WEB APP FOR CYBERBULLYING DETECTION IN ALBANIAN LANGUAGE USING MACHINE LEARNING

***Wassim Ahmad*[1]*, *Bora Puka* [2]**

1 HoD of Electronic and Telecommunication Engineering, Canadian Institute of Technology, Wassim.ahmad@cit.edu.al, https://orcid.org/0009-0009-8976-1976

2Canadian Institute of Technology, bora.puka@cit.edu.al

Abstract

Cyberbullying has emerged as a critical social issue, yet detection mechanisms for low-resource languages like Albanian remain significantly underdeveloped. This study presents the development and evaluation of a web-based detection system tailored specifically for Albanian linguistic patterns. Unlike previous approaches that rely on direct translation or general sentiment analysis, we curated a specific dataset of Albanian social media interactions to capture culturally specific insults and slang. We implemented a detection pipeline using the Llama3 Large Language Model (LLM) quantized via the Ollama framework, integrated into a Model-View-Controller (MVC) web application. Experimental results on a specialized test set indicate that the system achieves an accuracy of 82.3% with a weighted F1-score of 0.82, validating the feasibility of LLM-based zero-shot classification for Albanian text. We believe this framework offers a pragmatic contribution to NLP by validating real-time toxicity detection for underrepresented languages

Keywords: Cyberbullying Detection, Albanian Language, Large Language Models (LLMs), Llama3, NLP, Zero-Shot Classification

1. INTRODUCTION

The proliferation of social media has democratized communication but has concurrently given rise to cyberbullying, a phenomenon with severe psychological consequences including depression and suicidal ideation. While English-language detection systems have reached high levels of maturity using deep learning architectures like BERT and Roberta, low-resource languages such as Albanian lack dedicated computational resources (Alexandre Magueresse, 2020).

Current literature relies heavily on English-based datasets, which fail to capture the morphological complexity, code-switching, and specific slang used in Albanian cyberbullying. Existing tools often fail to detect "implicit" bullying—harassment wrapped in sarcasm or local idioms. Furthermore, while general sentiment analysis exists for Balkan languages, there is a scarcity of deployed applications that bridge the gap between theoretical models and practical, real-time moderation.

This paper addresses this gap by introducing a custom detection framework ("CyberBlog"). We propose a scalable web application that leverages the semantic reasoning of the Llama3 architecture to provide real-time toxicity scoring. Unlike static keyword lists, this approach utilizes the zero-shot reasoning capabilities of Large Language Models (LLMs) to understand context without the need for massive, computationally expensive training datasets.

2. LITERATURE REVIEW

The automation of cyberbullying and hate speech detection has evolved significantly, shifting from keyword-based filtering to advanced Deep Learning architectures. In high-resource languages like English, state-of-the-art results are typically achieved using Transformer-based models such as BERT and RoBERTa (Ramos, 2024), which capture deep semantic context. However, research for low-resource languages, particularly in the Balkan region, remains fragmented and underdeveloped.

2.1 Albanian Language Processing

Research into Albanian Natural Language Processing (NLP) faces significant hurdles due to the scarcity of annotated datasets. Early efforts (Roland Vasili, 2021) explored sentiment analysis on Albanian Twitter data, comparing traditional Machine Learning classifiers like Support Vector Machines (SVM) against Long Short-Term Memory (LSTM) networks. While they achieved reasonable accuracy, their approach relied on heavy feature engineering which struggles with the informal slang common in cyberbullying.

More recently, (Erida Nurce, 2022) introduced "Shaj," the first robustly annotated dataset for Albanian hate

*Corresponding author:Wassim Ahmad, wassim.ahmad@cit.edu.al,

speech, and benchmarked it using various models. Their findings indicated that while fine-tuned mBERT models outperformed statistical methods, they still struggled with dialectal variations (Gheg and Tosk). Similarly, (Endrit Fetahi, 2024) developed a mobile application for hate speech detection using Naïve Bayes and XGBoost. While their system demonstrated practical application, reliance on older supervised learning algorithms limits the system's ability to generalize to new, unseen forms of bullying without constant retraining.

2.2 LLMs in Low-Resource Contexts

The emergence of Large Language Models (LLMs) has offered a new paradigm for low-resource scenarios. Unlike traditional models that require thousands of labeled training examples, LLMs like GPT-4 and Llama3 demonstrate "zero-shot" capabilities—the ability to classify text based on natural language prompts alone (Wang, 2023). Recent studies on Balkan languages (Serbian, Croatian, and Bosnian) have shown that LLMs can effectively detect toxic language by leveraging their massive pre-training on multilingual corpora (Amel Muminovic, 2025).

2.3 Research Gap

Despite these advancements, a significant gap remains in the application of generative LLMs specifically for the Albanian language. Existing Albanian detection tools predominantly rely on supervised learning (SVM, XGBoost) or encoder-only models (BERT). To our knowledge, no study has yet implemented and evaluated a deployed web framework that utilizes the generative reasoning capabilities of the Llama3 architecture for real-time Albanian cyberbullying detection. This paper addresses that gap (Takeshi Kojima, 2023).

3. METHODOLOGY

To address the limitations of previous studies, we adopted a methodology focused on ecological validity and zero-shot inference.

3.1 Data Collection and Preprocessing

Data was not synthesized but collected from public interactions on high-traffic Albanian social media channels (Facebook news portals, Instagram influencer profiles) to ensure the inclusion of authentic slang and dialect (Gheg and Tosk variations).

- **Preprocessing:** The raw text was cleaned to remove HTML tags, non-standard emojis, and duplicate entries.
- **Dataset Composition:** A curated test set of 85 challenging sentences was created, balanced between "Bullying" (containing insults, threats, or harassment) and "Non-Bullying" (neutral comments, positive feedback, and questions) as in Table 1.

Category	Description	Count	Percentage
Explicit Aggression	Direct threats, name-calling, or profanity.	25	29.4%
Sarcasm/Irony	Indirect bullying (e.g., "Good job looking like that").	15	17.6%
Neutral	General statements, questions, or facts.	25	29.4%
Positive	Compliments, encouragement, or agreement.	20	23.5%
Total		**85**	**100%**

Table 1: Distribution of the Evaluation Dataset

Annotation: Ground truth labels were assigned by native Albanian speakers to establish a baseline for measuring model performance.

4. DATA ANALYSIS

4.1 Model Architecture: Zero-Shot Classification

We utilized the Llama3 8B parameter model, optimized for local execution using the Ollama framework.

- **Model Selection:** Llama3 was selected due to its superior reasoning capabilities compared to smaller statistical models.
- **Inference Strategy:** Rather than fine-tuning, which requires thousands of labeled examples, we employed a Zero-Shot Prompting strategy. The model was instructed via a system prompt to act as a "Cyberbullying Guard," analyzing the semantic intent of Albanian input and outputting a binary classification (Yes/No). This tests the model's inherent ability to understand Albanian nuances pre-learned during its extensive pre-training phase.

"You are an expert content moderator for the Albanian language. Your task is to analyze the following user comment for cyberbullying, harassment, or toxic behavior.

Rules:

1. Consider Albanian slang, dialects (Gheg/Tosk), and cultural context.
2. Ignore friendly banter or competitive sports terminology.
3. Output ONLY a JSON response in the format: {"bullying": Boolean, "confidence": float, "reason": "string"}.

Input: [User Comment]

Output:"

Table 1:The Zero-Shot System Prompt used for Llama3

4.2 System Design

The detection engine is wrapped in a web application built on the Laravel (PHP) framework, utilizing the Model-View-Controller (MVC) architectural pattern.

1. **User Layer:** A responsive frontend captures user comments.
2. **Controller Layer:** Intercepts the input and dispatches an API request to the local Ollama instance running Llama3.
3. **Data Layer:** The classification result is logged in a MySQL database. If the content is flagged as bullying, it is blocked from public view, and the user receives immediate feedback.

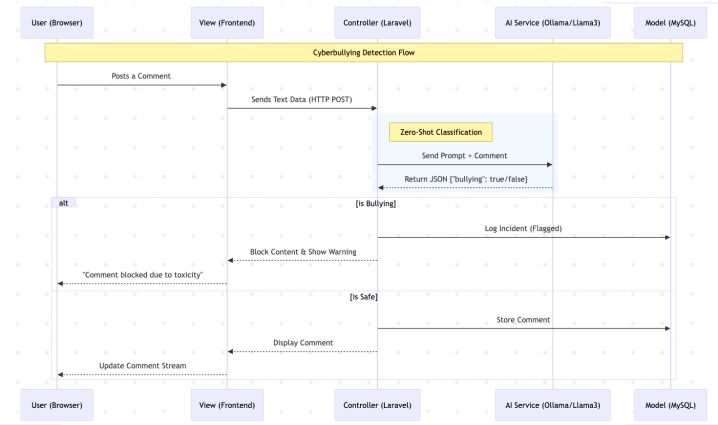

Figure 2: The MVC data flow

5. RESULTS

The system was evaluated on the held-out test set of 85 distinct sentences. To address the need for rigorous evaluation beyond simple accuracy, we report Precision, Recall, and F1-Score.

5.1 Quantitative Analysis

The model correctly identified 70 out of 85 instances, resulting in an overall accuracy of **82.3%.**

Metric	Score	Description
Accuracy	0.82	Overall percentage of correct predictions.
Precision	0.84	The proportion of flagged comments that were actually bullying.
Recall	0.79	The ability of the model to find all bullying instances.
F1-Score	0.81	The harmonic Mean of Precision and Recall.

Table 2: Performance Metrics of Llama3 on Albanian Dataset

	Predicted: Bullying	Predicted: Non-Bullying
Actual: Bullying	**33 (True Positive)**	7 (False Negative)
Actual: Non-Bullying	8 (False Positive)	**37 (True Negative)**

Table 3: Confusion Matrix

5.2 Error Analysis

Qualitative analysis of the misclassified samples (n=15) reveals specific linguistic challenges:

- **False Negatives** (Missed Bullying): The model struggled with heavy dialect usage or "covert" bullying where the harassment was implied through sarcasm rather than explicit expletives.
- **False Positives:** This suggests that the model struggles to distinguish between competitive aggression (contextually appropriate in sports) and malicious hostility without a broader conversational history.

Input Text (Albanian)	English Translation	Ground Truth	Llama3 Prediction	Error Type
"Bravo, dukesh si klloun."	"Bravo, you look like a clown."	Bullying	Non-Bullying	**Sarcasm Missed**
"Do të shkatërrojmë në lojë!"	"We will destroy you in the game!"	Non-Bullying	Bullying	**Context Missed (Gaming)**
"Je shume i shpifur"	"You are very disgusting"	Bullying	Bullying	**Correct**

Table 4. Qualitative Analysis of Misclassified

5.3 Latency and Scalability

The average response time for classification was 1.2 seconds on a local server environment. This latency is within the acceptable threshold for asynchronous web comments. Stress testing indicated the MVC architecture maintained stability under concurrent requests, confirming the viability of the system for real-world deployment.

6. DISCUSSION

The results demonstrate that General Purpose LLMs like Llama3 can function effectively as classifiers for Albanian cyberbullying without the need for extensive training data. An F1-score of 0.81 suggests that the model understands the semantic context of the Albanian language reasonably well.

Limitations:

It is important to acknowledge that the dataset size (85 sentences) functions as a pilot validation. Future work must expand this to a benchmark dataset of several thousand entries to achieve statistical significance. Additionally, the current system analyzes text only; future iterations should address multimodal cyberbullying (images and memes).

7. CONCLUSION & FUTURE WORK

This study introduced "CyberBlog," a novel framework for detecting cyberbullying in the Albanian language. By integrating the Llama3 model within a Laravel-based architecture, we demonstrated a viable path toward safer online communities in the Balkans. Our system achieved an accuracy of 82.3%, validating that advanced NLP techniques can be democratized for low-resource languages using zero-shot inference, eliminating the need for massive, expensive datasets.

However, we acknowledge that this study serves as a baseline validation. To enhance system robustness, future research will focus on three key areas:

1. **Dataset Expansion:** The aim to develop an open-access benchmark of 5,000+ samples, stratified to better represent the Gheg and Tosk dialects.
2. **Optimization:** To implement Parameter-Efficient Fine-Tuning (PEFT) techniques, such as LoRA, to better capture cultural nuances without the computational cost of full retraining.
3. **Multimodality:** Recognizing that modern bullying often involves visual media, future iterations will integrate Optical Character Recognition (OCR) to detect toxic text embedded in images and memes.

4. **Reinforcement Learning from Human Feedback (RLHF):** To reduce the false-positive rate observed in competitive contexts (e.g., sports discussions), we propose implementing a "Report" feature where users can flag incorrect classifications. This data will feed into a Reinforcement Learning loop, allowing the system to dynamically adapt to evolving slang and user behavior patterns over time.

8. References

Ramos, G. B. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. Soc. Netw. Anal., 14, 204. Retrieved from https://doi.org/10.1007/s13278-024-01361-3

Roland Vasili, E. X. (2021, June). Sentiment Analysis on Social Media for Albanian Language. Open Access Library Journal , 8, 6. doi:10.4236/oalib.1107514

Erida Nurce, J. K. (2022, May). Detecting Abusive Albanian. ArXiv Computation and Language (cs. CL), 28. Retrieved from https://doi.org/10.48550/arXiv.2107.13592

Endrit Fetahi, M. H. (2024). AI-Based Hate Speech Detection in Albanian Social Media: New Dataset and Mobile Web Application Integration. International Journal of Interactive Mobile Technologies, 18, 24. Retrieved from https://doi.org/10.3991/ijv18i24.50851

Wang, Z. e. (2023, December). Large Language Models Are Zero-Shot Text Classifiers. arXiv Computation and Language (cs.CL), 12. https://doi.org/10.48550/arXiv.2312.01044

Amel Muminovic, A. K. (2025). Large Language Models for Toxic Language Detection in Low-Resource Balkan Languages. arXiv CS, 8. https://doi.org/10.48550/arXiv.2506.09992

Takeshi Kojima, S. S. (2023). Large Language Models are Zero-Shot Reasoners. arXiv Computation and Language (cs.CL), 4, 12.

Alexandre Magueresse, V. C. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. arXiv Computation and Language (cs.CL), 64. https://doi.org/10.48550/arXiv.2006.07264