

Context-Aware Cyberbullying Detection in Chats

Using previous and current messages to understand potentially harmful language

Team members

Hiba Fathima T K (23MID0057)

Bharat Palanisamy Mettukadai (23MID0014)

Asad Irfan K (23MID0442)

Introduction

- Online chatting is a big part of everyday communication
- Cyberbullying through messages can hurt mental well-being and make online spaces unsafe
- Detecting it automatically is hard, because the meaning of a message depends on the conversation
- We use Natural Language Processing (NLP) to find potentially abusive messages

Main idea

Understand the current message using the previous message as context.

Problem Statement

Traditional detection often looks only at the words in one message, without the conversation around it.

Message 1

“I’m damn tired today.”

“damn” does not make this cyberbullying

Message 2

“You are useless.”

may be abusive, depending on context

- Problem: a system that checks each message alone can misread the meaning and predict wrongly
- Our approach: use both the previous and the current message

Objectives and Proposed System

1. Classify chat messages as SAFE or ABUSIVE
2. Build a baseline model that looks at the current message
3. Build a context-aware model that uses the previous and current messages
4. Compare the performance of both models
5. Create a simple web interface to show the predictions

Proposed system

The user enters the previous message and the current message. The system processes the text and gives a prediction from each model.

The two results are compared to see whether adding context changes the classification.

Who Can Use It and For What?

Potential users

- Social media platforms: help moderators find abusive messages
- Schools and universities: review cyberbullying reports in student chats
- Community moderators: decide which conversations need attention
- Researchers: study how context affects abuse detection

What it is used for

- Classifying current chat messages as SAFE or ABUSIVE
- Comparing message-only and context-aware predictions
- Spotting potentially harmful conversations
- Helping human moderators choose what to review

Important: the system supports human reviewers. It should not automatically punish users or make final judgments about their intentions.

Dataset and Preprocessing

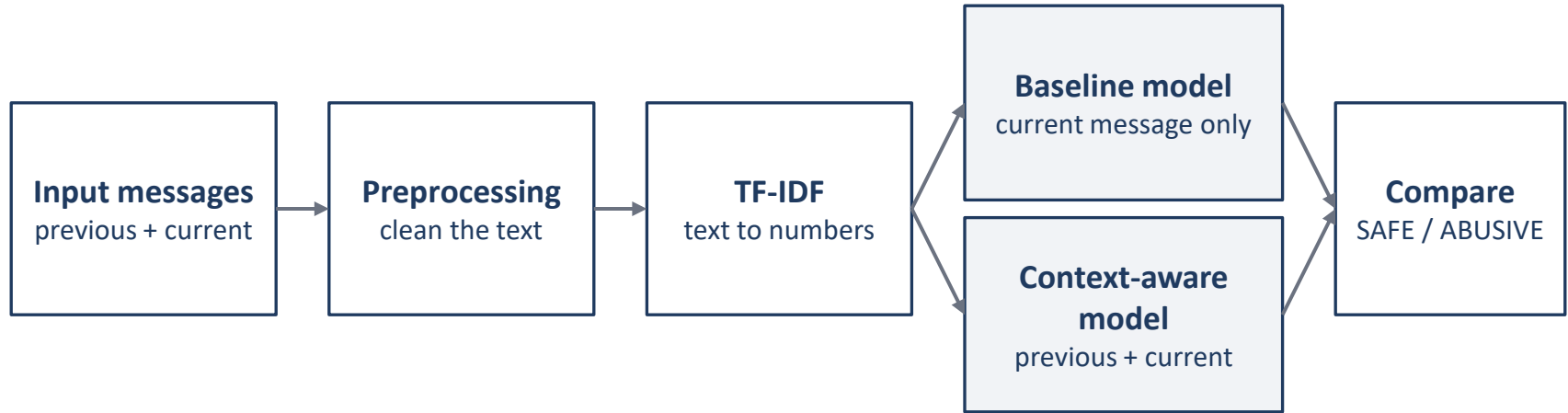
Dataset: CAD_context_aware.csv

1. Data loading: load the dataset in Python
2. Text cleaning: handle missing values, normalise the text
3. Data splitting: training set and testing set
4. Feature extraction: convert text to numbers using TF-IDF
5. Model training: train the baseline and context-aware classifiers

What is TF-IDF?

Term Frequency–Inverse Document Frequency. It turns text into numbers based on how important each word is in the dataset, so a machine-learning model can work with it.

System Architecture



Technologies used: Python, Pandas, NumPy, Scikit-learn (TF-IDF and classifiers), Flask for the local web app, HTML and CSS for the interface.

Evaluation and Results

Metric	Baseline	Context-aware
Accuracy	77.60%	78.82%
Abusive precision	44.97%	47.24%
Abusive recall	51.45%	46.47%
Abusive F1-score	47.99%	46.85%

Bold = better value in that row. Accuracy and precision improved a little with context, but recall and F1-score went down.

Understanding the Results

Interpretation

- The context-aware model has slightly higher accuracy
- Its precision for abusive messages is also higher
- Its recall and F1-score are slightly lower
- So adding context does not improve every metric
- The approach is promising, but it needs more work to detect abusive messages reliably

What the metrics mean

- Accuracy: percentage of all predictions that are correct
- Precision: of the messages predicted abusive, how many really are
- Recall: of the truly abusive messages, how many were found
- F1-score: a combined measure of precision and recall

Application Demonstration

Input example

Previous message

“You never listen to me.”

Current message

“You are so annoying.”

The application shows

- The baseline model’s prediction
- The context-aware model’s prediction
- A comparison of the two results

Predictions can still be wrong when the language is ambiguous, so the outputs are meant to help a human reviewer.

Limitations and Future Work

Current limitations

- Sarcasm and teasing are hard to tell apart from abuse
- The dataset may not cover every kind of real conversation
- The model gives false positives and false negatives
- Context did not improve all the metrics

Future work

- Improve recall for abusive messages
- Improve the quality of the dataset
- Test on more realistic conversations
- Improve the web interface and error handling
- Try better ways of using conversation context

Responsible use: the system should support human moderation. It should not be the only basis for penalizing someone.

Conclusion

- We explored context-aware cyberbullying detection using NLP
- A baseline model was compared with a model that uses the previous and current messages
- A simple web interface shows and compares the predictions
- Accuracy and precision improved slightly; recall and F1-score still need work
- Using conversation history is a useful direction, but more evaluation is needed

Thank You

Hiba Fathima T K | Bharat Palanisamy Mettukadai | Asad Irfan K