# EVALUATING SENTIMENT MODELS FOR CYBERSHIELD ABUSIVE LANGUAGE DETECTION SYSTEM

Binisa Giri [1], Hashmath Fathima [1], Kelechi Nwachukwu [2], Bikesh Regmi [3] and Kofi Nyarko [4]

[1] Co-primary authors
Department of Electrical and Computer Engineering, Morgan State University, Baltimore, USA

***ABSTRACT***

*Cyber Shield is an automated graph augmented abusive language and interaction detection system designed to identify harmful content including toxic interaction, hate speech, and general negative sentiment that is prevalent on social media platforms. As part of integrating a robust sentiment component into the system, we evaluated four widely used sentiment analysis models: BERT, RoBERTa, VADER, and Text Blob based on their complementary strengths and methodological diversity. BERT and RoBERTa represent string transformers architectures capable of capturing contextual meaning in noisy social media texts. VADER provides a lexicon based model optimized for informal online communication, offering a lightweight alternative to transformers. TextBlob is a traditional NLP baseline to benchmark improvements offered by more contemporary models. Together, this combination allows for a comprehensive comparison across model families, ensuring evidence-based model selection for the Cyber Shield project. These models were evaluated on a Kaggle dataset containing social media comments labeled with three sentiment classes (i.e., negative, positive, and neutral) serving as the ground truth. Each model's performance was measured using confusion matrices, accuracy, macro F1, weighted F1, and per class F1 scores. Our findings show that with an initial sample of 3000 texts, classical lexicon based models (i.e. VADER) and the traditional NLP baseline model (i.e., Text Blob), significantly outperformed transformer based models. TextBlob achieved the strongest performance results in this phase, underscoring the challenges of applying general pre-trained transformers to real world sentiment classification without domain specific fine tuning. However, after expanding the dataset to 18000+ samples per sentiment class and rerunning the evaluation with the updated RoBERTa sentiment model, the performance trend shifted. The updated RoBERTa model demonstrated substantial improvement and outperformed the earlier transformer results.*

***KEYWORDS***

*Abusive Language Detection, Sentiment Analysis, Transformer Models, Lexicon-based models, Social Media Moderation, Performance Metrics.*

## 1. INTRODUCTION

The CyberShield project aims to create an automated graph system for detecting abusive language and mitigating such interactions in online environments. Given the high volume and dynamic nature of user generated content, manual moderation is often impractical, prior work has shown that automated approaches combining both textual content and conversational structure

David C. Wyld et al. (Eds): CNSA, CCSEIT, NLAI, SIGML, ICCIOT, AIAP – 2026
pp. 69-75, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.160508

can significantly improve abuse detection performance [1]. While abusive language detection traditionally focuses on explicit hate speech or offensive expressions, sentiment signals provide additional contextual cues that can enhance classification robustness [2]. For example, negative sentiment frequently co-occurs with hostile or abusive language [3, 4] whereas positive sentiment can help distinguish non harmful messages [5].

To select an appropriate sentiment analysis model for integration into Cyber Shield, we conducted an empirical evaluation of four models spanning lexicon based, classical NLP and transformer based. The goal of this evaluation is to identify which model is best to detect sentiment patterns found in real world social media text and provides the most reliable performance when incorporated into the Cyber Shield pipeline.

## 2. LITERATURE REVIEW

Automated detection of abusive language on social media has been widely studied, often combining lexical, syntactic, and contextual features to identify harassment, hate speech, and toxicity [6].

More recent research highlights the value of incorporating emotion and sentiment signals into abusive language detection models. For example, Rajamanickam et al. [4] propose a multi-task learning framework that jointly models emotion and abuse detection, showing that affective features significantly improve abuse classification. Samghabadi et al. [3] similarly demonstrate that leveraging emotion-aware features helps capture both explicit and implicit abusive content.
Stylometric and emotion based features have also been found to be robust indicators of hate speech across languages and domains. Markov et al. [7] show that emotion expression, when combined with stylistic markers, improves cross-domain hate speech detection. More recently, Mnassri et al. [5] propose a multi-task model that uses a shared transformer encoder to jointly learn emotional representations and abusive/offensive language detection, leading to more reliable detection across datasets.

However, many of these studies do not compare different sentiment analysis methods such as lexicon based (e.g., VADER), classical NLP (e.g., TextBlob), and transformer based models on the same real world dataset using a three way sentiment taxonomy (positive, negative, neutral). That gap motivates our study by empirically evaluating models across these families. We aim to identify which sentiment model is best suited for integration into the CyberShield abusive language detection pipeline.

## 3. METHODOLOGY

This study evaluates four widely used sentiment analysis models: VADER, TextBlob, BERT and RoBERTa selected to represent three methodological families: lexicon based, classical NLP, and transformer based approaches. The objective is to determine which model is most suitable for integration into CyberShield's pipeline.

### 3.1. Models Evaluated

#### 3.1.1. VADER (Valence Aware Dictionary And Sentiment Reasoner)

VADER is a rule-based, lexicon-driven sentiment model, designed specifically for social media language including slang, emoticons/emojis, capitalization, punctuation, and other informal features common in short online posts. It uses a validated lexicon of roughly 7.5k tokens (words,

emoticons, acronyms, slang) with associated valence scores, and applies syntactic/grammatical heuristics (e.g., handling degree modifiers, negations, punctuation, capitalization) to infer overall sentiment intensity and polarity. It provides a lightweight, rule-based baseline to compare against heavy transformers models.

#### 3.1.2. Text Blob

Text Blob is a classical NLP tool that performs sentiment analysis using a combination of rule-based and probabilistic (e.g., Naïve Bayes–style) methods. It serves as a traditional baseline in many sentiment analysis applications. This was included to benchmark how a classical, easy-to-use tool compares against both lexicon-based and transformer-based models in our social media context.

#### 3.1.3. Bert

BERT (Bidirectional Encoder Representations from Transformers) was selected as a strong general purpose transformer baseline, known for its contextual embeddings and robust performance on diverse NLP tasks. It provides a benchmark for evaluating how well a standard transformer handles informal and abusive social media text.

#### 3.1.4. Robert A (Robust BERT)

RoBERTa is a robust, improved version of BERT that modifies the original pretraining regime to optimize performance on a wide range of NLP tasks. Because of its strong performance across many benchmarks, it is often regarded as a state-of-the-art transformer encoder. We include it to observe how a strong general transformer performs on sentiment classification especially without domain-specific fine-tuning. It helps determine whether a more optimized transformer yields better performance.

### 3.2. Data Set

To evaluate the model performance a publicly available sentiment annotated dataset from Kaggle named “Twitter Sentiment Analysis”[8] was used. This dataset contains tweets annotated with sentiment labels (i.e., positive, negative and neutral) for entity level sentiment classification. Each record includes a tweet, the associated entity, and a sentiment label, making it well suited for evaluating sentiment models on real social media content.

From this dataset, initially random samples of 3000 tweets were evenly distributed across the three primary sentiment categories (i.e., negative, positive, neutral) to form a balanced evaluation set. To assess the effect of increased sample size on model performance and to simulate a more realistic deployment scenario for CyberShield, the data size was expanded to 18,318 samples, again maintaining representation across sentiment classes. This expanded set provided a larger and more diverse sample of tweets for reevaluating model performance.

### 3.3. Evaluation Procedure

Each model received the same input datasets and produced predictions mapped to the three sentiment classes (i.e., negative, positive, neutral). For both the 3,000 and 18,318 versions of the sample dataset. For each we computed the following evaluation metrics:

- Accuracy
- Macro F1 Score

- Weighted F1 Score
- Per Class F1 Scores

These metrics help determine not only overall correctness but also how well each model distinguishes between the three classes.

## 4. Results And Discussion

### 4.1. 3000 Sample Evaluation

In the 3000 sample evaluation, lexicon based sentiment models outperformed pretrained transformer models on this dataset. VADER achieved moderate performance, with an accuracy of 44.8%, balanced macro F1 of 45.0% and per class F1scores of 42.8% (negative), 50.1% (positive), and 42.2% (neutral). TextBlob performed even better, achieving 48.9% accuracy and macro and weighted F1 of 49.0%, particularly excelling on negative (53.6%) and positive (55.5%) classes, though its neutral F1 remained relatively low (34.9%). These results align with the known strengths and limitations of lexicon based models, which are simple and fast but can struggle with nuanced language, sarcasm, or domain specific context [9].

In contrast, BERT performed poorly with 32.9% accuracy and macro F1 of 17.88%, heavily favoring the negative class (positive F1=4.47%, neutral=0.0%). Similarly, RoBERTA predicted only the positive class, yielding 33.3% accuracy and macro F1 of 0.167%, indicating a clear domain mismatch or checkpoint misalignment. Such performance issues are common when transformers models are applied out of the box on datasets outside their pretraining domain [10].

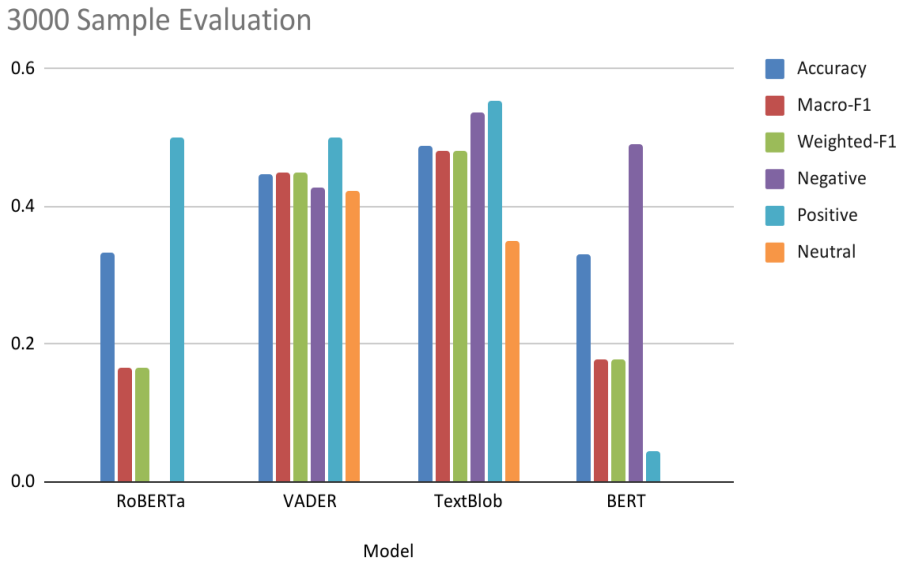



Figure 1: Accuracy, Macro-F1, Weighted-F1, Negative, Positive, and Neutral for 3000 Samples

Table 1. Performance Metrics for 3,000 Samples

| Model | Accuracy | Macro F1 | Weighted F1 | Positive F1 | Negative F1 | Neutral F1 |
|---|---|---|---|---|---|---|
| TextBlob | **0.4893** | **0.4799** | **0.4799** | **0.5546** | **0.5357** | 0.3495 |

| | | | | | | |
|---|---|---|---|---|---|---|
| VADER | 0.4480 | 0.4504 | 0.4504 | 0.5011 | 0.4278 | 0.4223 |
| ROBERTA | 0.3297 | 0.1788 | 0.1788 | 0.0447 | 0.4917 | 0.0000 |
| BERT | 0.3333 | 0.1667 | 0.1667 | 0.5000 | 0.0000 | 0.0000 |

Overall, these findings suggest that, despite the sophistication of pretrained transformers, simpler lexicon based models like VADER and TextBlob can provide more reliable sentiment predictions on domain specific or out of distribution datasets, especially under conditions of noise, limited data, or lack of domain specific fine tuning [11].

### 4.2. 18,000+ SAMPLE EVALUATION

As our first evaluation the result was from all the models were not even close to 50 %, so we reevaluated the models to validate the results, same metrics were used to calculated same dataset as well but the size was changed to 18,318 samples, RoBERTa emerged as the strongest overall model, this time achieving the highest accuracy (0.6483) and F1 scores, particularly excelling in identifying Positive( F1=0.7090) and Negative (F1=0.7335) sentiments, indicating a strong capability to capture polarized opinion. Notably, this improved performance was achieved using the updated cardiffnlp/twitter-roberta-base-sentiment-latest model [12] , which is specifically trained on social media data and better aligned with the character of the dataset. Despite its overall strength, RoBERTa exhibited relatively lower performance on the Neutral class (F1=0.4004), highlighting the ongoing difficulty in distinguishing neutral content from sentiment bearing text.

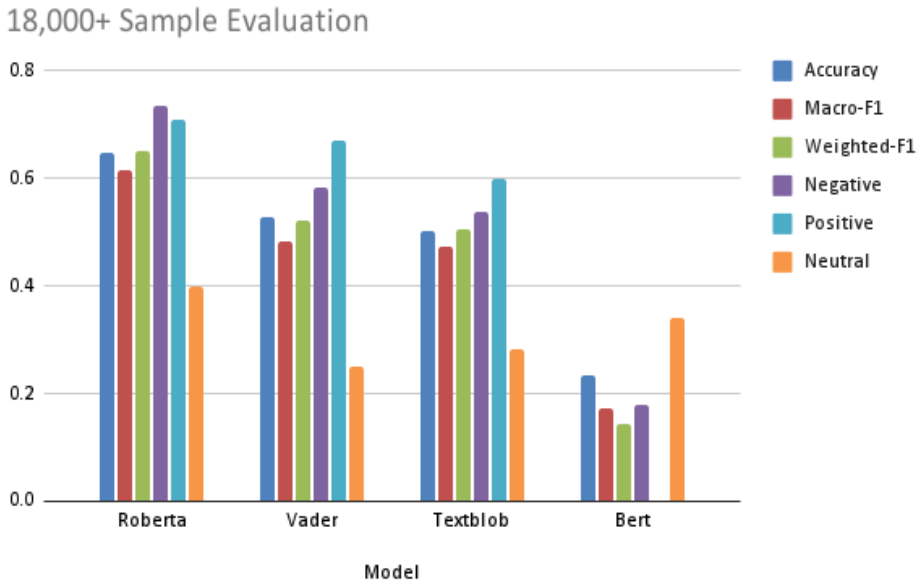



Figure 2: Accuracy, Macro-F1, Weighted-F1, Negative, Positive, and Neutral for 3000 Samples

Table 2. Performance Metrics for 18,000 Samples

| MODEL | ACCURACY | MACRO F1 | WEIGHTED F1 | POSITIVE F1 | NEGATIVE F1 | NEUTRAL F1 |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| RoBERTa | **0.6483** | **0.6143** | **0.6509** | **0.7090** | **0.7335** | **0.4004** |
| Vader | 0.5269 | 0.4826 | 0.5221 | 0.67117 | 0.5846 | 0.2515 |
| Textblob | 0.5040 | 0.4732 | 0.5058 | 0.5981 | 0.5388 | 0.2826 |
| Bert | 0.2326 | 0.1737 | 0.1449 | 0.0000 | 0.1790 | 0.3422 |

VADER demonstrated moderate effectiveness, especially for Positive and Negative sentiments, benefiting from its lexicon and rule based desing, but showed a pronounced weakness in Neutral classification (F1= 0.2515) reflecting a bias toward polarized sentiment. Textblob offered reasonable but weaker performance than VADER across all metrics, with limited capability in identifying Neutral sentiment (F1=0.2826).

In contrast, BERT performed poorly across nearly all evaluation metrics, with particularly severe failure in identifying Positive sentiment (F1= 0.0000). This suggests that the pretrained BERT model, without task specific fine tuning, is not well aligned with the characteristics of the dataset. Overall, the results highlight the superiority of transformer based models when appropriately aligned with the task, while also emphasizing the persistent challenge of accurately modeling Neutral sentiment across all approaches.

## ACKNOWLEDGEMENTS

The authors would like to express gratitude to everyone who contributed to the success of this work. Special thanks are extended to Morgan State University, CEAMLS (DEPA Lab) for their guidance and support throughout the project.

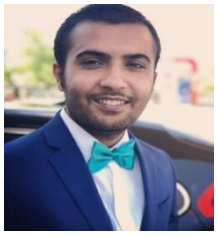
## REFERENCES

[1] M. Cécillon et al., "Abusive Language Detection in Online Conversations," 2019.
[2] K. Shanmugavadivel et al., "Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data," Scientific Reports, vol. 12, p. 21557, 2022.
[3] N. S. Samghabadi, A. Hatami, M. Shafaei, S. Kar, and T. Solorio, "Attending the emotions to detect online abusive language," in Proc. 4th Workshop on Online Abuse & Harms, 2020, pp. 79–88.
[4] S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova, "Joint modelling of emotion and abusive language detection," in Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4270–4279.
[5] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "HateBERT: Retraining BERT for abusive language detection in English," arXiv preprint arXiv:2302.08777, 2023.
[6] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," Information, vol. 13, no. 6, p. 273, 2022.
[7] Markov, N. Ljubešić, D. Fišer, and W. Daelemans, "Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection," in Proc. 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2021, pp. 149–159.
[8] J. P. Sharma, “Twitter Entity Sentiment Analysis,” Kaggle Dataset, 2020. [Online]. Available: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis.
[9] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Proc. ICWSM, vol. 8, no. 1, 2014, pp. 216–225.
[10] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?" arXiv preprint arXiv:1902.08648, 2019
[11] P. Nandwani and R. Verma, "Review on sentiment analysis and opinion mining," 2021. (Note: Please verify the publication venue for this source as it was missing from your bibliography).

[12] CardiffNLP, "twitter-roberta-base-sentiment-latest model card," Hugging Face. [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest (Accessed: Dec. 2025)
[13] R. Baly, G. Karadzhov, J. Glass, and P. Nakov, "We Can Detect Your Bias: Predicting the Political Ideology of News Articles," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.
[14] T. Caselli, V. Basile, E. Mitrofanova, and M. Sanguinetti, "HateBERT: Retraining BERT for Abusive Language Detection in English," arXiv preprint arXiv:2010.12472, 2021.
[15] S. Gururangan et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," in Proceedings of ACL 2020, 2020
[16] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment Analysis of Short Informal Texts," Journal of Artificial Intelligence Research, vol. 50, pp. 723–762, 2014.
[17] M. Shiny, "Sentiment Analysis and Offensive Language Detection in Social Media," i-manager's Journal on Computer Science, vol. 10, no. 2, pp. 1–7, 2022.

## AUTHORS

Binisa Giri is one of the primary authors for the Cyber Shield initiative, leading the conceptual development, technical narrative, and integration of research objectives across the project. Her work focuses on advancing cybersecurity resilience through interdisciplinary approaches that bridge cyberinfrastructure, risk mitigation, and community-oriented solutions. As lead author, Binisa coordinates cross-team contributions, translates complex technical concepts into clear and compelling research narratives, and ensures alignment with funding priorities and programmatic goals. She brings a strong commitment to rigor, clarity, and impact, with particular emphasis on secure systems, scalable architectures, and inclusive cyber defense strategies..



**Hashmath Fathima** is a core researcher and Project Manager for the REU/RET program, where she plays a central role in coordinating research activities, mentoring support, and program operations. She contributes to project planning, implementation, and assessment while ensuring alignment with research goals and educational objectives.
In her project management role, Hashmath oversees timelines, reporting, and supporting both undergraduate researchers and educators. Her work emphasizes effective collaboration, structured research experiences, and inclusive participation, helping translate research initiatives into impactful training outcomes.

**Bikesh Regm**i is a researcher at CEAMLS.