Personalised Abusive Language Detection Using LLMs and Retrieval-Augmented Generation

Tsungcheng Yao, Ernest Foo**, Sebastian Binnewies***
**School of ICT, Griffith University Nathan Campus, Brisbane, QLD, AU
*School of ICT, Griffith University Gold Coast Campus, Gold Coast, QLD, AU
tsungcheng.yao@griffithuni.edu.au

Abstract

Large language models (LLMs) can be useful tools for detecting abusive language on social media. However, LLMs are not always effective as they can overlook the diversity among individuals, which can lead to severe consequences. This paper proposes a novel solution that incorporates psychological knowledge into an out-of-the-box LLM using the retrieval augmented generation (RAG) method. Two rule sets were extracted and transferred to the LLM via query prompts. Experiment results showed that our solution improves LLM's performance in generating personalised detection by 1.5% to 4.4% weighted F1 score points.

1 Introduction

Abusive language detection systems play a significant role in addressing cyberbullying. Most detection systems function by identifying patterns of abusive messages, such as combinations of letters, keywords, or phrases (Jahan and Oussalah, 2023; Chhabra and Vishwakarma, 2023; Festus Ayetiran and Özgöbek, 2024). In addition, studies have proven that determining abusive language can be greatly impacted by individuals' subjectivity, including attitude, belief and experience (Sap et al., 2022; P.Y.K.L et al., 2024; Wan et al., 2023; Larimore et al., 2021). To make more personalised detection, prior studies have integrated various attributes into the systems, such as Balakrishnan et al. (2020) enhanced detection systems by introducing psychological attributes - Big Five and Dark Triad measurement scales. Kocoń et al. (2021) incorporated user demographic features into their detection systems to make adjusted predictions based on personal profiles.

Leveraging vast training data, LLMs are useful tools for abusive language detection; however, some studies have demonstrated that LLMs are not always effective in detecting such language (Kolla et al., 2024; Kruschwitz and Schmidhuber, 2024). In addition, when dealing with diversity between individuals, Park et al. (2024) found that LLMs can generate near-zero response variation in certain conditions. Overlooking individuals' diversity in abusive detection on LLMs can lead to severe consequences (Cheng et al., 2023; Gallegos et al., 2024). As a result, a novel solution is required to enhance LLMs in generating personalised abusive language detection.

This paper proposes and evaluates a novel solution incorporating psychological knowledge into an LLM (GPT-3.5 Turbo) through the RAG method, initially introduced by Lewis et al. (2020) and later extended for various applications (Fan et al., 2024). Two sets of rules were extracted from a dataset that incorporated psychological features, using association rule mining and a decision tree classifier. Then, these rule sets were provided as extra knowledge to enhance an out-of-the-box LLM's ability to generate personalised detection through the RAG approach. Our experimental results suggest that (i) our solution improves performance and (ii) it is reasonably robust with contradictory inputs. Lastly, the complete code, rules, and data are available on our repository page (here).

2 Method

An experimental approach is adopted to evaluate the effectiveness of the proposed solution (see Figure 1). Two groups, experimental and control, are created. The prompts for the experimental group are enhanced with rules derived from a dataset that includes psychological features, while the prompts for the control group are not enhanced. If the experimental group outperforms the control group in generating personalised detection, we may suggest that the proposed solution is effective and vice versa. The following sections will elaborate on the details of the experiment setup.

Notably, our experiments use a simulator as a preliminary study to assess the feasibility of the

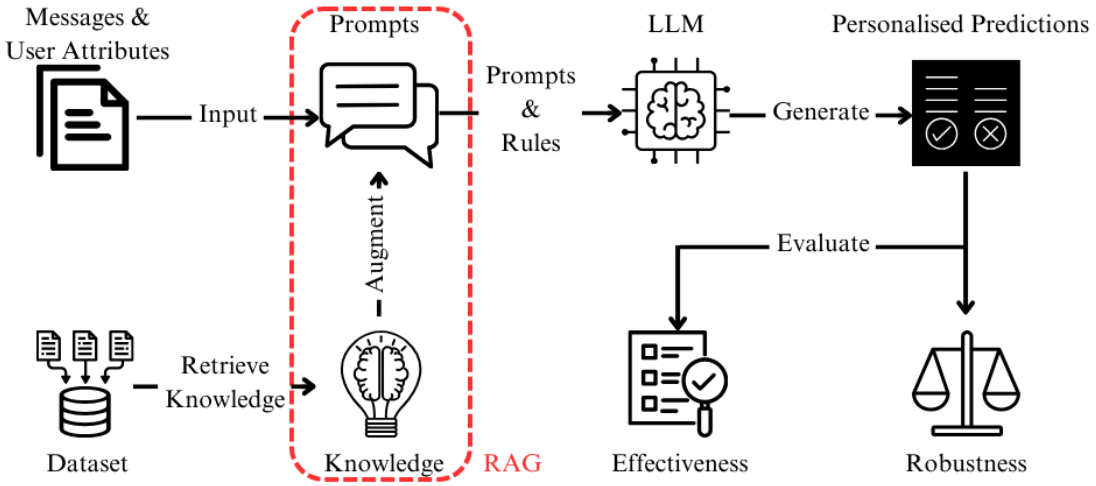

Figure 1: LLM and RAG Personalisation Framework

proposed solution. The input is replaced with test data, and the RAG component is simplified by embedding selected rules into prompts.

2.1 Data

ALDIPF: An Abusive Language Dataset that Includes Psychological Features was used (Yao et al., 2024). ALDIPF denotes 505 users' personality traits and their emotional reactions towards a series of messages with three features: personality traits (user attributes), emotional reactions (class labels), and messages. These user attributes were measured by the Five-Point Shortened General Attitude and Belief Scale, used in clinical settings (Turner et al., 2018). 65.6% of the messages belong to the Neutral class, and the rest are in the Harmful class.

This dataset was created based on the two understandings from psychological studies. First, people's emotional reactions towards messages are co-created by the messages and user attributes. Second, people with similar attributes can share a similar tendency to interpret messages (DiGiuseppe et al., 2013; Ciarrochi and Bailey, 2009). Thus, a single message can be associated with two different class labels, distinguishable only by the users' attributes.

Furthermore, recognising the limitations of LLMs in processing numeric data, user attributes were filtered and transferred into textual tokens. Three attributes, namely Rationality, Irrationality and Self-Down, were selected due to their significance in previous psychological studies (DiGiuseppe et al., 2018; David et al., 2019). After that, the original attributes were transferred from numeric values into 8 buckets (Appendix A) according to the mean and standard deviations (Owings et al., 2013). A data examples are shown in Table 1.

Message	User Attributes	Class
You're SO SMART	Low Ra High SD High Ir	1
You're SO SMART	High Ra Low SD Low Ir	0

Table 1: The same message can be associated with two class labels. Note 1: Class 0 is Neutral, and Class 1 is Harmful. Note 2: Ra indicates Rationality. Ir is Irrationality. SD refers to Self-Down.

2.2 Knowledge Extraction

To extract knowledge from ALDIPF, association rule mining and decision tree approaches were adopted. These processes can establish a correlation between certain user attributes and class labels. In this paper, we are particularly interested in indicators and rules for personalised abusive language detection.

2.2.1 Association Rule Mining Approach

Association rule mining can discover items' co-occurrence probability by identifying frequently occurring item sets and generating rules among them. This approach has been proven effective in extracting rules in various study settings (Diaz-Garcia et al., 2023; Shu and Ye, 2023).

Nine apriori algorithm-generated rules were selected after evaluating their support, confidence and lift. Five were associated with the Harmful class

Item Set	Conseq	Conf	Lift
Low Ra Ex High SD	Harmful	0.78	2.31
Very High Ir Low Ra	Harmful	0.72	2.13
Very High Ir Low Ra Ex High SD	Harmful	0.78	2.31
Low Ra, Ex High SD	Harmful	0.78	19
Very High Ir Very High SD Low Ra	Harmful	0.66	1.95

Table 2: Rules based on association rule mining approach

(see Table 2), and the rest belonged to the Neutral class.

2.2.2 Decision Tree Approach

Decision trees are widely used for knowledge representation due to their high interpretability (Shu and Ye, 2023; Sarailidis et al., 2023). This approach creates a tree-like model of decisions by splitting data into subsets and repetitive recursion.

The decision tree classifier algorithm generated rules, thirteen of which were selected after evaluating class distribution and probabilities. Four were associated with the Harmful class (see Table 3), and the rest belonged to the Neutral class. A complete decision tree is available on our GitHub page.

2.3 LLM and Prompt Optimization

The experiments were conducted on GPT-3.5 Turbo via the OpenAI API with a temperature setting of 0. Our prompts followed the framework proposed by Eager and Brunton (2023). Additionally, we used the attributed prompt technique to co-create prompts with the LLM, improving performance, particularly in the transformation of rules (Yu et al., 2023). The complete prompt is provided in Appendix B, and the design of the prompts is as follows:

- Role of the LLM: It was assigned as a classifying bot that helps users classify messages.
- The task with details: The LLM was asked to analyze the following messages and determine whether it was abusive. A message is considered abusive if it might make a reader uncomfortable after reading it. In other words, we focus on not only explicit but also implicit abusive messages.
- Additional Knowledge: The LLM was informed that the decision could be made based on the textual message and user attributes. Then, the rules discussed in Section 2.2 were presented. In our experiment, selected rules were embedded into the prompts.
- Input: Consists of (i) a message and (ii) user attributes.
- Output: A score from 0 to 1, where 0 means absolutely not harmful, and 1 means definitely harmful.

Condition	Class	Prob
In: Below Avg. Ir Ex High SD Not in: Higher Ra	Harmful	0.561
In: Very High SD Below Avg. Ra	Harmful	0.798
In: Very High Ra Not in: Below Avg. SD Very High Ir	Harmful	0.583
In: Low SD Low Ir Not in: High Ra	Harmful	0.861

Table 3: Rules based on Decision Tree. Note: these rules have been simplified for better readability.

Three prompt architectures were created by modifying the *Additional Knowledge* layer. DT_M denotes the architectures enhanced by the decision tree rules. ARM_M is enriched by association rule mining. Lastly, there is no augment for N_M, and the user attributes were removed from the *Input* layer.

3 Experiment and Result

Two experiments were conducted to evaluate the effectiveness and robustness of the proposed solu-

Archt.	ACC	Weighted F1	TP Rate
N_M	0.600	0.556	0.192
DT_M	**0.630**	**0.600**	**0.258**
ARM_M	0.608	0.571	0.225

Table 4: Result for Experiment 1. Note: TP rate is defined as Number of True Positives / Number of Positive Samples

tion.

3.1 Experiment 1: Effectiveness

This experiment evaluates whether the proposed approach can enhance the LLM's ability to generate personalised detection.

Implementation: 500 data points containing messages and user attributes were randomly selected from ALDIPF, and 36% were Harmful class. Then, these selected data points were conveyed into the *Input* layer of the prompt framework. Importantly, messages and user attributes were passed to DT_M and ARM_M, while only messages were passed to N_M. After that, the LLM's responses were cleaned and rounded to 0 or 1. Lastly, all responses were evaluated against the ordinary class labels.

Result: The experimental group consistently outperformed the control group at every metric (see Table 4). Importantly, the experimental group yielded higher true positive rates (TP rate), which implies that the experimental group can identify more abusive messages than its counterpart.

3.2 Experiment 2: Robustness

Prior studies suggested that individual subjectivity should be less influential in determining abusive messages when the messages usually have only one clear meaning (Sandri et al., 2023; Plank, 2022). Therefore, this experiment assesses whether the proposed solution can handle contradictory inputs, such as extremely positive or negative messages paired with attributes that strongly contrast the messages.

Implementation: 100 joyful messages were randomly selected from the HappyDB (Asai et al., 2018), a corpus of 100,000 happy moments. Then, these messages were joined with attributes strongly related to the Harmful class (Appendix C).

Regarding profane messages, ChatGPT created 100 samples containing at least one swear word. Then, these samples were joined with attributes strongly related to the Neutral class (Appendix C).

Type	N_M	DT_M	ARM_M
Joyful	[100,0]	[98,2]	[97,3]
Profane	[0,100]	[1,99]	[**9,91**]

Table 5: Result for Experiment 2. Note: [Neutral class, Harmful class]

Similar to experiment one, both messages and attributes were passed to DT_M and ARM_M, while only messages were passed to N_M. As a result, N_M was not impacted by manipulated attributes. In this instance, N_M serves as the baseline to evaluate the extent to which the proposed solution would be affected by contradictory inputs.

Result: For N_M, both joyful and profanity messages were accurately classified according to their nature. Nevertheless, the experimental group encountered different levels of disturbance (see Table 5). In particular, the predictions in Profanity were flipped by 9% in ARM_M.

4 Discussion

Comparison of Rules. Although the knowledge extraction approaches differ, the two rule sets still share similarities. Extremely High Self-Down is always associated with the Harmful class. In addition, Low and Lower Rationality are generally linked with the Harmful class. Nevertheless, the correlation between the Irrationality and Harmful class is unclear due to the contradictory implications of the two rules. Specifically, association rule mining indicates a positive correlation between Irrationality and the Harmful class, whereas decision tree analysis shows the opposite.

Effectiveness of Solution. The experimental group showed consistent improvement on every metric. Importantly, part of the improvement stems from identifying more abusive language (higher TP rate). As a result, the experimental group can provide more benefits for users, as identifying abusive language is the primary advantage users gain from detection systems (Hardt et al., 2016).

Robustness of Solution. Despite extremely joyful or profane messages, the results of Experiment 2 did not align with the expectation that individual subjectivity would be less influential when messages have a clear meaning.

Regarding the joyful message, our solution is robust to noise, effectively addressing contradictory inputs. However, more research is needed on profane messages. When explicit swear words were

presented, DT_M predictions were flipped by only 1%, while ARM_M predictions were flipped by 9%.

5 Conclusion and Future Work

This paper evaluates a novel solution that enhances LLMs for personalised abusive language detection by retrieving and incorporating psychological knowledge into an out-of-the-box LLM. Experiment results showed that our solution outperformed its counterpart and withstood noise reasonably well. For better re-productivity, the complete code, rules, and data are available on our repository page (here).

In our future work, we aim to address two key limitations. First, more detailed decision rules will be developed to cover a broader range of user attributes, making the retrieval component essential and requiring additional evaluation of its effectiveness. Second, further research should explore the compound effects between messages and user attributes. While this paper focuses on testing the robustness of the proposed solution in highly contradictory scenarios, the major challenge of robustness, in our view, lies in handling neutral messages across diverse individual attributes.

References

Akari Asai, Sara Evensen, Behzad Golshan, Alon Halevy, Vivian Li, Andrei Lopatenko, Daniela Stepanov, Yoshihiko Suhara, Wang-Chiew Tan, and Yinzhan Xu. 2018. HappyDB: A corpus of 100,000 crowdsourced happy moments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vimala Balakrishnan, Shahzaib Khan, and Hamid R. Arabnia. 2020. Improving cyberbullying detection using twitter users' psychological features and machine learning. *Computers & Security*, 90:101710.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Joseph Ciarrochi and A. Bailey. 2009. *A CBT-practitioner's Guide to ACT: How to Bridge the Gap between Cognitive Behavioral Therapy and Acceptance and Commitment Therapy*, volume 50.

Daniel O. David, Raymond DiGiuseppe, Anca Dobrean, Costina Ruxandra Păsărelu, and Robert Balazsi. 2019. *The Measurement of Irrationality and Rationality*, pages 79–100. Springer International Publishing, Cham.

J.A. Diaz-Garcia, M.D. Ruiz, and M.J. Martin-Bautista. 2023. A survey on the use of association rules mining techniques in textual social media. *Artificial Intelligence Review*, 56:1175–1200.

Raymond DiGiuseppe, Russell Leaf, Bernard Gorman, and Mitchell W. Robin. 2018. The development of a measure of irrational/rational beliefs. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, 36(1):47–79.

Raymond A. DiGiuseppe, Kristene A. Doyle, Windy Dryden, and Wouter Backx. 2013. *A Practitioner's Guide to Rational-Emotive Behavior Therapy*. Oxford University Press.

B. Eager and R. Brunton. 2023. Prompting higher education towards ai-augmented teaching and learning practice. *Journal of University Teaching & Learning Practice*, 20(5).

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Eniafe Festus Ayetiran and Özlem Özgöbek. 2024. A review of deep learning techniques for multimodal fake news and harmful languages detection. *IEEE Access*, 12:76133–76153.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, pages 1–79.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate

speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Udo Kruschwitz and Maximilian Schmidhuber. 2024. LLM-based synthetic datasets: Applications and limitations in toxicity detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 37–51, Torino, Italia. ELRA and ICCL.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Larry R. Owings, Gregory L. Thorpe, Evan S. McMillan, Ronald D. Burrows, Scott T. Sigmon, and Dawn C. Alley. 2013. Scaling irrational beliefs in the general attitude and belief scale: An analysis using item response theory methodology. *SAGE Open*, 3(2).

P.S. Park, P. Schoenegger, and C. Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Srinivas P.Y.K.L, Amitava Das, and Viswanath Pulabaigari. 2024. Racists spreader is narcissistic; sexists is machiavellian influence of psycho-sociological facets in hate-speech diffusion prediction. *Expert Systems with Applications*, 247:123211.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Georgios Sarailidis, Thorsten Wagener, and Francesca Pianosi. 2023. Integrating scientific knowledge into machine learning using interactive decision trees. *Computers Geosciences*, 170:105248.

Xiaoling Shu and Yiwan Ye. 2023. Knowledge discovery: Methods from data mining and machine learning. *Social Science Research*, 110:102817.

Martin J. Turner, Mark S. Allen, Matthew J. Slater, Jamie B. Barker, Charlotte Woodcock, Chris G. Harwood, and Ken McFayden. 2018. The development and initial validation of the irrational performance beliefs inventory (ipbi). *European Journal of Psychological Assessment*, 34(3):174–180.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Tsungcheng Yao, Sebastian Binnewies, Ernest Foo, and Masoumeh Alavi. 2024. See the words through my eyes: The role of personality traits in abusive language detection. *SSRN Electronic Journal*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Buckets

The numeric psychological features were transferred to 8 buckets as shown in Table 6.

Buckets	Description
Extremely High	Any $> (\overline{X} + 3\sigma_X)$
Very High	$(\overline{X} + 2\sigma_X)$ to $(\overline{X} + 3\sigma_X)$
High	$(\overline{X} + \sigma_X)$ to $(\overline{X} + 2\sigma_X)$
Above Average	$\overline{X}$ to $(\overline{X} + \sigma_X)$
Below Average	$\overline{X}$ to $(\overline{X} - \sigma_X)$
Low	$(\overline{X} - \sigma_X)$ to $(\overline{X} - 2\sigma_X)$
Very Low	$(\overline{X} - 2\sigma_X)$ to $(\overline{X} - 3\sigma_X)$
Extremely Low	Any $< (\overline{X} - 3\sigma_X)$

Table 6: Buckets and Ranges of Values

B Prompt Details

System Role: You are a classifying bot that helps users classify sentences. Please analyze the following sentence and determine if it is abusive or not. A sentence is considered abusive if a reader may feel uncomfortable after reading it. Provide a score from 0 to 1, where 0 means absolutely not harmful, and 1 means definitely harmful. Lastly, just return the score and do not provide any explanation. Additional Knowledge Base:
Here are the association rules mined from user attributes and their impact on identifying harmful messages:

Rule 1:
Antecedents: Low Rationality, Ex High SD
Consequents: discomfort
Support: 0.0014
Confidence: 0.78
Lift: 2.31
Explanation: When users exhibit low rationality and extremely high standard deviation in identity, there is a strong association with message content leading to discomfort (discomfort), with a lift of 2.31.

Rule 2:
Antecedents: Very High Irrationality, Low Rationality
Consequents: discomfort
Support: 0.0025
Confidence: 0.72
Lift: 2.13
Explanation: The combination of very high irrationality and low rationality significantly correlates with messages causing discomfort (discomfort), with a lift of 2.13.

Rule 3:
...
...
note: complete code and rules are available on our repository page (here).

User Role: Sentence: this is a sample message. Attributes: this is a sample attribute.

C User Attributes for Experiment 2

For ARM_M, joyful messages were attached to Low Rationality and Extremely High Self-Down, strongly associated with Harmful class, while profane messages were linked to Low Self-Down, strongly associated with Neutral class.

For DT_M, joyful messages were attached to Very High Self-Down and Low Rationality, strongly associated with Harmful class, while profane messages were linked to Below Average Self-Down and High Rationality, strongly associated with Neutral class.