

Kajal Singh
Department of Artificial Intelligence
Noida Institute of Engineering and Technology
Greater Noida, India
kajalss1807@gmail.com

Aparna Pandey
Department of Artificial Intelligence
Noida Institute of Engineering and Technology
Greater Noida, India
aparnapandey29@gmail.com

***Abstract*—Cyberbullying detection is a challenging issue because harmful online messages can be unclear and unpredictable. On social media platforms, abusive intent often appears in sarcasm, mixed-language writing, slang, spelling changes, or short coded phrases [2], [3]. The problem becomes even more difficult with Hinglish communication, which mixes English and Hindi in the same message. In this work, we propose a context-aware hybrid cyberbullying detection framework based on XLM-RoBERTa [4]. The system combines transformer-based contextual understanding with lexical indicators, word-level TF-IDF, character-level TF-IDF, and subword-aware cues. A two-stage classification process first determines whether the content is harmful and then identifies the specific type of harm. Focal loss helps improve learning for minority harmful classes [5]. We created the dataset for this study by merging public toxic-language datasets with custom Hinglish samples. We analyzed it further using entropy, Gini impurity, imbalance ratio, coefficient of variation, and lexical diversity. The final model achieved a test accuracy of 0.8099, a weighted F1 score of 0.8113, and a macro F1 score of 0.8192. The results indicate that a hybrid multilingual approach can provide a stronger and more effective way to detect cyberbullying in real online environments.**

***Index Terms*—Cyberbullying detection, XLM-RoBERTa, multilingual NLP, Hinglish, hybrid features, TF-IDF, focal loss, context-aware classification.**

I. Introduction

The evolution of online communication has increased its speed and accessibility to users while creating new pathways for online harassment to develop. The most severe outcome of this transformation operates through cyberbullying. Dangerous messages spread through online channels which maintain their visibility permanently and create emotional and social damage to people. Online bullying exhibits a different pattern from traditional bullying because it enables aggressors to target multiple online spaces while their victims experience ongoing victimization.

The process of detecting cyberbullying from natural language processing requires more than basic classification because earlier research discovered that abusive language identification becomes more challenging when users write in informal styles and need to understand context and user intent [8], [9]. People use sarcasm and informal language with spelling changes and punctuation repetition and coded phrases to express harmful intentions. Users face increasing difficulties when they operate in multilingual environments that require language switching. Indian social media platforms demonstrate a typical pattern where users blend English and Hindi and Hinglish in their online discussions. The real world eliminates numerous English-only systems from effective operation. Jigsaw and TweetEval function as standard resources which help researchers study harmful language although these systems fail to capture all aspects of code-mixed communication on social media platforms [2], [3], [11].

Recent transformer models like BERT, RoBERTa, and XLM-RoBERTa have improved how we understand text and have shown strong results in classification problems [4], [6], [7]. However, many cyberbullying systems still simplify the task by using binary labels, ignoring context, or focusing mainly on one language. Another issue is class imbalance. Severe harmful categories like threats and identity attacks are less frequent than non-bullying content [5].

This work proposes a context-aware hybrid XLM-RoBERTa framework for multilingual cyberbullying detection. The model combines transformer-based contextual learning with lexical and statistical feature support. It uses a two-stage prediction pipeline to improve the separation of harmful content. The goal is to create a stronger classifier and to design a solution that better matches the type of language found on real social media platforms.

II. Contributions

The primary contributions of this work are:

- Multi-lingual framework for detection of cyberbullying in the text which is noise and code-mixed.
- A context-sensitive input design, where the model can utilize previous conversation information when such information is available.
- A hybrid feature strategy that attaches transformer output embeddings with lexical indicators, word level TF-IDF, character level TF-IDF and subword-aware signals.
- A two-stage classification setup, where harmful detection is separated from harmful subtype prediction.
- Combined and analyzed dataset containing public toxic-language datasets as well as our in-house Hinglish data.

III. Review of Literature

Cyberbullying detection has come a long way from traditional machine learning based systems to transformer-based models. Most early work used either bag-of-words, TF-IDF or n-gram features and then models no more complicated than Logistic Regression, Naive Bayes or Support Vector Machines. These have worked well as they're easy to train and relatively interpretable, but they've often struggled with nuanced, informal language as well as context-dependent meaning [8]–[10]. The introduction of pretrained transformer models significantly altered this field. BERT enhanced numerous NLP tasks by acquiring contextual text representations rather than relying solely on superficial word statistics [6]. Later, RoBERTa improved this method and did better on a lot of text classification tasks [7]. This change was important in harmful-language research because the meaning of a word can change depending on the context around it.

The researchers developed RoBERTaNET as a new method which combined RoBERTa processing with GloVe-based enhancements to detect cyberbullying incidents according to the study from [1]. The concept proves useful because offensive content possesses both semantic elements and distinct lexical indicators. XLM-RoBERTa became important for multilingual environments because it was created to learn how multiple languages should be represented simultaneously according to the study from [4]. The technology proves valuable because it enables organizations to examine code-mixed data which combines multiple languages into single sentences.

The available public resources have proven to be valuable for the project. Researchers commonly use the Jigsaw Toxic Comment dataset to study toxicity while TweetEval serves as the reference standard for detecting offensive language in tweets which require tweet-level assessment [3]. The SemEval code-mixed benchmarks demonstrate how important it is to conduct abusive language research which involves analyzing multiple languages and different writing systems according to the study from [11]. Existing datasets lack sufficient multilingual data and authentic conversation patterns and they lack appropriate distribution of classes between different groups. Cyberbullying detection becomes easier when methods which benefit from focal loss to manage data imbalance issues are applied according to the study from [5]. The literature indicates more effective cyberbullying systems should integrate three essential components which include contextual language understanding and multilingual language support and effective control of harmful minority groups. The current study follows that trajectory by focusing on user text which includes noisy elements and code-mixed content and Hinglish-style language use.

IV. Dataset Construction and Statistical Validation

The project requires multiple datasets to improve study realism instead of using a single dataset. The created merged dataset results from combining various public resources together with custom Hinglish cyberbullying samples. The goal was to show the model a wider range of bad words and ways of speaking.

The sources used in this work are:

- Jigsaw Toxic Comment Classification Challenge dataset [2]
- TweetEval Offensive dataset [3]
- Code-Mixed Offensive Language Detection dataset
- Custom Hinglish cyberbullying samples

After cleaning, deduplication, and label alignment, the final label set was defined as:

- non_bullying
- cyberbullying
- insult
- identity_attack
- threat

TABLE I
Original Merged Label Distribution

Class	Samples
non_bullying	216,524
cyberbullying	44,618
insult	6,488
identity_attack	1,302
threat	477
Total	269,409

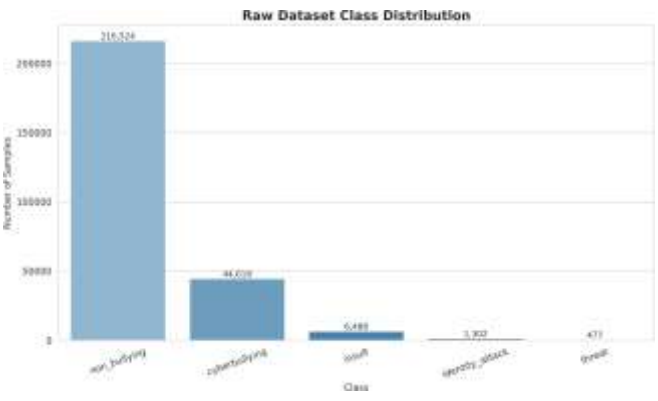


Fig. 1. Raw class distribution in the merged dataset.

The distribution is definitely skewed towards non-bullying content. While this is true to life when dealing with online content, it does create a problem for training, especially if the rare classes of harmful content will be ignored by the algorithm.

This problem is alleviated by balancing the distribution in the training set.

TABLE II
BALANCED TRAINING DISTRIBUTION

Class	Samples
non_bullying	50,000
cyberbullying	44,618
insult	6,488
identity_attack	4,000
threat	4,000
Total	109,106

Fig. 2. Balanced training distribution used during model development.

To better examine dataset quality, several statistical measures were considered. Class proportion is defined as:

$$p_i = \frac{n_i}{N} \tag{1}$$

Shannon entropy is:

$$H = -\sum_{i=1}^{C} p_i \log_2 p_i \tag{2}$$

Normalized entropy is:

$$H_{norm} = \frac{H}{\log_2 C} \tag{3}$$

Gini impurity is:

$$G = 1 - \sum_{i=1}^{C} p_i^2 \tag{4}$$

Imbalance ratio is:

$$IR = \frac{\max(n_i)}{\min(n_i)} \tag{5}$$

Coefficient of variation is:

$$CV = \frac{\sigma}{\mu} \tag{6}$$

Lexical diversity is:

$$LD = \frac{\text{Unique Tokens}}{\text{Total Tokens}} \tag{7}$$

These factors contribute to explaining why this problem cannot be assessed based on accuracy only. The database is characterized by the presence of class imbalance and wide variation in vocabulary, which means that the algorithm needs to be evaluated accordingly.

V. METHODOLOGY

The approach to the research in this paper has been grounded in reality because detecting cyberbullying can be done better by integrating context with linguistic indicators. Transformers are extremely effective in identifying context in adjacent texts; however, abusive language on the internet also relies on lexical features, orthographic patterns, and stylistic traits. Therefore, an integrated model had to be developed.

The overall pipeline is:

Context + Message → XLM-RoBERTa → Hybrid Feature Fusion → Two-Stage Classification → Final Label

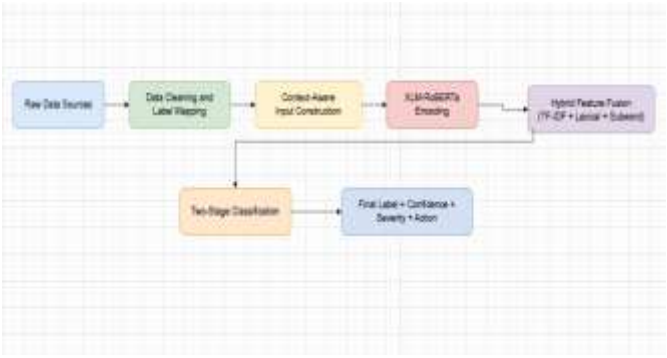


Fig. 3. Overall methodology flow of the proposed cyberbullying detection framework.

A. Preprocessing

The preprocessing phase involves the standardizing of the raw data before starting the training process. Preprocessing may include text cleaning, whitespaces normalizations, duplicates elimination, sources unification, and labels standardizations. Considering the fact that the data originates from various sources, such a step becomes inevitable.

B. Context-Aware Input Construction

There are many harmful messages that may appear to be harmless if viewed by themselves. The actual meaning is revealed only once the preceding message exchange is known. This can be achieved using:

Context: <previous_text> [SEP] Message: <current_text>

This makes the system more suitable for realistic social-media interactions.

C. Transformer Backbone

The primary encoder employed in this research is the XLM-RoBERTa [4] model. This model was selected due to its capability of doing multilingual context-based learning, making it more appropriate for code-mixed text as well as Hinglish [12].

D. Hybrid Feature Fusion

To support the transformer output, the framework also uses:

- word-level TF-IDF
- character-level TF-IDF
- lexical cyberbullying indicators
- subword-aware feature cues

TF-IDF on a word level can be useful in spotting negative phrases. On a character level, it can assist in detecting misspellings and changes made to abusive terms as well as long expressions. Lexical features retain the actual information conveyed by threats, insults, stressed punctuation, and identity-centric terms.

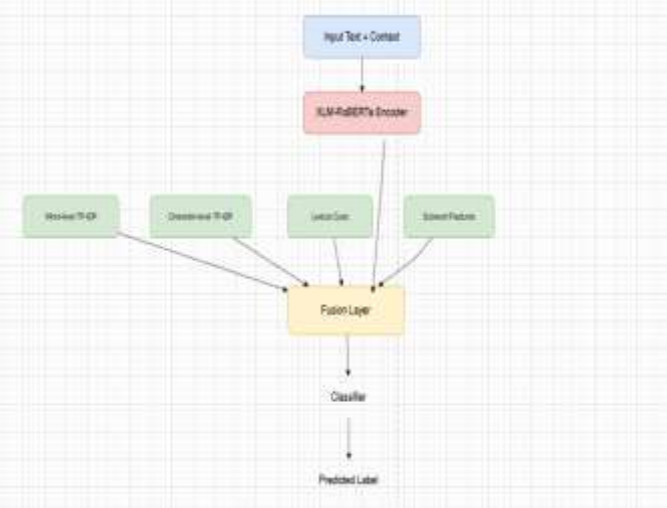


Fig. 4. Hybrid model architecture combining XLM-RoBERTa with lexical and statistical features.

E. Two-Stage Classification

The classifier does not classify all classes in one round but rather takes two steps to reach its final result. In the first step, it determines whether the content is either harmful or non-harmful, and then in the second step, the classifier classifies harmful content into one of the four harmful classes, namely, cyberbullying, insult, identity attack, and threat.

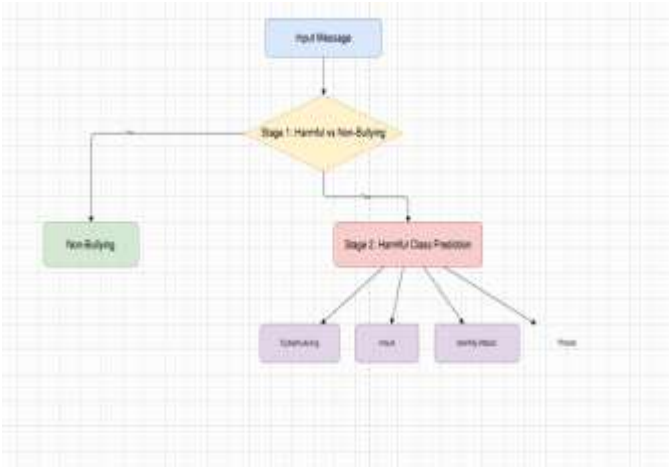


Fig. 5. Two-stage classification strategy used in the proposed system.

F. Loss Function

Since the harmful classes are imbalanced, focal loss is used [5]:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \quad (8)$$

where $\gamma = 2.0$ in this project. This helps increase the influence of difficult and underrepresented samples.

VI. Experimental Setup

All the experiments were conducted using the following libraries: Python, Hugging Face Transformers, PyTorch, Scikit-learn, Datasets, and Streamlit. The transformer architecture used for this work is `xlm-roberta-base`. Focal loss, hybrid features, and two-stage classification were applied in the training process.

The main experimental setup was:

- Backbone: XLM-RoBERTa base
- Loss: focal loss
- Focal gamma: 2.0
- Batch size: 4
- Epochs: 2
- Training samples: 60,000
- Validation samples: 8,000
- Test samples: 8,000
- Hybrid features: enabled
- Two-stage classification: enabled

The model was evaluated using:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

$$Macro\text{-}F1 = \frac{1}{C} \sum_{i=1}^{C} F1_i \quad (13)$$

$$Weighted\text{-}F1 = \sum_{i=1}^{C} \frac{n_i}{N} F1_i \quad (14)$$

Macro F1-score is especially useful in this study because it prevents low-frequency harmful classes from being hidden behind majority-class performance.

VII. Results and Discussion

The final model produced stable results on both validation and test data. The overall performance is shown in Table III.

TABLE III
Overall Model Performance

Split	Accuracy	Weighted F1	Macro F1
Validation	0.8088	0.8105	0.8071
Test	0.8099	0.8113	0.8192

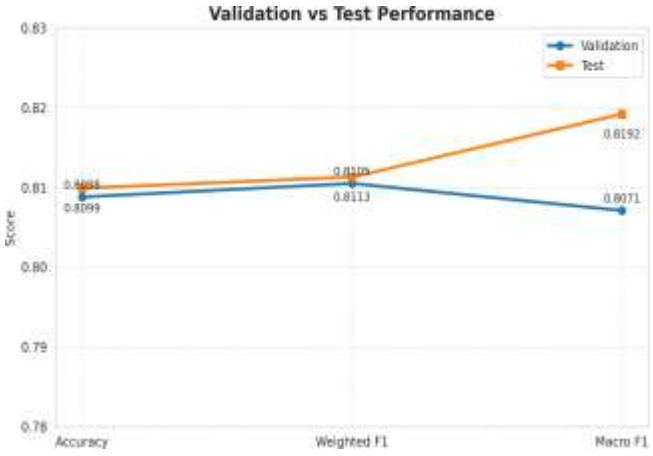
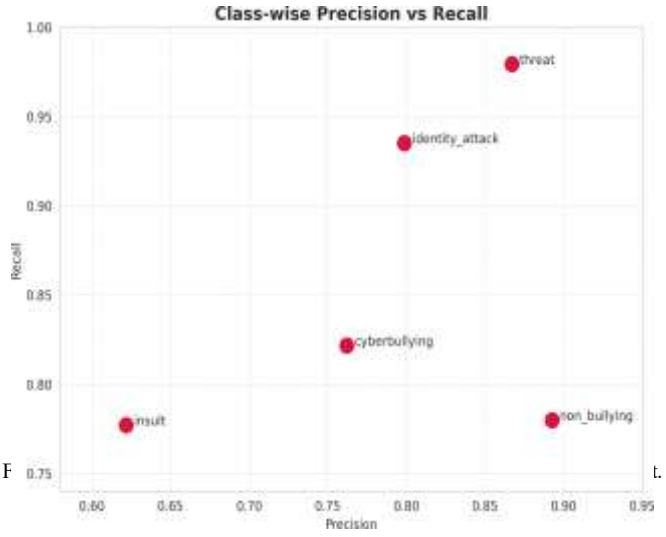


Fig. 6. Validation and test comparison across major evaluation metrics.

The class-wise results are shown in Table IV.

TABLE IV
Per-Class Test Performance

Class	Precision	Recall	F1-score
cyberbullying	0.7622	0.8218	0.7909
identity_attack	0.7988	0.9352	0.8616
insult	0.6218	0.7773	0.6909
non_bullying	0.8926	0.7799	0.8324
threat	0.8671	0.9795	0.9199



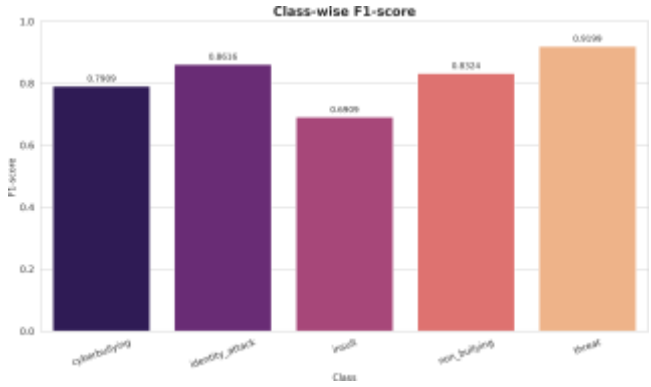


Fig. 8. Class-wise F1-score comparison for the final model.

In terms of classification accuracy, threat and identity_attack achieved the best performance. It is also crucial due to their rarity and significance for practical moderation use cases. Insult classification proved to be more challenging than the others because of its inherent association with sarcasm or humorous or casual forms of offense.

It is worth mentioning a particular observation from the experimental results obtained in this paper: the use of the hybrid architecture resulted in higher stability when processing noisy or mixed texts. Texts characterized by differences in spelling and language mixtures were processed better in combination with transformer features and lexical/TF-IDF-based information.

VIII. Limitations

While the suggested approach demonstrates strong performance overall, there are still several challenges with it. The first challenge relates to the more difficult classification of the insult category compared to the rest of harmful classes. The second challenge is the smaller size of the Hinglish part of the dataset compared to production-size multilingual datasets.

Another challenge is that context can be used if it is available; however, at this point, deeper conversation modeling is not being performed.

IX. Comparison with Existing Approaches

The proposed solution differs from RoBERTa-based approaches like RoBERTaNET [1] in that it incorporates some useful modifications to the current design. They include utilizing XLM-RoBERTa for dealing with the multilinguality of the input; being capable of processing contextually-aware inputs; and integrating outputs of semantic transformers with the results of TF-IDF and sub-word analysis.

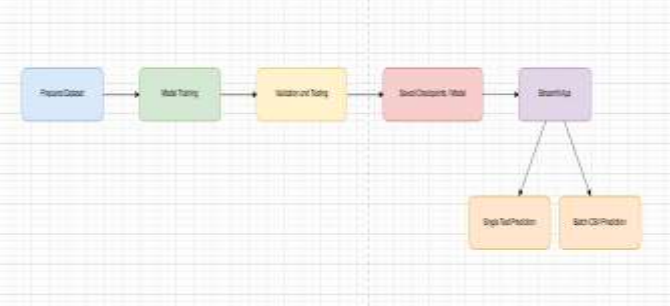


Fig. 9. Training and deployment workflow of the proposed cyberbullying detection system.

X. Conclusion

The paper above described the context-aware hybrid framework based on XLM-RoBERTa model, which was proposed to tackle the problem of cyberbullying detection on multilingual and code-switched texts. This research was driven by the necessity to create more advanced cyberbullying classification systems than those working just for the English language or only in a binary setting. It was demonstrated that context learning along with additional features significantly increases the effectiveness of harmful content detection.

It was also emphasized that it is vital to utilize a diverse dataset and analyze it properly instead of relying on only one source of benchmarks for the task. From the analysis of the experimental results, it follows that hybrid multilingual approaches can be highly beneficial for solving the problem of cyberbullying classification in real-world scenarios [10], [11]. The further research can include exploring Hinglish resources and conversations modeling.

Acknowledgment

The authors would like to sincerely thank the guide, department, and institution for their support and encouragement throughout this work.

References

1. A. M. Alzahrani, M. Alharthi, S. Alsubaie, et al., "RoBERTaNET: Enhanced RoBERTa Transformer-Based Model for Cyberbullying Detection With GloVe Features," 2024.
2. Jigsaw, "Toxic Comment Classification Challenge Dataset," 2018.
3. F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Findings of EMNLP*, 2020.
4. A. Conneau, K. Khandelwal, N. Goyal, et al., "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of ACL*, 2020.
5. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolla´r, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019.
[7] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[8] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Modeling the Detection of Textual Cyberbullying," in *The Social Mobile Web*, 2011.
[9] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improved Cyberbullying Detection Using Gender Information," in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop*, 2013.
[10] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection Using Natural Language Processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 2017.
[11] P. Patwa, G. Sharma, S. PYKL, et al., "SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020.
[12] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?," in *Proceedings of ACL*, 2019.