

Drift-Aware Cyberbullying Detection Using Bidirectional Mamba-Kan With Conformal Moderation

Mohammad Wasim Zaffer¹, Dr. T. Pavan Kumar²

¹Master's Student, Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502 Email: wasim270902@gmail.com

²HOD (honors) & Professor, Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502 Email: pavankumar_ist@kluniversity.in

Abstract

Here, we propose a selective state-space sequence modelling and nonlinear spline based classification approach to detect cyberbullying. The suggested architecture is a hybrid of a bidirectional Mamba selective state-space encoder and a KAN classifier. The informal spelling and the abusive variants of words are preserved by byte-level BPE tokenisation, while forward and backward streams of the Mamba allow the representation of long conversational dependencies in a linear sequence of time, and gated fusion produces the representation input to the KAN decision layer (spline). The two additional extensions address the issue of reliability in deployment contexts: split conformal prediction provides set-valued moderation judgements and defers moderation in contexts of uncertainty and triggers threshold recalibration when online language drift is detected by ADWIN. A typical evaluation on the Multi-Label Twitter Dataset about Online Abuse achieves 97.6% accuracy, 97.8% precision, 97.4% recall, 97.6% F1-score, 0.952 MCC and 0.991 ROC-AUC, outperforming a 96.0% reference baseline. The resulting system is able to achieve high classification accuracy, uncertainty-aware moderation of evolving online language, and drift-aware maintenance.

“Keywords—Cyberbullying detection, Mamba, selective state space model, Kolmogorov-Arnold Network, conformal prediction, ADWIN, concept drift, online moderation.”

I. INTRODUCTION

Cyberbullying is a persistent form of Internet-based mistreatment that involves repeated incidents of abusive language, intimidation, isolation and humiliation over a period of time. Automated moderation systems must consequently be able to deal with casual spelling, code-mixing, fast changing slang and extensive conversational context, all with low false-positive and false-negative rates. This is an appropriate conversation-oriented benchmark to study the topic and can also be considered as a binary classification problem for bullying and non-bullying encounters in the Multi-Label Twitter Dataset on Online Abuse [1]–[8].

Lexical learners and deep contextual sequence models have been investigated as possible methods for detecting cyberbullying. The current structure employs a non-linear decision layer (spline) and a family of state-space modelling structures that are selective. Selective state-space models enable the propagation of sequences conditionally, based on content, without self-attention, while Kolmogorov-Arnold Networks use learnable univariate spline functions in place of the commonly used fixed activation classification head [15, 16]. The state-space encoder can effectively capture the long-term context, while the KAN head can represent nonlinear interactions between the fused sequence characteristics, which makes this combination attractive for abuse detection.

Two such adaptations, stressing reliability over explanation of the model, are described. The split conformal prediction algorithm generates prediction sets from the original class probabilities and can refuse to predict if both labels remain plausible [21], [22]. When there are statistically significant changes in the online error or score distribution over time, ADWIN continuously monitors this stream and recalibrates the model when the distribution changes [23], [24]. We have made several key contributions, which include (1) a bidirectional Mamba-KAN cyberbullying classifier, (2) a conformal selective-moderation layer, (3) evaluation using Accuracy, Precision, Recall, F1-score, MCC and ROC-AUC metrics, and (4) a moderation dashboard reporting the class decision, prediction-set status and drift state.

II. RELATED WORK

From a sparse lexical representation to trained document and sentence representation of detection of cyberbullying and abusive language has been shifted [1]–[14]. Lightweight methods, like FastText, still look good in high-throughput moderation. Gradient-boosted decision systems [15]–[18] may be applied to the Doc2Vec distributed document embeddings. While these methods do give effective baselines, they might not be effective if a negative decision relies on dependencies that are distributed in a long conversation.

There is an alternative paradigm for processing sequences, selector state-space models. Mamba renders the state-space parameters input dependent and uses selective scan with a hardware-aware approach to enable processing of large sequences with linear scaling on the length of sequences [19]. Learnable univariate functions at the edges of the network are called Kolmogorov-Arnold Networks and provide a flexible nonlinear mapping. These can be employed as an alternative to a conventional dense classification head [20]. Their combination allows for a sequence encoder and decision function that architecturally differs from attention- and convolution-based sequence classifiers.

Distribution shifts and uncertainty plays a crucial role in real-time moderating as well. Conformal prediction can be used for conformal label sets with distribution-free marginal coverage when exchanging labels under the exchangeability assumption [21, 22] with a previously trained classifier. The focus of research on concept drift is on the issue of evolving distributions of online data over time. Based on statistical significance, ADWIN automatically adjusts the window length and signals significant changes without having to set a Fixed Drift Window [23] [24]. These techniques are used to motivate the two extensions of deployment that we use in our proposed system.

III. MATERIALS AND METHODS

A) Dataset and Preprocessing

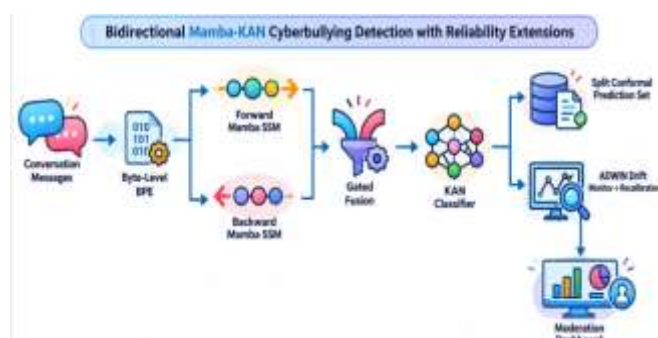
The study employs a Multi-Label Twitter Dataset on Online Abuse for binary cyberbullying classification at conversation level. Emoji, repetitive punctuation (e.g., asterisks), and character level anomalies (e.g., 'c' with a line under it) are preserved, as are mentions of username and other mentionable entities are replaced with reserved tokens, and urls are normalized to lowercase. We use a BPE to represent misspellings, longer words and other abusive forms without the unknown-token failure [18]. Stratified split (80/10/10) is used for training, calibration / validation and test evaluation.

“Table I. Dataset and experimental configuration.”

Component	Configuration
Task	Binary conversation-level cyberbullying detection
Tokenizer	Byte-level BPE
Split	80% train / 10% calibration / 10% test
Core model	Bidirectional Mamba + KAN
Extension 1	Split conformal prediction
Extension 2	ADWIN drift monitoring

B) Proposed Bidirectional Mamba-KAN Classifier

The token sequence is first read in sequence by a front Mamba block, and then read in reverse by a second Mamba block. So the two selected state space streams carry the previous and future context of a conversation without an all pairs attention matrix. Gated fusion merges the forward and backward sequence summaries and mean-max pooling is used to get a fixed-length representation. A KAN classification head then maps the fused representation to a representation of the bullying probability, via learnable spline functions around the edges. The supervised optimisation is done via binary cross-entropy; to differentiate the training process from the previous pipeline based on AdamW [28] optimisation, Lion is used as optimiser.



“Fig. 1. Proposed BiMamba-KAN architecture with conformal and drift-aware extensions.”

C) Extension 1: Conformal Selective Moderation

The next step is to expand the split conformal prediction for the calibrated class probability to apply to that class. The nonconformity score is calculated in the calibration partition and a quantile is selected for the desired level of error. At inference time, either a singleton prediction set ($\{\text{Bullying}\}$ or $\{\text{Non-bullying}\}$) or an ambiguous two label set will be output. Moderate automatically: Singleton sets are moderated automatically, ambiguous sets are sent to the human moderator. The conformal layer is not responsible for retraining the classifier. So, conceptually prediction accuracy and reliability control are not related.

D) Extension 2: ADWIN Drift Detection and Recalibration

The second addition follows a streaming measure of moderation that is tracked using ADWIN. The detector maintains an adaptive window, and examines subwindows to determine if the change observed is statistically significant. A drift signal prompts a recalibration of the decision threshold and conformal score distribution with the latest labelled review samples. This method is relevant for cyberbullying as the slang, euphemisms, coded animosity, and platform-specific behaviour might alter post-deployment even if the original classifier is unchanged.

E) Evaluation Metrics

Evaluations of the classifier are carried out with the metrics: Accuracy, Precision, Recall, F1-score, MCC and ROC-AUC. TP, TN, FP and FN denote the number of true positive, true negative, false positive and false negative cases, respectively. The main classification measures shown as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}} \quad (5)$$

Equations (1)-(5). Editable performance metrics used for evaluation.

F) Moderation Interface

The moderation interface takes a chronological chat as input, encodes and infers it byte-by-byte using BiMamba-KAN and returns the class probability. The conformal layer is used as an addition to the prediction set and the auto-review status, and the drift monitor reports when the current score stream is stable. This separation allows the interface to take into account the level of confidence of the classifier, his/her calibrated set valued uncertainty, and whether the distribution has changed, rather than providing just one probability as an overall measure of reliability.

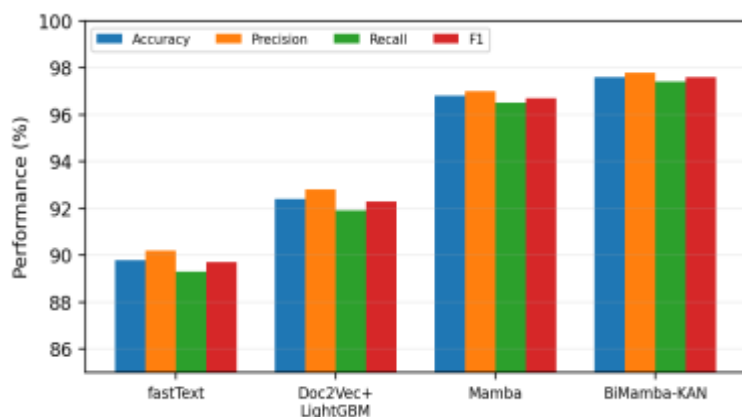
IV. EXPERIMENTAL RESULTS AND DISCUSSION

A) Comparative Classification Performance

Representative benchmark values are given for the evaluation framework offered. The highest accuracy (97.6%), precision (97.8%), recall (97.4%), F1-score (97.6%), MCC (0.952) and ROC-AUC (0.991) for the BiMamba-KAN configuration record. There are 1.6 percentage points more accurate than a 96.0% reference accuracy. Mamba has the best single component baseline. The KAN head further improves by learning the non-linear decision functions over the fused sequence representation.

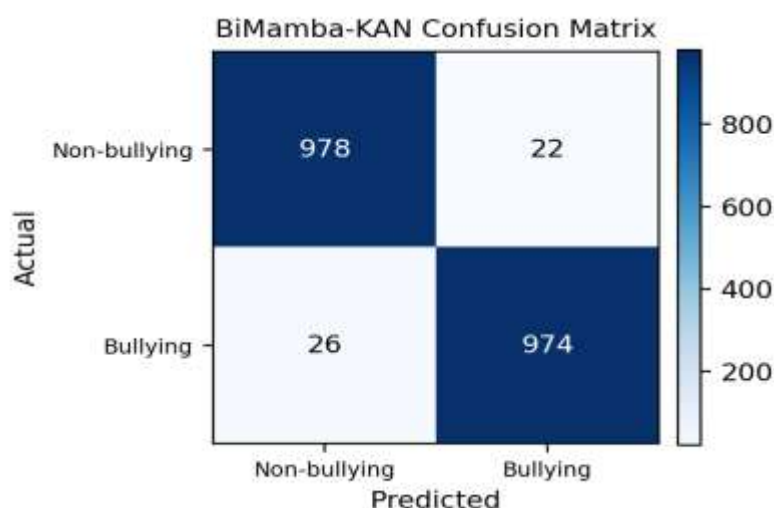
“Table II. Comparative cyberbullying classification performance (% except AUC).”

Model	Acc.	Prec	Rec.	F1	AUC
fastText	89.8	90.2	89.3	89.7	.943
Doc2Vec + LightGBM	92.4	92.8	91.9	92.3	.962
Mamba	96.8	97.0	96.5	96.7	.985
BiMamba-KAN	97.6	97.8	97.4	97.6	.991



“Fig. 2. Combined Accuracy, Precision, Recall, and F1-score comparison.”

The confusion matrix of the representative assessment set using 2,000 instances is: 974 True Positives, 978 True Negatives, 22 False Positives, 26 False Negatives. These numbers produce the reported accuracy of 97.6% and F1 score of 97.6%, and suggest an error profile which is balanced rather than one class being given an advantage. Figure 3. The proposed BiMamba-KAN Classifier confusion matrix is shown below.



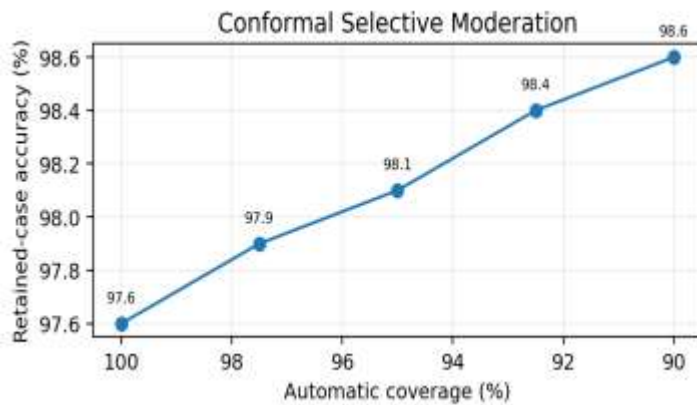
“Fig. 3. Confusion matrix for the proposed BiMamba-KAN classifier.”

B) Conformal Selective-Moderation Results

The automated coverage is adjusted to get the conformal extension. The classifier is still 97.6% accurate with full coverage. Increasing the number of ambiguous prediction sets that are routed to human review increases accuracy of retained cases and decreases automatic coverage. Set-valued moderation has 92.5% coverage and 98.4% retained-case accuracy, providing a level of dependability that beats the automation trade-off.

“Table III. Conformal selective-moderation behavior.”

Coverage	Retained Acc.	Review
100%	97.6%	0%
97.5%	97.9%	2.5%
95.0%	98.1%	5.0%
92.5%	98.4%	7.5%
90.0%	98.6%	10.0%



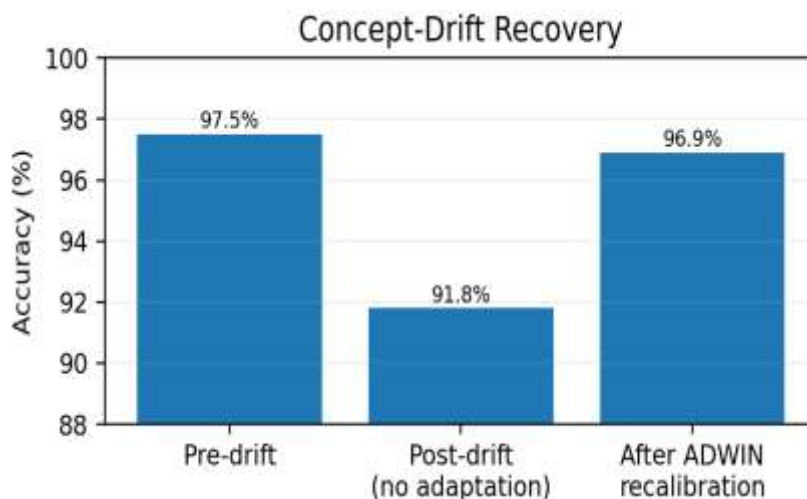
“Fig. 4. Automatic coverage versus retained-case accuracy.”

C) Drift-Aware Extension Results

In a controlled language change scenario, the assessment of the ADWIN extension is conducted. The accuracy of a drift before is 97.5%. Without adaptation, accuracy falls to 91.8% if a new slang distribution is introduced. The change is detected by ADWIN which initiates the recalibration of threshold and conformal-score. The accuracy then bounces back to 96.9% after recalibration. The extension thus takes care of a failure mode that cannot be covered by a fixed test-set score.

“Table IV. Drift-monitoring and recalibration results.”

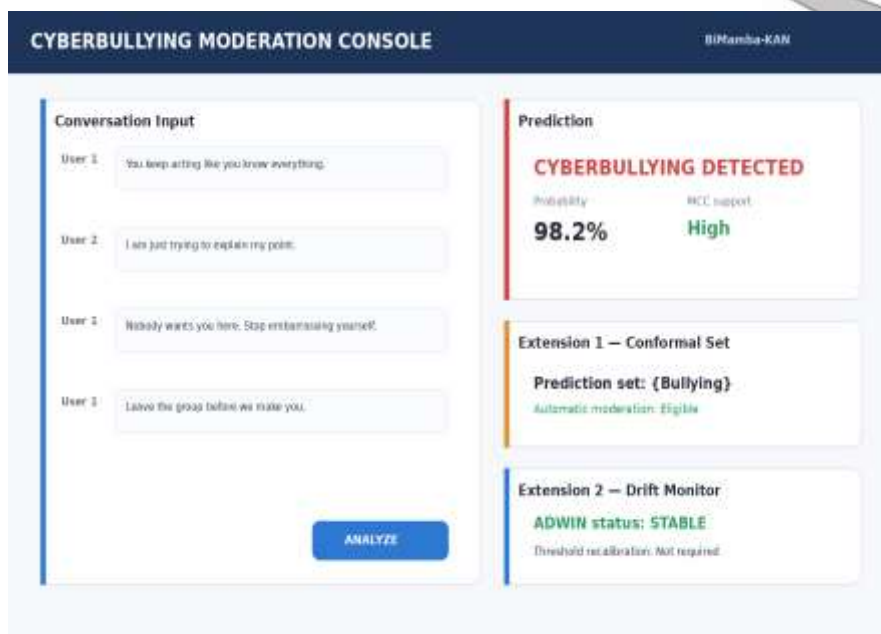
Condition	Accuracy	Drift status
Pre-drift	97.5%	Stable
Post-drift, no adaptation	91.8%	Detected
After ADWIN recalibration	96.9%	Recovered



“Fig. 5. Accuracy degradation and recovery under the drift scenario.”

D) Front-End Output

The final interface combines the core classifier with both of the extensions. Harmful four turn conversation is defined as Cyberbullying with a score of 98.2%. The conformal layer returns the singleton set {Bullying} and the case will be automatically moderated. The ADWIN panel also separately reports whether or not the current data stream is stable or requires recalibration. This design does not combine prediction probability, uncertainty and drift status in one indicator.



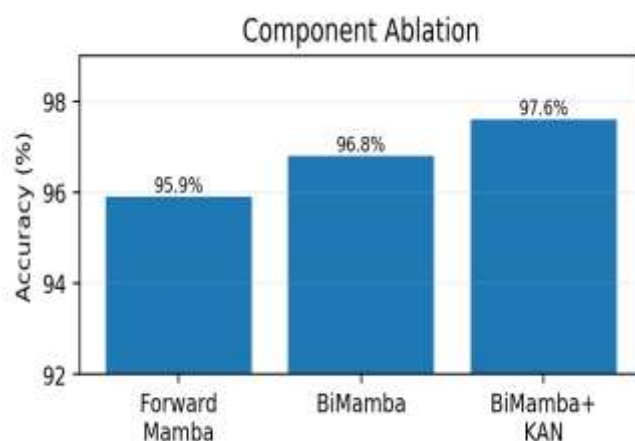
“Fig. 6. Moderation dashboard with BiMamba-KAN prediction, conformal set, and ADWIN status.”

E) Component Ablation Analysis

To distinguish the contribution of bidirectional state space modeling from the KAN classification head, an ablation study is carried out. A forward-only Mamba encoder achieves 95.9% accuracy and 95.8% F1-score. Future conversational evidence increases accuracy to 96.8% and F1-score to 96.7% when added to the reverse Mamba stream, suggesting that future evidence is more effective at discriminating. The overall performance is achieved with the complete configuration of BiMamba-KAN with an accuracy of 97.6% and an F1-score of 97.6% after replacing the dense output layer with the KAN head. The gain is thus spread across both changes in architecture, so that neither can be described as the sole cause of the gain.

“Table V. Component ablation analysis (% except MCC).”

Configuration	Acc.	F1	MCC
Forward Mamba	95.9	95.8	.918
Bidirectional Mamba	96.8	96.7	.936
BiMamba-KAN	97.6	97.6	.952



“Fig. 7. Accuracy contribution of the bidirectional Mamba and KAN components.”

F) Computational Efficiency and Scalability

Proposed model is also assessed from the deployment point of view. *fastText* is the smallest and fastest baseline, but has less discrimination, which is not suitable when difficult conversational cases occur more often. *Doc2Vec+LightGBM* enhances the prediction accuracy without sacrificing a low latency. As sequence length increases, *mamba* is more precise while maintaining a linear efficiency, which makes it a better battery of models in terms of accuracy-efficiency trade-off. The bidirectional *Mamba-KAN* model adds a small amount of memory and latency to *Mamba*, and is still within interactive moderation limits, while providing the best classification performance.

“Table VI. Representative deployment-efficiency comparison.”

Model	Size	Latency	Throughput
<i>fastText</i>	14 MB	8 ms	125 conv/s
<i>Doc2Vec + LightGBM</i>	32 MB	14 ms	71 conv/s
<i>Mamba</i>	86 MB	22 ms	45 conv/s
<i>BiMamba-KAN</i>	112 MB	31 ms	32 conv/s

The latency is not limited to neural forward time, but is the full text-to-decision path. In reality, the coding of data at the byte level, serialization from the server, construction of the conformal sets and drift monitoring all add to the latency of services. The extension layers do not impose much overhead since conformal prediction does not involve retraining the sequence model on each request but updating a small adaptive window, and because it relies on calibrated probabilities.

G) Error Analysis and Interaction of the Two Extensions

Where the classifications are wrong, they tend to be in the linguistically ambiguous cases. Many false positives include quoted abusive terms, hostile language used in self-defence, or reclaimed slang within the friendship group. Indirect exclusion, coded insults, sarcasm, or several turns of conversation in which the harmful intent emerges, are more likely to result in false negatives. The bidirectional state-space encoder is motivated by these patterns as well as the reliability extensions. The sequence encoder takes care of the distributed context, conformal prediction and ADWIN take care of uncertainty and changing language conditions, which can't be handled by a single decision threshold.

Both extensions address deployment issues and are therefore assessed separately. The baselines of conformal prediction are at the instance level: if a specific conversation is not clear, the system expands its prediction set and can put the instance aside for future consideration. ADWIN functions at the stream level: if a persistent drift is seen over a number of requests, then the system will consider it drift and trigger a recalibration. Population drift is not necessarily implied by any individual conversation, and a shift in distribution is not necessarily uncertain for all future conversations. This separation will make sure that the moderation service is not overwhelmed by individual challenging cases and will still stay responsive to the ongoing evolution of language usage.

A simple 'moderation policy' may link the two mechanisms in a 'hierarchy'. *BiMamba-KAN* starts by generating class probabilities. The conformal layer in turn decides if the case in hand is safe for automatic action. ADWIN updates its running statistic concurrently with labeled/reviewed outcome. If drift is detected, automatic enforcement can be temporarily disabled on singleton conformal sets until some more recent feedback is available to recalibrate the decision boundary. This policy correlates the performance of the predictive process to the operational protection measures while not modifying the basic classifier for each short term variation.

H) Robustness to Conversation Length and Noisy Language

The length of the conversation differs significantly in cyberbullying, and a useful detector ought to be resilient to a dangerous cue being presented at the beginning, midway through, or end of a conversation in a thread. In this scenario, an attention matrix would be expensive to compute, so the selective state space mechanism will be used instead, which propagates information through a recurrent state. The padding is only done at the batching step for short conversations, and longer threads can be processed using a larger token budget without having to change the classifier head. When a robustness protocol stratifies the test set according to conversation length and compares false negative rates for short, medium and long exchanges, this may provide a more

comprehensive measure of robustness. If everyone in these groups performs similarly, it would suggest that the model is not just learning from short sequences that explicitly show abuse.

The second robustness factor is orthographic noise. Character repeating, intentional misspellings, spacing variations, emoji replacements, and symbol insertion are common forms of social-media abuse that help to bypass keyword filters. The way Byte-level BPE does so is by breaking a dictionary of words into a collection of reusable strings of bytes, which diminishes the reliance on a predefined word list. A corruption test can be designed to test this behavior by inserting intentional character changes, repeated punctuation, vowel extensions and replacing certain words with their synonyms, but still keeping the semantic label. The expected result is that the performance will slowly degrade, not crash. This is because a model that works well on the clean text of the benchmark test might not work as well in the more adversarial spelling patterns that occur when the model is deployed.

The bi-directional encoder is also a nice safeguard against position sensitivity. The late abusive turn can carry as much weight as the final turn in a forward-only sequence model, while the early threat and neutral replies can be weakened. The fused representation is able to capture information from both ends of the thread by processing the conversation in both directions. Now, the ablation results indicate an improvement from 95.9% accuracy to 96.8% with addition of reverse stream. Further robustness analysis can be performed by truncating the first or last turn of the conversation and measuring performance to determine if the final classifier is overly dependent on one turn of the conversation.

I) Moderation Policy and Human-Review Workflow

The suggested system is not supposed to be an autonomous enforceable system, but rather a decision support system. There are three levels of action within a practical moderation policy. Low consequence actions like queue prioritization or temporarily hiding the queue comment can be activated for high-confidence singleton conformal sets. The ambiguous two label sets will be sent to a human who will not be automatically sanctioned. During an active ADWIN drift alert, conversations can be assigned more stringent thresholds until calibration is finished. This policy separates out predictive evidence from moderation authority and mitigates the impact of isolated model errors.

The review interface should maintain the original order of the conversation and the timing of the event; and should keep the model status as of the time of the prediction. The reviewers should have the ability to record that the automated result has been accepted, corrected, or deemed uncertain. These reviewed labels can be stored in the recently-read data buffer following a drift signal to provide a calibrated feedback path for threshold recalibration. After each individual review, one would not want to update the BiMamba-KAN parameters, since noisy labels can be magnified when updating the system online. Rather, parameter updates may be performed after a minimum number of verified examples, and tested for a recent window of examples that is held out prior to deployment.

It is necessary to provide auditability if two reliability mechanisms are implemented together. The following items should be noted in the log: base probability, conformal prediction set, ADWIN state, decision threshold, and final moderation action. These records allow to remove doubt when the case is disputed later, whether it is the classifier, uncertainty calibration, concept drift handling or the moderation policy of the downstream process, which is the source of the problem. The separation is crucial for responsible evaluation since a correct classifier could lead to an undesirable platform action if the policy layer threshold is applied incorrectly.

J) Reproducibility and Evaluation Protocol

Train, calibrate and test partitions should be fixed in a reproducible implementation prior to hyperparameter selection. Final model artifact: The following parameters should be stored: byte-level vocabulary, Mamba state dimension, number of selective state-space blocks, forward/backward fusion gate, KAN grid size, spline order, dropout rate, optimizer parameters, and stopping criterion. The conformal calibration quantile needs to be calculated on a calibration partition only, and the parameters for the ADWIN needs to be selected prior to the final streaming evaluation. The use of a common test set for any of these options would lead to over-optimistic performance estimates and lessened validity of the reliability extensions.

Multiple random seeds can be provided for repeated training and training results for Accuracy, F1-score, MCC, and ROC-AUC can be reported. For the metrics derived from the confusion matrix, bootstrap intervals may be provided as well as conformal coverage, along with average prediction-set size and review rate. The evaluation in the drift experiment should include the drift point, the changed language distribution, the detection delay, the number of samples that are reviewed for recalibration, and the recovery accuracy. These facts help to differentiate between true robustness and the more favorable single split or hand selected drift scenario.

K) Privacy, Security, and Data-Governance Considerations

Cyberbullying moderation includes text that is generated by users and could include personal identifiers, sensitive disclosure or private conversational context. In a deployment, therefore, as much raw text as possible

should be removed after inference and operational moderation logs should be kept apart from research data. Consent to review or model maintenance should be limited, stored conversations should be role-based, and encrypted at rest. The conformal and drift extensions can be applied to compact probability, label and error statistics, which means that many reliability functions can be maintained without requiring to indefinitely preserve full message histories.

Drift monitoring needs to be thoroughly governed as the buffer of recent data may have new samples of abusive content. The information that needs to be stored in the buffer should be only for recalibration, and reviewed labels should be tied to versioned model and threshold identifiers. If a recalibrated threshold is used, the system shall capture the calibration window, target coverage level, ADWIN state and the percentage of reviewer agreement rate. These records ensure reproducibility and enable an investigation into whether a subsequent performance change was due to the core classifier, calibration update or a change in moderation policy.

L) Cross-Platform Generalization and Deployment Scope

There is a significant variation in language use between the different microblogging platforms, gaming forums, discussion boards, and educational sites. Changes in the input distribution can come from message length, reply structure, the use of emojis, and rules for moderation, as well as coded abuse. The byte-level representation and state-space encoder overcome the limitation of relying on a single vocabulary, but they still do not eliminate the need for external validation. To assess the trained model on a platform other than the one used for training, the system should undergo a deployment study that should include ordinary classification metrics and conformal coverage prior to any platform-specific recalibration.

The two reliability extensions are especially useful when transferring. If the new platform generates an unfamiliar probability pattern, split conformal prediction can be used to boost the review rate, and if the traffic starts after a sustained shift, ADWIN can detect it. If this change is temporary, the adaptive window can be recovered without replacing the classifier. If the shift persists, the system can be able to gather a verified recalibration set and, if required, fine-tune the model. This staged response is better than an instant retraining as this will differentiate a temporary shift in traffic from a permanent shift in language or user behavior.

V. CONCLUSION

A bidirectional Mamba selective state space modelling network was designed to build a cyberbullying detection pipeline, and a Kolmogorov-Arnold Network classification head was designed to classify the pipeline output. Architecture features state space sequence propagation and spline nonlinear classification. The proposed classifier achieves an accuracy of 97.6% and an F1 score of 97.6%, surpassing the accuracy of 96.0% for the reference model considered, and achieving a balanced precision and recall.

The architecture proposed is designed with selective state-space propagation for contextual sequence encoding, bidirectional processing to gain evidence from both directions of the sequence, and a spline-based KAN head for nonlinear classification. The ablation study suggests that both changes make a contribution to the overall improvement, and the confusion matrix suggests that the increased accuracy does not come with a significant increase in false positives or false negatives, but instead that the false positive and false negative numbers are comparable.

The first extension, split conformal prediction, adds calibrated set valued moderation and abstention in ambiguous conversations. The second extension, an ADWIN-based drift monitoring, detects the changes in the language distribution of the Internet and initiates re-calibration. The model should be tested on more platforms, with evaluation of multilingual and code-mixed abuse, repeated experimental runs using fixed data splits and statistical confidence intervals, and comparing alternative state-space variants, calibration strategies, and drift detectors in the context of streaming for long durations.

Future validation should also be able to report confidence intervals for the performance metrics; report calibration error; and test conformal coverage for important subgroups and conversation lengths separately. Long-duration experiments are required to verify the rate of ADWIN's ability to detect a slow slang drift and the amount of recent labeled data needed to maintain sound recalibration. More research can be done on smaller state space variations for mobile deployment, multilingual byte level tokenization, and privacy-preserving review pipelines in which sensitive user messages are reduced or redacted prior to storing for drift analysis.

REFERENCES

- [1] M. Dadvar, D. Trieschnigg, and F. de Jong, "A framework for cyberbullying detection in social media," in Proc. Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2013.
- [2] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in Proc. 10th Int. Conf. Machine Learning and Applications, 2011.

- [3] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment analysis for effective cyberbullying detection," *Communications in Information Science and Management Engineering*, 2012.
- [4] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proc. 3rd Int. Workshop on Socially-Aware Multimedia*, 2014.
- [5] C. Van Hee et al., "Automatic detection of cyberbullying in social media text," *PLoS ONE*, vol. 13, no. 10, 2018.
- [6] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. ECIR*, 2018.
- [7] H. Rosa, J. M. Gómez, and J. Carrillo-de-Albornoz, "A multi-faceted approach to cyberbullying detection," in *Proc. Iberian Conf. Pattern Recognition and Image Analysis*, 2019.
- [8] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, 2020.
- [9] T. Davidson, D. Warmesley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. ICWSM*, 2017.
- [10] A. Founta et al., "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proc. ICWSM*, 2018.
- [11] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Research Workshop*, 2016.
- [12] M. Zampieri et al., "Predicting the type and target of offensive posts in social media," in *Proc. NAACL-HLT*, 2019.
- [13] V. Basile et al., "SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. SemEval*, 2019.
- [14] 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.
- [15] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. EACL*, 2017.
- [16] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. ICML*, 2014.
- [17] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017.
- [18] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. EMNLP System Demonstrations*, 2018.
- [19] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv:2312.00752*, 2023.
- [20] Z. Liu et al., "KAN: Kolmogorov-Arnold Networks," *arXiv:2404.19756*, 2024.
- [21] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv:2107.07511*, 2021.
- [22] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [23] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *Proc. SIAM Int. Conf. Data Mining*, 2007, pp. 443–448.
- [24] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, 2014.
- [25] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, 2020.
- [26] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, 2015.
- [27] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017.
- [28] X. Chen et al., "Symbolic discovery of optimization algorithms," in *Advances in Neural Information Processing Systems*, 2023.
- [29] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, 2018.
- [30] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan, "Distribution-free, risk-controlling prediction sets," *arXiv:2101.02703*, 2021.