SafeSpeech: A Comprehensive and Interactive Tool for Analysing Sexist and Abusive Language in Conversations

Xingwei Tan[1,2], **Chen Lyu**[1], **Hafiz M. Umer**[1], **Sahrish Khan**[1], **Mahathi Parvatham**[1],
Lois Arthurs[4], **Simon Cullen**[4], **Shelley Wilson**[4], **Arshad Jhumka**[1,3], **Gabriele Pergola**[1]

[1]Department of Computer Science, University of Warwick, UK
[2]School of Computer Science, University of Sheffield, UK
[3]School of Computer Science, University of Leeds, UK
[4]Forensic Capability Network, UK

{xingwei.tan, chen.lyu, sahrish.khan, gabriele.pergola.1}@warwick.ac.uk
{lois.arthurs, simon.cullen, shelley.wilson}@dorset.pnn.police.uk
h.a.jhumka@leeds.ac.uk

Abstract

Detecting toxic language, including sexism, harassment, and abusive behaviour, remains a critical challenge, particularly in its subtle and context-dependent forms. Existing approaches largely focus on isolated message-level classification, overlooking toxicity that emerges across conversational contexts. To promote and enable future research in this direction, we introduce *SafeSpeech*, a comprehensive platform for toxic content detection and analysis that bridges message-level and conversation-level insights. The platform integrates fine-tuned classifiers and large language models (LLMs) to enable multi-granularity detection, toxic-aware conversation summarization, and persona profiling. *SafeSpeech* also incorporates explainability mechanisms, such as perplexity gain analysis, to highlight the linguistic elements driving predictions. Evaluations on benchmark datasets, including EDOS, OffensEval, and HatEval, demonstrate the reproduction of state-of-the-art performance across multiple tasks, including fine-grained sexism detection.[1]

1 Introduction

The dissemination of toxic content has become a pervasive issue across digital communication platforms. Toxic language, including sexism, harassment, and abusive behaviour, is not only widespread but also often amplified by the ease of dissemination in online environments. While explicit forms of harmful language, such as direct insults or threats, have been the focus of much research (Zampieri et al., 2019a; Vidgen et al., 2021), there is increasing awareness that toxicity frequently manifests in more subtle or context-dependent forms. Subtle toxic behaviours include

coercive language, controlling speech, and microaggressions (Wijanarko et al., 2024). These may seem benign in isolation but reveal their toxicity through recurring patterns in conversations, as observed in contexts such as violence against women and girls (VAWG). Recent advances, such as the EDOS dataset for fine-grained sexism detection (Kirk et al., 2023) and large language models (LLMs) capable of analyzing multi-turn dialogues (Suhara and Alikaniotis, 2024), highlight the potential for bridging this gap. Yet, existing approaches to toxic content detection mainly operate at the message level (Basile et al., 2019; Caselli et al., 2020; Mathew et al., 2020), using classifiers trained to identify the harmful language in isolated text (Camacho-collados et al., 2022; Wijanarko et al., 2024). And to date, there is no comprehensive platform that integrates message-level classification with conversation-level analysis to systematically evaluate and benchmark model performance across these dimensions.

Therefore, we introduce *SafeSpeech*, a unified platform designed for toxic content detection and analysis. *SafeSpeech* bridges the gap between message-level classification and conversation-level analysis, providing users with the tools to evaluate both explicit and context-dependent forms of toxicity. The platform supports a broad range of tasks, from granular single-message classification to the contextual understanding of multi-turn conversations. It leverages recent advances in dataset availability (Kirk et al., 2023; Basile et al., 2019; Zampieri et al., 2019b; Caselli et al., 2020) and large language models (LLMs) capable of processing extended sequences and capturing conversational context.

SafeSpeech incorporates prompt-based customization, enabling users to apply tailored *prompt*

---

[1]SafeSpeech Demo Website

templates for classification tasks. To better understand the models' prediction, *SafeSpeech* integrates a soft-explainability mechanism. This module, based on perplexity gain analysis (Suhara and Alikaniotis, 2024), identifies and visually highlights the phrases that most contributed to a classification output.

The platform extends its capabilities with a toxic-aware summarization module designed for long, multi-turn conversations. By combining semantic chunking with instruction-based summarization (Wang et al., 2023), this module identifies recurring or overlapping topics, ensuring that semantically related content is grouped together, even when interrupted by unrelated dialogue, and produces concise summaries. These summaries are conditioned on toxic classification outputs, to explicitly highlight harmful content. In addition to summarization, *SafeSpeech* includes a persona analysis module that goes beyond detecting toxic content to characterize individuals' behavioural tendencies. Existing methodologies offering automated persona analysis for conversations typically provide only general predictions (Han et al., 2023; Liu, 2024; Lyu and Pergola, 2024b; Wen et al., 2024). In contrast, our platform represents the first approach to leverage the Big Five Personality Traits framework (Costa and McCrae, 2008) and toxic-aware summaries to provide trait-based insights conditioned on toxic content classifications.

SafeSpeech has been thoroughly evaluated across diverse datasets and practical use cases. It provides state-of-the-art model performance within benchmark tasks for sexism, abusive, and hate speech detection, along with advanced tools for conversation summarization and persona analysis. The contributions of this paper can be summarized as follows:

1. *SafeSpeech*, an integrated platform combining message-level and conversation-level analysis, provides state-of-the-art classifiers and datasets for fine-grained and context-dependent forms of toxicity.

2. An AI assistant for dynamic exploration of the classified data via prompt-based interactions, and a soft-interpretability mechanism based on perplexity gain, enabling users to better trace model predictions.

3. Modules for toxic-aware conversation summarization and for persona analysis, facilitating
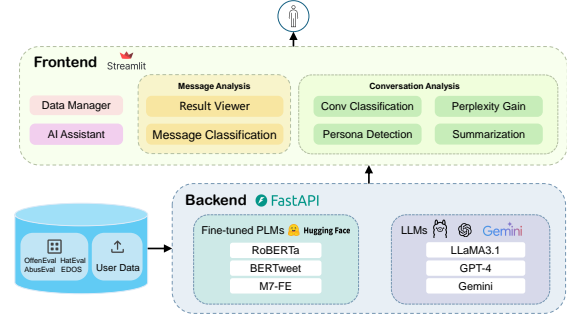

Figure 1: The structure of *SafeSpeech* platform.

in-depth examination of speaker behaviour and dynamics within multi-turn dialogues.

4. A comprehensive evaluation across benchmark datasets, demonstrating its state-of-the-art performance in identifying and analyzing nuanced toxic behaviours on public datasets.

2 Related Work

Existing approaches to toxic content detection mainly operate at the message level (Basile et al., 2019; Caselli et al., 2020; Mathew et al., 2020; Pergola et al., 2021; Tan et al., 2023, 2024; Lyu and Pergola, 2024a; Sun et al., 2024), which use classifiers to identify the harmful language in the isolated text without considering the complicated interactions in conversation settings. These systems commonly framed toxic content detection as sentence classification, then followed by extensive feature engineering and supervised training of statistical machine learning models or deep neural networks (Zhang and Luo, 2019; Zhu et al., 2021). Perspective API (Lees et al., 2022) is based on a BERT model pre-trained on "subword" sequences (UTF-8 bytes) and thus is capable of multilingual hate speech detection. Camacho-collados et al. (2022) provide a comprehensive analysis platform for tweets with language identification modules for hate and offensive language detection, but can only produce binary labels on sentence-level. Wijanarko et al. (2024) presents a system that can detect hate speech and identify its 5 subtypes, even though they only provide IndoBERTweet (Koto et al., 2021) as the backbone, which is trained only on Indonesia toxic speech data (Susanto et al., 2024).

3 *SafeSpeech* Platform

3.1 Platform Architecture Overview

SafeSpeech is designed as an interactive platform that combines data sources with analysis models,

including neural classifiers and large language models (Figure 1). Four core modules connect the data, analysis methods, and interpretation tools together:

- **Data Management** Host multiple toxic speech detection benchmark datasets. Manage user-uploaded data.

- **AI Assistant** Provide an LLM-based conversation interface, which is aided with carefully-designed prompts and classification labels, for interactive detection of toxic content. Users can also input custom prompts for detection and ask follow-up questions.

- **Message Analysis** Provide state-of-the-art fine-tuned classifiers on various toxic content detection subtypes, returning message-level predictions.

- **Conversation Analysis** Utilize LLMs for conversation-level toxic content detection. On top of the prediction results, we 1) provide perplexity gain to trace the source of the predictions, 2) label-aware conversation summarization, and 3) persona detection.

3.2 User Interaction Design

SafeSpeech is an interactive, multi-functional platform for the comprehensive analyses of toxic speech in messages and conversations. The frontend is developed using *Streamlit*[2], providing an intuitive, browser-based interface that requires no additional software installation. The main modules are available via a dedicated sidebar and allow the user to (i) upload and manage datasets, (ii) explore the results interactively via an AI assistant, and perform fine-grained analyses at both (iii) message and (iv) conversation levels (see Appendix A.4 for screenshots).

3.2.1 Data Manager

Users can upload their data files or select from available benchmarks on the *Data Manager* page. *SafeSpeech* includes widely-used toxic content detection benchmarks: the SemEval2023 EDOS (*sexism*) (Kirk et al., 2023), HatEval (*hateful*) (Basile et al., 2019), AbusEval (*abusive*) (Caselli et al., 2020), and OffensEval (*offensive*) (Zampieri et al., 2019b). The *Data Manager* automatically parses the structure of the user-uploaded file based on column names, preprocessing them for message-level or conversation-level analyses

---

3.2.2 Message Level Analyses

To investigate message-level data, *SafeSpeech* provides state-of-the-art models and a customizable workflow in terms of labels and prompt templates. Users can select from multiple fine-tuned classification models to perform analyses across different levels of label granularity. E.g., in the default case of sexism detection, the platform allows classification on three levels of granularity: binary (sexist, non-sexist), four-class, and eleven-class subcategories (Kirk et al., 2023). Alongside fine-tuned models, users can enable the use of LLMs, which will automatically apply customized or predefined prompts to verify the classification results (See Appendix A.4 for screenshots). Additionally, users can process to the *Result Viewer* page for more visualization options.

3.2.3 Conversations Level Analyses

To investigate conversation-level data, the platform provides several components to (i) highlight relevant excerpts of text, conduct (ii) topic-aware summarization, and (iii) persona profiling.

Perplexity Gain To identify the most relevant phrases determining the message classification, the platform relies on the analyses of the *Perplexity Gain* (Suhara and Alikaniotis, 2024). This statistical method quantify the contribution of individual phrases to the model's output (i.e., classification) by measuring the changes in perplexity when a particular sentence is removed from the input. The results are visualised as a heatmap over text, where darker shades denoted higher relevance. Figure 2 demonstrates an example on a synthetically generated conversation, where the LLM give a positive.

Toxic-aware Speaker Summaries The *Summarization* component enables users to track the progression of multi-turn, multi-topic toxic conversations. First, long conversations are visualised to the users into semantically coherent chunks, ensuring that recurrent topics discussed at different time intervals are logically grouped. For example, in a conversation where a topic like "workplace harassment" is initially discussed, interrupted by unrelated small talk, and then revisited later, the system identifies both segments as semantically related and groups them together. The summaries are generated for each segment, preserving each speaker's perspectives, and distinguishing contributions from different participants. Additionally, the summarization process incorporated the message-level classifications as conditional inputs (*toxic-aware sum-*
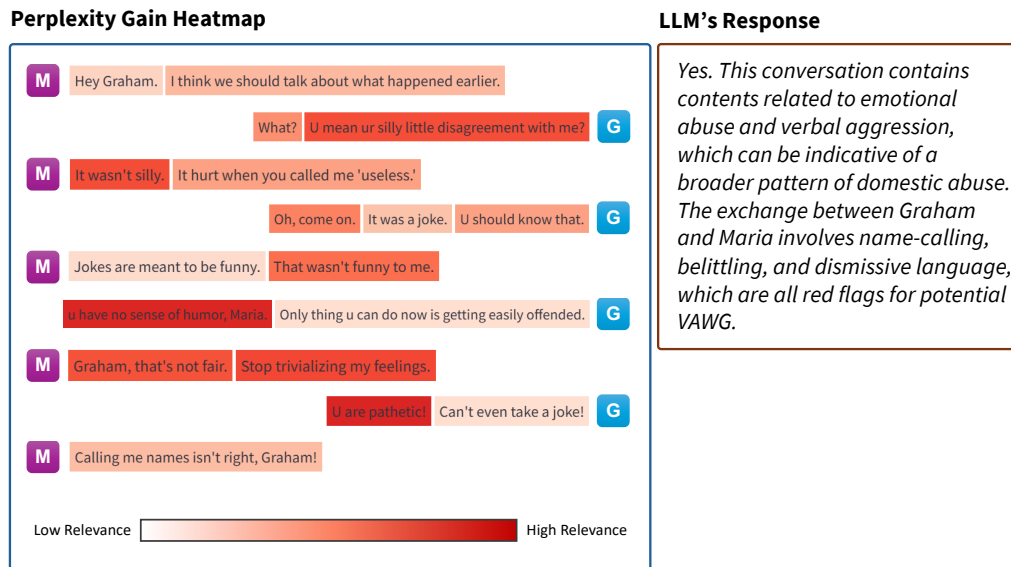
Perplexity Gain Heatmap

LLM's Response

Yes. This conversation contains contents related to emotional abuse and verbal aggression, which can be indicative of a broader pattern of domestic abuse. The exchange between Graham and Maria involves name-calling, belittling, and dismissive language, which are all red flags for potential VAWG.

(M) Hey Graham. I think we should talk about what happened earlier.

What? U mean ur silly little disagreement with me? (G)

(M) It wasn't silly. It hurt when you called me 'useless.'

Oh, come on. It was a joke. U should know that. (G)

(M) Jokes are meant to be funny. That wasn't funny to me.

u have no sense of humor, Maria. Only thing u can do now is getting easily offended. (G)

(M) Graham, that's not fair. Stop trivializing my feelings.

U are pathetic! Can't even take a joke! (G)

(M) Calling me names isn't right, Graham!

Low Relevance ▬▬▬ High Relevance

Figure 2: An example of perplexity gain analysis on a synthetically generated conversation. Based on the response, we computed perplexity gains and represented them as a heatmap.

maries). This is to highlight abusive messages within the condensed summaries and ensure that toxic or abusive content remains identifiable, a feature particularly relevant when dealing with subtle behaviours such as coercive control, backhanded compliments, or escalation patterns. To the user, intermediate results are visualised along with a progress bar, indicating the summarization is still in progress. Once processed, summaries generated for each semantically related chunk will be presented and grouped by participants to make it easier for users to navigate long conversations, understand the dynamics of the dialogue, and gain insights into how topics evolve (See Appendix A.4 for screenshots).

Persona Analysis The *Persona Analysis* component predicts the speaker's persona using the Big Five Personality Framework: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Costa and McCrae, 2008). Unlike general-purpose personality prediction approaches (Wen et al., 2024; Han et al., 2023; Liu, 2024), this *SafeSpeech*'s component incorporates toxic-aware conditional summaries to focus on relevant toxic contexts. The module outputs numerical scores for each trait and explanatory justifications, linking linguistic features observed in the summaries to the predicted traits[3]. This informa-

tion is presented via interactive radar plots, together with a detailed explanation of each trait, presenting a compressive analysis of each speaker's persona (Figure 3). For example, high scores in *Neuroticism* might be associated with frequent emotional instability or escalation in toxic exchanges, while lower *Agreeableness* could signal antagonistic tendencies in abusive or coercive contexts.

3.2.4 AI Assistant

This module provides a chat-style interface where users can flexibly interact with the classified data via LLMs. This allows users to input prompts, request clarifications, or generate responses based on the classifier's prediction. The responses from the LLM are displayed via streaming, facilitating an iterative exploration process that helps refine the understanding of detected content. Users can also opt to apply predefined prompts or create custom ones, making this module adaptable for users with different levels of familiarity with prompt engineering. The interaction history can then be conveniently downloaded.

3.3 Backend Implementation

The backend of *SafeSpeech* is designed to efficiently combine multiple data sources with fine-tuned neural classifiers and state-of-the-art LLMs.

---

[3]The Persona Analysis module is intended for research purposes only. The generated profiles are exploratory and not definitive. Results rely on automated models that may reflect biases and inaccuracies inherent in the data or methods. They should not be used as a substitute for professional evaluation or in operational settings, but as a starting point to explore the limitations and potential of automated personality profiling in a controlled research environment.
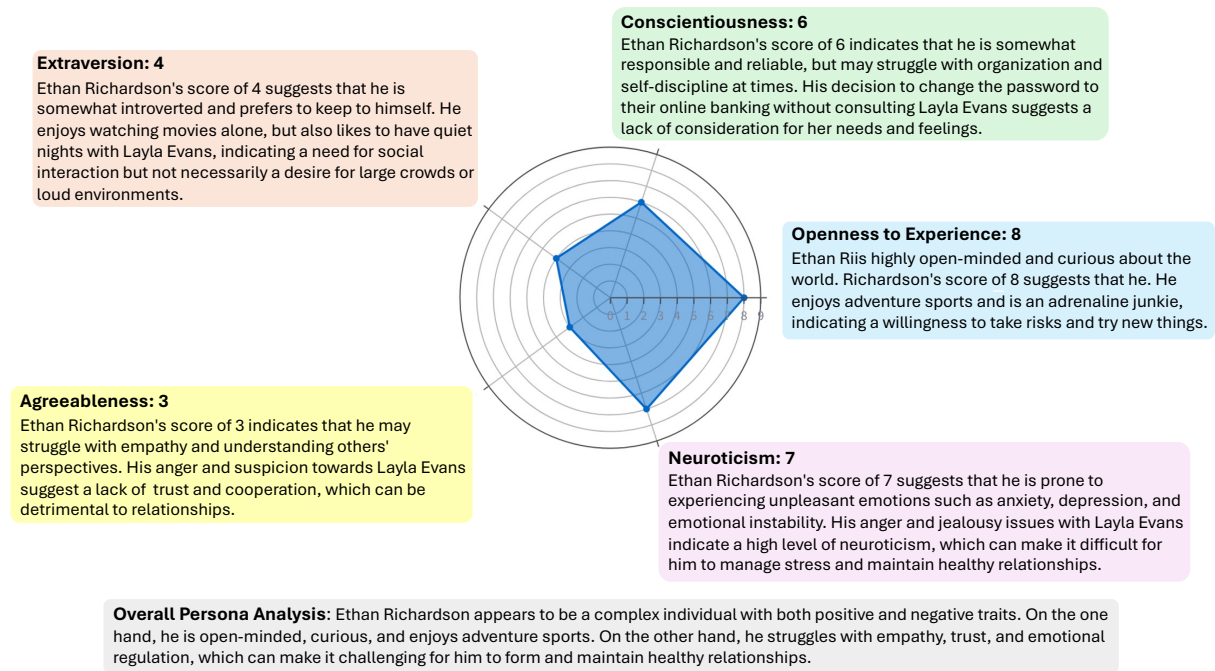
Conscientiousness: 6
Ethan Richardson's score of 6 indicates that he is somewhat responsible and reliable, but may struggle with organization and self-discipline at times. His decision to change the password to their online banking without consulting Layla Evans suggests a lack of consideration for her needs and feelings.

Extraversion: 4
Ethan Richardson's score of 4 suggests that he is somewhat introverted and prefers to keep to himself. He enjoys watching movies alone, but also likes to have quiet nights with Layla Evans, indicating a need for social interaction but not necessarily a desire for large crowds or loud environments.

Openness to Experience: 8
Ethan Riis highly open-minded and curious about the world. Richardson's score of 8 suggests that he. He enjoys adventure sports and is an adrenaline junkie, indicating a willingness to take risks and try new things.

Agreeableness: 3
Ethan Richardson's score of 3 indicates that he may struggle with empathy and understanding others' perspectives. His anger and suspicion towards Layla Evans suggest a lack of trust and cooperation, which can be detrimental to relationships.

Neuroticism: 7
Ethan Richardson's score of 7 suggests that he is prone to experiencing unpleasant emotions such as anxiety, depression, and emotional instability. His anger and jealousy issues with Layla Evans indicate a high level of neuroticism, which can make it difficult for him to manage stress and maintain healthy relationships.

Overall Persona Analysis: Ethan Richardson appears to be a complex individual with both positive and negative traits. On the one hand, he is open-minded, curious, and enjoys adventure sports. On the other hand, he struggles with empathy, trust, and emotional regulation, which can make it challenging for him to form and maintain healthy relationships.

Figure 3: An example of persona analysis generated by Llama3.1 based on the conversation summary and the Big Five personality traits framework.

It is based on *FastAPI*[4], a modern web framework that facilitates real-time communication between the frontend and backend, ensuring smooth data processing, model inference, and user interaction. The neural classifiers are implemented using *PyTorch* and Huggingface's *transformers*. The LLMs are managed by *Ollama* or via their publicly available APIs (e.g., OpenAI and Gemini). Below, we describe how each module's methodologies, models, and tools are integrated into the backend.

AI Assistant *SafeSpeech* enables interactive exploration of toxic content detection by integrating LLMs through API-based communication with a conversation interface. Backend requests are managed via FastAPI, which routes user-defined or predefined prompts to the selected LLM for response generation. Open-weight LLMs, such as *Llama3.1-70B-instruct*, are deployed using *Ollama*, leveraging *llama.cpp* for efficient model loading and inference. Users can flexibly choose LLMs based on model size, quantization levels, or available external APIs, such as OpenAI and Gemini.

Message Level Analysis To produce message-level predictions, we utilize Transformer-based pre-trained language models, specifically *DeBERTa-v3-large* (He et al., 2021), *RoBERTa-Large* (Liu et al., 2019), and *Mistral 7B* (Jiang et al., 2023).

These models are supervised fine-tuned on the respective task-specific datasets. Moreover, *SafeSpeech* integrates a voting-based ensemble model that combines the strengths of *DeBERTa-v3-large*, *RoBERTa-Large*, and *Mistral 7B*, called *M7-FE*, (Khan et al., 2024) which outperforms existing models on sexist detection of English tweets.

Conversations Level Analysis For the analyses of toxic conversations, we designed a backbone of LLMs, connecting the LLMs' APIs to the uploaded conversation files and to their message classifications at the conversation level via prompts. This analysis is a prerequisite for *Perplexity Gain*, *Summarization*, and *Persona Analysis*.

Perplexity Gain The contribution of individual phrases to the model's generated results are computed via perplexity gain (Suhara and Alikaniotis, 2024). It was originally proposed as a source identification approach for abstractive summarization, and it is based on the assumption that a language model should be more perplexed (i.e., less confident) to generate the same output (e.g., summary, or classification label) if an important sentence is removed. In the analyses of conversations, when the LLM predicts that the input contains toxic contents, we adapted the perplexity gain to highlight the phrases that contribute the most to this conclusion. In particular, the method first generates an

---

[4]https://github.com/fastapi/fastapi

output Y from the complete input X, which consists of N sentences $(s_1, ..., s_i, ..., s_N)$. Then, for each sentence s_i in the input X, compute the perplexity of the original output Y given an input X in which the sentence s_i is removed. The difference between the initial perplexity and the perplexity after removing the sentence is used as a measure to evaluate how relevant the sentence s_i is to the output Y. The score is defined as:

$$R(s_i, Y|X; \theta) = PPL(Y|X_{\setminus s_i}; \theta) - PPL(Y|X; \theta) \quad (1)$$

A heatmap visualizes perplexity gains, with darker shades indicating sentences most critical to the output. However, it is worth noticing that for negative predictions, perplexity gains are less informative as they rely on the entire conversation.

Toxic-aware Conditional Summarization Our backend facilitates the generation of summaries for conversations among specific groups of participants. It first identifies and organizes messages belonging to the same conversation by matching the list of dialogue participants. Then, given that conversational flows often span multiple topics and temporal ranges, summarization is performed on utterances discussing the same topic (Pergola et al., 2019). This process requires segmenting lengthy conversations into semantically coherent chunks. To achieve this, we employ *semantic chunking*[5], a method that adaptively determines segmentation points based on cosine similarity between sentence embeddings, generated using the *sentence-transformers*[6] library. We customized *semantic chunking* to enhance its performance in processing conversational data. Once the conversation is segmented, we employ an instructional summarization model, InstructDS (Wang et al., 2023), to generate summaries for each chunk. InstructDS is a state-of-the-art instruction-following dialogue summarization model fine-tuned on Flan-T5-XL (Chung et al., 2022). The segmented content is input into the model with specific instructions to generate summaries considering messages previously categorized as toxic. The final summaries, categorized by segmented chunks, are transmitted to the frontend for display.

Summary-based Persona Analysis This module leverages dialogue summaries to predict speaker

personas based on the Big Five personality traits framework (Costa and McCrae, 2008). The backend of this module involves three key components: summary concatenation, persona prediction, and output generation. First, all relevant toxic-aware summaries generated from prior stages are concatenated. This provides a comprehensive context for a speaker's linguistic style, behavioural patterns, and interpersonal dynamic. By leveraging summaries instead of lengthy raw messages, the method not only minimizes information loss on critical toxic-related patterns or utterances, but also ensures computational efficiency. The concatenated summaries are subsequently processed by a persona prediction model, instructed via prompt to identify linguistic and behavioural patterns based on the Big Five personality traits theory. Detailed prompts are shown in the Appendix 5. Due to data availability constraints, we assume for simplicity and to reduce sparsity, a consistent persona across all messages for each speaker. This assumption enables the generation of a unified and cohesive persona profile for each individual within the dialogue.

4 Evaluation

4.1 Quantitative Analysis

We evaluate the performance of *SafeSpeech*'s message-level classification models across widely-used benchmarks for toxic content detection. Table 1 reports results on benchmark datasets including EDOS - SemEval 2023 (Kirk et al., 2023), OffensEval 2019, AbusEval, and HatEval.

For sexism detection (EDOS), we evaluate classifiers on three levels of granularity: binary detection (Subtask A), category-level classification (Subtask B), and fine-grained vector-level classification (Subtask C). *SafeSpeech* integrates state-of-the-art models, including our ensemble method **M7-FE** (Mistral-7B Fallback Ensemble). The results demonstrate that **M7-FE** achieves top performance across all subtasks, improving over the winning system (Zhou, 2023) by 1% F1 on Subtask A and 4% F1 on Subtask C. On Subtask B, **M7-FE** matches the best reported performance while outperforming publicly available baselines such as *ISEGURA/roberta*[7] by 14% F1. **M7-FE** outperforms *NLP-LTU/bertweet*[8] by 2% and prompting with the 4-bit quantised version of Llama3.1.

For hate speech, abusive speech, and offen-

Dataset / Subtask	Model	Macro F1
	HateBERT_offeneval	0.53
	HateBERT_abuseval	0.59
	HateBERT_hateval	0.66
	llama3.1-4bit (in-context learning)	0.60
EDOS Subtask A (Binary Sexism Detection)	*NLP-LTU/bertweet*	0.86
	PingAnLifeInsurance (Zhou, 2023)	0.87
	M7-FE (Our)	**0.88**
	ISEGURA/roberta	0.58
EDOS Subtask B (Category Detection, 4 Classes)	PingAnLifeInsurance (Zhou, 2023)	0.72
	M7-FE (Our)	**0.72**
	ISEGURA/roberta	0.43
EDOS Subtask C (Vector Detection, 11 Classes)	PingAnLifeInsurance (Zhou, 2023)	0.56
	M7-FE (Our)	**0.60**
	M7-FE (Our)	0.48
OffensEval 2019	Llama3.1-4bit (in-context learning)	0.67
	BERT	0.80
	HateBERT	**0.81**
	M7-FE (Our)	0.49
AbusEval	Llama3.1-4bit (in-context learning)	0.65
	BERT	0.73
	HateBERT	**0.77**
	M7-FE (Our)	0.51
HatEval	Llama3.1-4bit (in-context learning)	**0.59**
	BERT	0.48
	HateBERT	0.52

Table 1: The test performance of the available models on the benchmark datasets on *SafeSpeech*. BERT and HateBERT results are obtained from Caselli et al. (2021).

sive speech classification, we evaluate our models on the OffensEval, AbusEval, and HatEval benchmarks. Notably, *HateBERT* (Caselli et al., 2021), a model specifically designed for this domain, achieves top-ranking performance on the OffensEval 2019, AbusEval, and HatEval datasets. Specifically, *HateBERT* demonstrates the highest F1 scores on OffensEval 2019 and AbusEval when compared to baselines such as vanilla BERT and a 4-bit quantized Llama 3.1 model. Conversely, the 4-bit Llama 3.1 model exhibits superior performance on the HatEval dataset. Additionally, we assess the M7-FE model, trained on the EDOS dataset, to evaluate its performance in slightly out-of-domain settings across the same benchmarks. While M7-FE does not surpass the performance of Llama 3.1 or supervised fine-tuned BERT models, it achieves comparable results on HatEval, suggesting a significant overlap between sexism and hate speech as represented in these datasets.

5 Conclusion and Future Work

SafeSpeech is the first platform to provide tools for conveniently analysing hate speech or violence against women and girls in a conversational format. It connects with state-of-the-art supervised fine-tuned models and large language models, offering a comprehensive and interactive interface for researchers to test their new ideas without the need for coding. In future work, we will try to expand the platform and make it easy to use for the general public, making it an accessible tool for social media moderation.

Limitations

Bias Analysis Our platform has not analyzed the bias analysis of LLMs on detecting toxic content. Social media content makes up for a big portion of the pre-training data of the LLMs. Although most of the LLMs include filtering about toxic content when selecting their training data, there may still be implicit bias in the training data.

Model Transparency   Although we try to generate explanations and perplexity gains for the classifications, the transparency of the classifiers and LLMs remains an area for further improvement. The generated explanations are external references instead of a direct interpretation of the classifiers' internal mechanism. The perplexity gain was originally designed for abstraction summarization, and its effect on toxic content detection still requires more examination. The generated personality traits are predicted merely based on the context correlation between the inputs and the prompt, it has not been clinically validated and the output should not be used for professional diagnosis. Despite all these limitations, we still think the *SafeSpeech* platform would promote the exploration of toxic content detection by providing convenient access to these functions.

Model Variety   The demo currently provides a limited selection of models due to computational constraints. However, users can load additional models from the HuggingFace Model Hub, subject to size restrictions. Additionally, users can also integrate the use of external LLMs via their API keys.

Ethics Statement

Data Security and Privacy   The current platform is a prototypical demo which has not implemented extensive cybersecurity measures. The benchmarking data used in our platform are from publicly available datasets which do not contain confidential data. We recommend that users only upload non-sensitive data that can be publicly shared for research purposes.

The Risks in Monitoring Hate Speech   This paper focuses on toxic content detection and analysis, including sexism, racism, and xenophobic language etc. However, the definitions of such content are hugely controversial and are changing rapidly over time. The actions of monitoring and censoring hate speech without accurate and socially agreed definitions might bring risks of limiting the freedom of speech. Moreover, the training of LLMs and other neural models might already incorporate social biases which might affect the classification and explanation of the detection system. An additional risk is that the monitoring and censoring of toxic content might shift the form of such content and make understanding it more challenging.

References

Enrique Amigo and Agustín Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi

Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198.

Manuel Guerrero-García, Manuel Cerrejón-Naranjo, Jacinto Mata Vázquez, and Victoria Pachón Álvarez. 2024. I2C-UHU at EXIST2024: learning from divergence and perspectivism for sexism identification and source intent classification. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 1026–1042. CEUR-WS.org.

Wenjing Han, Yirong Chen, Xiaofen Xing, Guohua Zhou, and Xiangmin Xu. 2023. Speaker-aware hierarchical transformer for personality recognition in multiparty dialogues. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Sahrish Khan, Gabriele Pergola, and Arshad Jhumka. 2024. Multilingual sexism identification via fusion of large language models. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 1070–1079. CEUR-WS.org.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021

Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

J. Liu. 2024. Personality prediction from text. https://github.com/jcl132/personality-prediction-from-text. GitHub repository, accessed December 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Chen Lyu and Gabriele Pergola. 2024a. SciGisPy: a novel metric for biomedical text simplification via gist inference score. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 95–106, Miami, Florida, USA. Association for Computational Linguistics.

Chen Lyu and Gabriele Pergola. 2024b. Society of medical simplifiers. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*.

Gabriele Pergola, Lin Gui, and Yulan He. 2019. Tdam: A Topic-Dependent Attention Model for Sentiment Analysis. *Information Processing & Management*, 56(6):102084.

Gabriele Pergola, Lin Gui, and Yulan He. 2021. A disentangled adversarial neural topic model for separating opinions from plots in user reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2870–2883, Online. Association for Computational Linguistics.

Yoshi Suhara and Dimitris Alikaniotis. 2024. Source identification in abstractive summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 212–224, St. Julian's, Malta. Association for Computational Linguistics.

Zhaoyue Sun, Gabriele Pergola, Byron Wallace, and Yulan He. 2024. Leveraging ChatGPT in pharmacovigilance event extraction: An empirical study. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–357, St. Julian's, Malta. Association for Computational Linguistics.

Lucky Susanto, Musa Izzanardi Wijanarko, Prasetia Anugrah Pratama, Traci Hong, Ika Idris, Alham Fikri Aji, and Derry Wijaya. 2024. Indotoxic2024: A demographically-enriched dataset of hate speech and toxicity types for indonesian language.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2023. Event temporal relation extraction with Bayesian translational model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1125–1138, Dubrovnik, Croatia. Association for Computational Linguistics.

Xingwei Tan, Yuxiang Zhou, Gabriele Pergola, and Yulan He. 2024. Set-aligning framework for autoregressive event temporal graph generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3872–3892, Mexico City, Mexico. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023. Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore. Association for Computational Linguistics.

Zhiyuan Wen, Jiannong Cao, Yu Yang, Ruosong Yang, and Shuaiqi Liu. 2024. Affective-nli: Towards accurate and interpretable personality recognition in conversation. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 184–193. IEEE.

Musa Izzanardi Wijanarko, Lucky Susanto, Prasetia Anugrah Pratama, Ika Karlina Idris, Traci Hong, and Derry Tanti Wijaya. 2024. Monitoring hate speech in Indonesia: An NLP-based classification of social media texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 142–152, Miami, Florida, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Mengyuan Zhou. 2023. PingAnLifeInsurance at SemEval-2023 task 10: Using multi-task learning to better detect online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2188–2192, Toronto, Canada. Association for Computational Linguistics.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.

A Appendix

A.1 Experiment Results on EXIST

CLEF 2024 sEXism Identification in Social neTworks (EXIST)[9] is another competition on sexism content detection. Mistral-7B Fallback Ensemble (M7-FE) (Khan et al., 2024), which our team members develop, ranks first among 68 submitted models on English test data with an F1 of 0.76. Dual-Transformer Fusion Networks (DTFN) (Khan et al., 2024), which is also developed by our team members, ranks second among the 68. I2C-UHU (Guerrero-García et al., 2024) ranks third in the leaderboard. Table 1 shows the official results[10] of the test set evaluation. The ICM metric, proposed by (Amigo and Delgado, 2022), is based on information theory and measures the similarity between system classifications and gold standard labels. The normalized version, ICM-Hard Norm, takes the label imbalances into account. Higher values of ICM and ICM-Hard Norm indicate a stronger alignment between system outputs and the ground truth.

Model	ICM-HARD	ICM-Hard Norm	F1
I2C-UHU	0.58	0.80	0.76
DTFN *(Our)*	0.60	0.80	0.75
M7-FE *(Our)*	**0.62**	**0.82**	**0.76**

Table 2: The performance of the models on the EXIST task 1 English dataset.

A.2 Prompt Templates

Default VAWG content detection prompt

Instruction: Does the given conversation contain content about violence against women and girls?
VAWG definitions:
The followings are types of violence against women and girls:
1) Domestic Abuse; 2) stalking and harassment; 3) controlling or coercive behaviour; 4) child abuse; 5) rape; 6) honour based abuse; 7) forced marriage; 8) human trafficking, smuggling and slavery; 9) female genital mutilation; 10) prostitution.
Conversation:
"""{conversation}"""

Table 3: An example of the classification prompt when the toxic type is violence against women and girls.

Default sexism content detection prompt

Instruction: Does the given conversation contain content about sexism?
Sexism definitions:
The followings are types of sexism:
1) threats; 2) derogation; 3) animosity; 4) prejudiced discussion. The types can be further divided into 11 categories: Threats of harm; Incitement and encouragement of harm; Descriptive attacks; Aggressive and emotive attacks; Dehumanising attacks and overt sexual objectification; Causal use of gendered slurs, profanities and insults; Immutable gender differences and gender stereotypes; Backhanded gendered compliments; Condescending explanations or unwelcome advice; Supporting mistreatment of individual women; Supporting systemic discrimination against women as a group.
Conversation:
"""{conversation}"""

Table 4: An example of the classification prompt when the toxic type is sexism.

Default sexist content explanation prompt

Based on the analysis of a classifier, the conversation below contains sexist messages. The details are as follows:
{labels}
Based on the prediction, could you explain why? Is there any other sexist content in the conversation?
Conversation:
"""{conversation}"""

Table 6: An example of the explanation prompt when the toxic type is violence against women and girls.

A.3 Example Conversation

Figure 4 and Figure 5 show examples related to violence against women and girls. Such cases have become the main focuses of the Forensic Capability Network.

A.4 *SafeSpeech* Interface Screenshots

In this section, we provide a detailed overview of the main modules of *SafeSpeech* through interface screenshots. These images illustrate the platform's key functionalities.

---

[9]EXIST official website
[10]EXIST task 1 English leaderboard

Default persona generation prompt
Your will be given a concatenated summary of all the chat messages that [*speaker*] participated in. Your task is to read the given summary and provide numeric Big Five Personality prediction scores for [*speaker*]. The Big Five personality traits are a widely recognized model for describing human personality. The traits are: Openness to Experience: Reflects the degree of intellectual curiosity, creativity, and a preference for novelty and variety. Conscientiousness: Indicates a tendency for self-discipline, carefulness, and a goal-oriented behaviour. Extraversion: Involves energy, positive emotions, and a tendency to seek stimulation and the company of others. Agreeableness: Represents a tendency to be compassionate and cooperative towards others. Neuroticism: Indicates a tendency to experience unpleasant emotions easily, such as anxiety, depression, and emotional instability. Give integer prediction scores for each trait in Big-Five Persona for [*speaker*], based on the chat message summary that will be given. Generate your response following these steps: 1. Read the summary thoroughly, especially pay attention to those content related to toxic behaviour. 2. Generate list of INTEGERS (from 1 to 10) as Big Five Personality prediction scores, each associated with a trait in order, separated by commas, wrapping with square brackets. 3. Double check to ensure that only 5 INTEGER NUMBERS are provided at the BEGINNING of your response. 4. For each trait, provide a more detailed explanation for the score predicted. 5. Generate 1-2 sentences to give an overall persona analysis for [*speaker*]. Chat Messages Summary: [*summary*] Generate the respond in the following format: "[score 1, score 2, score 3, score 4, score 5] **Openness to Experience**: \<explanation\> **Conscientiousness**: \<explanation\> **Extraversion**: \<explanation\> **Agreeableness**: \<explanation\> **Neuroticism**: \<explanation\> **Overall Persona Analysis**: \<analysis\>"

Table 5: An example of the persona generation prompt for big five personality trait generation.
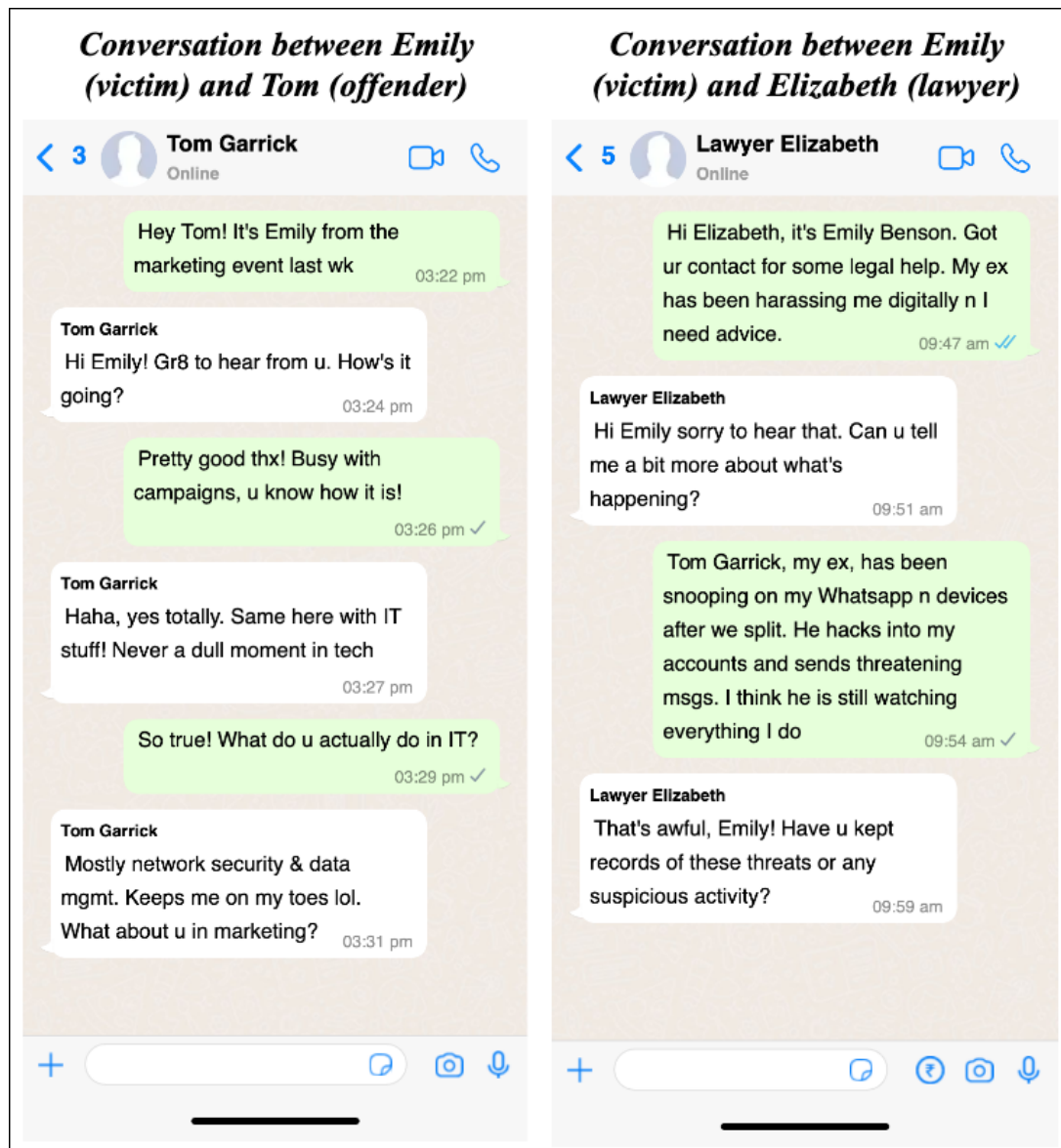
Figure 4: An example conversation about harassment.
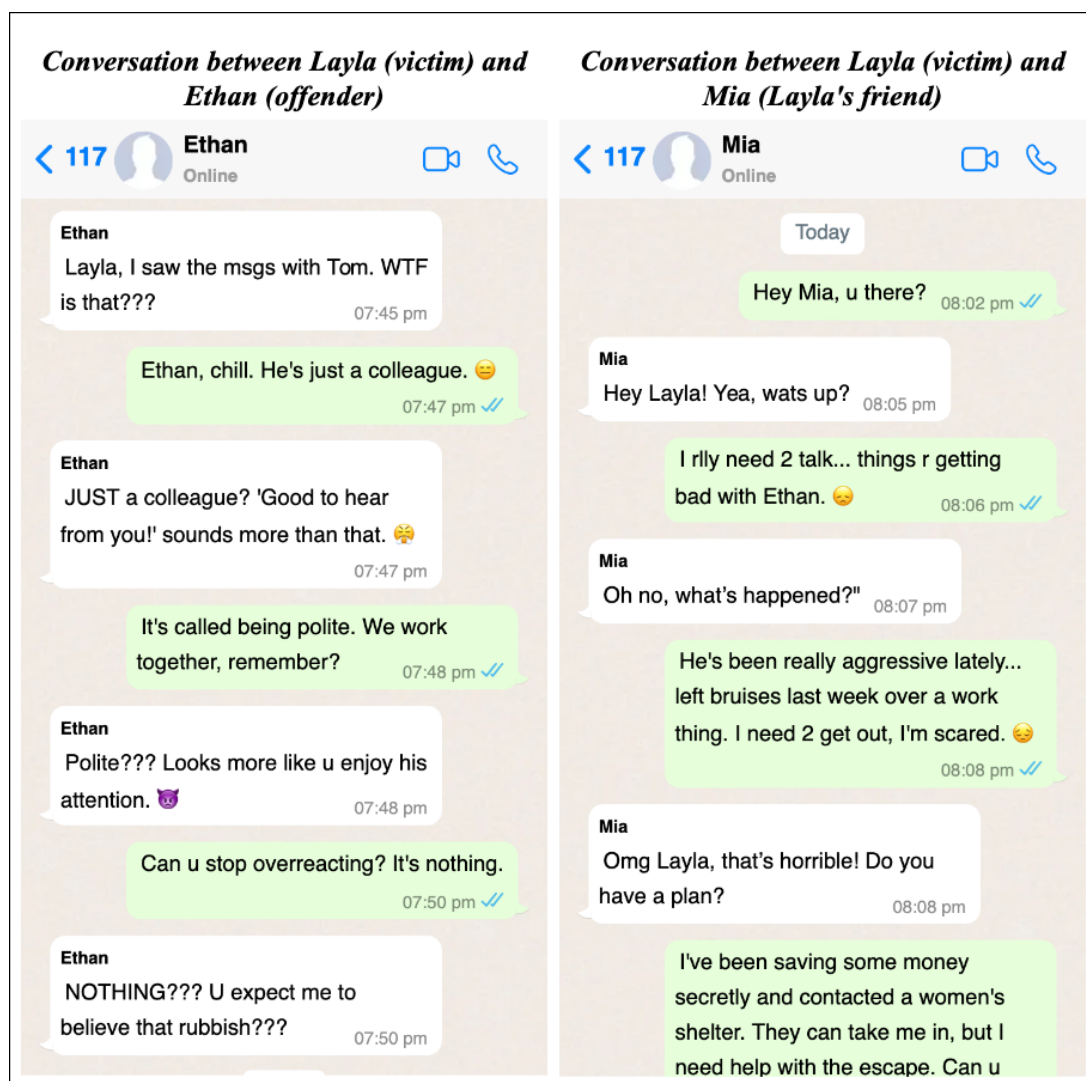
Figure 5: An example conversation involving controlling and coercive behaviours.

Figure 6: **A screenshot of the data manager page.** Users can upload their own data and preview the benchmark datasets on the platform. They can use the *delete* button to delete their own files.

Figure 7: **A screenshot of the classification page.** Users can select the classification models and change the hyperparameters. After completing the setup, they can hit the *start* button to start the analysis.

Figure 8: **A screenshot of the classification page after finishing the analysis.** The users can preview and download the output file. Our system will show a classification report if the selected file contains gold labels.
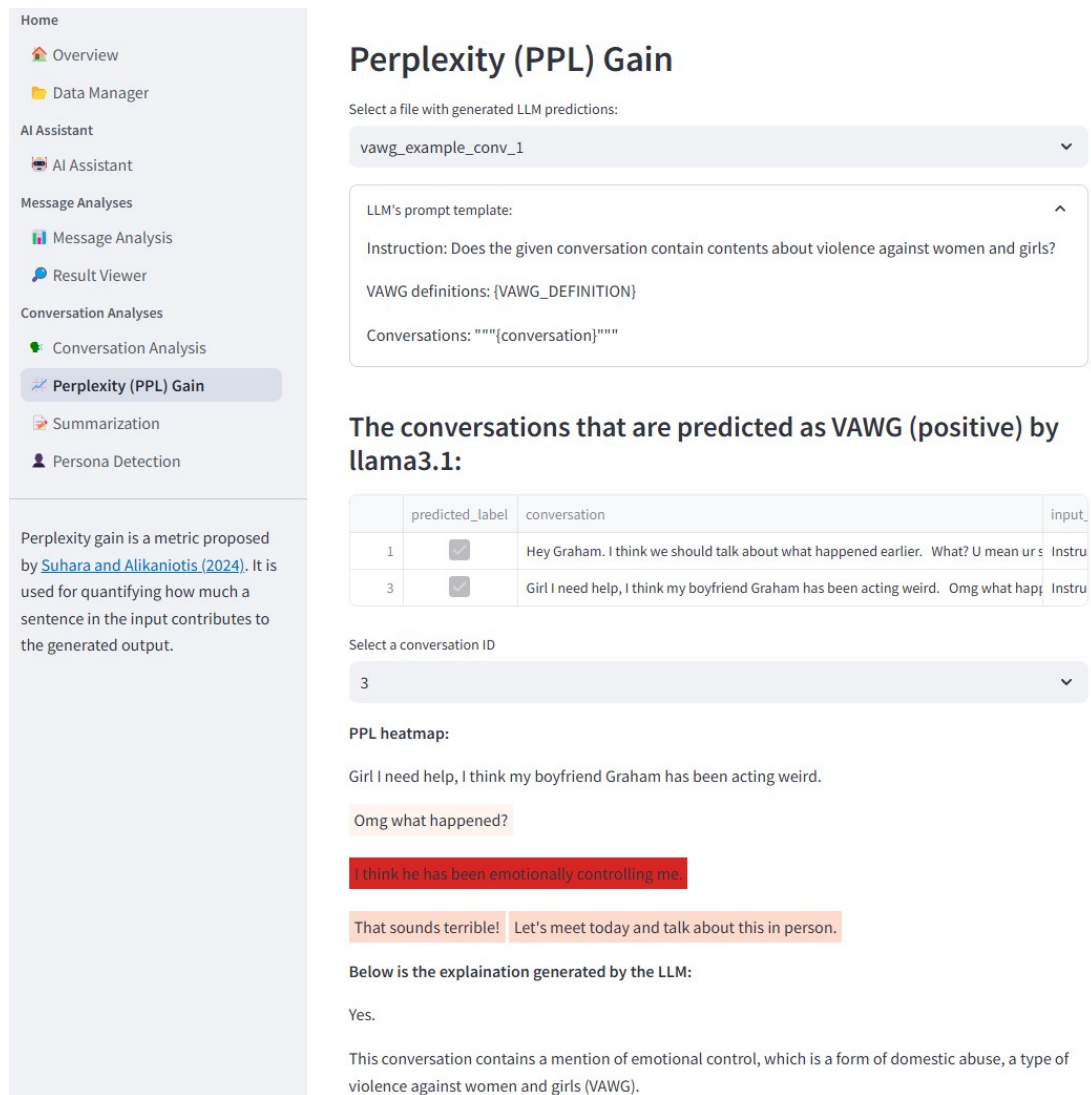
Figure 9: **A screenshot of the perplexity gain analysis page.** After choosing a conversation from the selected file, our system generates and then displays a heatmap of perplexity gains on the messages in that conversation.

Figure 10: **A screenshot of the result viewer page.** We provide a reader on our platform to let users view the result files without downloading them. Users can choose the data they would like to view, such as the 2nd most probable labels. We also integrate a conversation interface on this page which can provide explanations when the user indicate an index in the file.
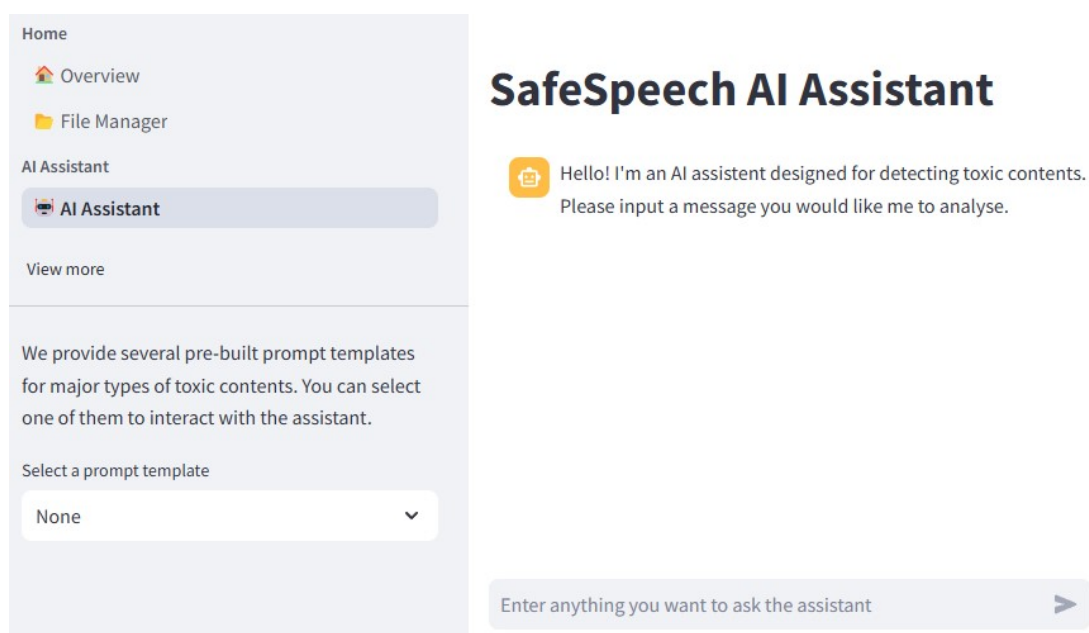
Figure 11: **A screenshot of the AI assistant page.** We provide a conversation interface, which is aided with carefully-designed prompts and classification labels, for interactive detection of toxic content. Users can also input custom prompts for detection. They can also ask follow-up questions.

Figure 12: **A screenshot of the Summarization page.** The interface displays conversation summaries segmented into logical chunks based on topics. Users can view concise, toxic-aware summaries that highlight relevant toxic or abusive content. The progress bar ensures users can track the status of the summarization process.
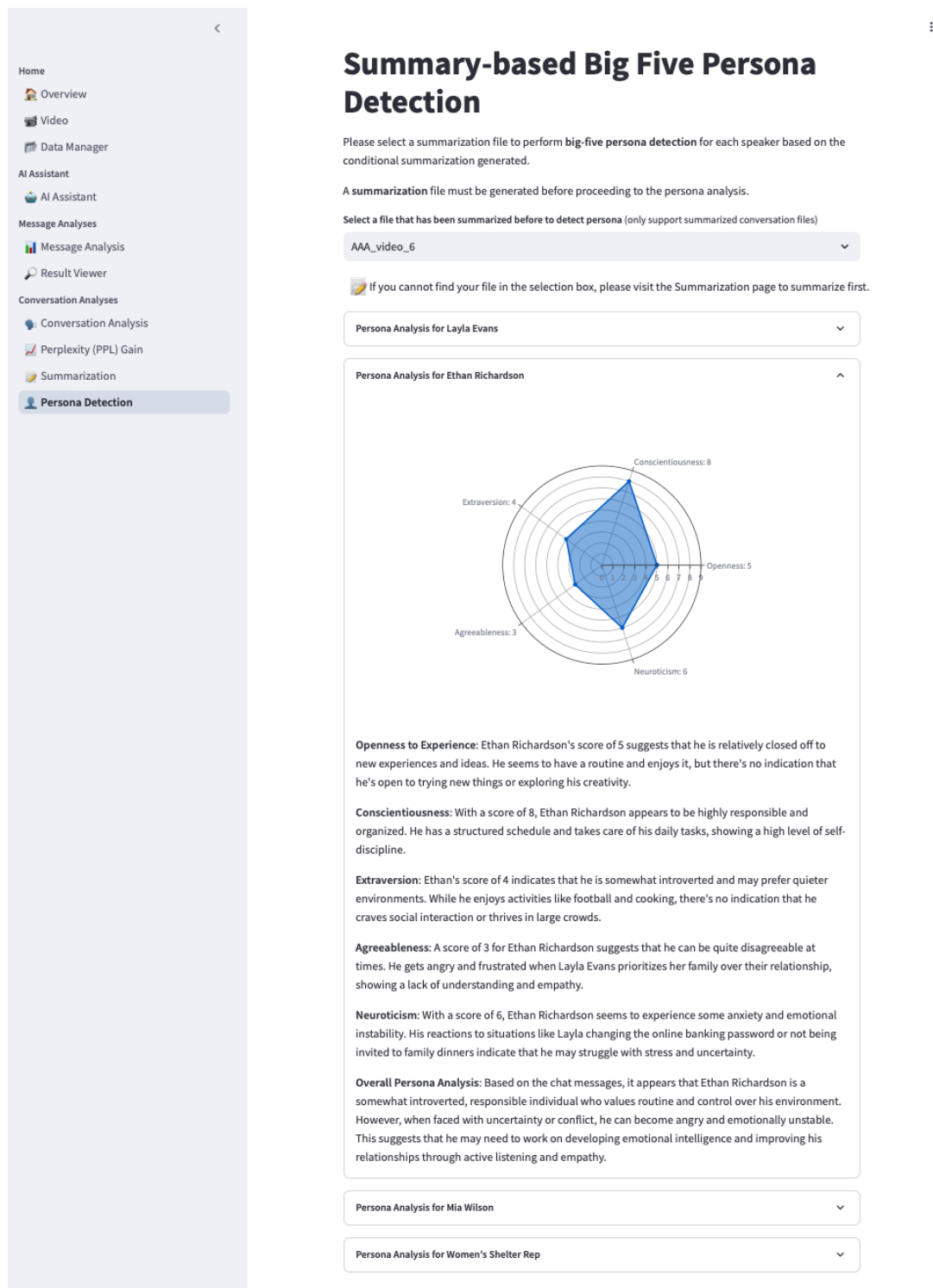
Figure 13: **A screenshot of the Persona Detection page.** This interface provides a visual representation of persona analysis results. Users can explore personality trait scores using an interactive radar plot and read detailed explanations for each trait. The clear layout helps users quickly interpret the predicted persona of conversation participants.