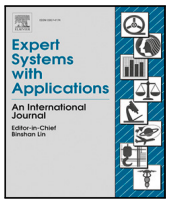
See the words through my eyes: The role of personal traits in abusive language detection

Expert Systems With Applications 276 (2025) 127188

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



See the words through my eyes: The role of personal traits in abusive language detection

Tsungcheng Yao [a,*], Sebastian Binnewies [b], Ernest Foo [a], Masoumeh Alavi [c]

[a] School of Information and Communication Technology, Griffith University Nathan Campus, 170 Kessels Rd Nathan, Brisbane, 4109 Qld, Australia
[b] School of Information and Communication Technology, Griffith University Gold Coast Campus, Parklands Dr Southport, Gold Coast, 4222 Qld, Australia
[c] School of Applied Psychology, Griffith University Mount Gravatt Campus, 176 Messines Ridge Rd Mount Gravatt, Brisbane, 4122 Qld, Australia

ARTICLE INFO

Keywords:
Cyberbullying
Ellis' ABC model
Psychological features
Dataset preparation framework
Abusive language dataset
Personalised abusive language detection

ABSTRACT

Abusive language detection systems play a significant role in addressing cyberbullying. However, conventional detection systems primarily focus on the textual pattern of messages for their prediction, overlooking the reality that the same message can usually provoke different consequences for different users. In other words, personalised predictions are essential but currently absent. To address this limitation, this paper considers the rationality of introducing a set of psychological features to personalise abusive language detection tasks. The foundation of our work lies in the Antecedents, Behaviours, Consequences model (ABC model), asserting that an individual's response to a trigger message is impacted not only by the trigger itself but also by their personal beliefs and attitudes. We develop a novel data preparation framework and construct a new abusive language dataset, incorporating psychological features from 505 online users. Logistic regression analysis illustrates that *Irrationality* and *Self-down* features positively correlate with the abusive class, while *Rationality* features exhibit a negative correlation. These results are supported by established psychological findings. Furthermore, a preliminary evaluation showed that our proposed psychological features improve a CNN-based detection system's Macro and Weighted F1 scores by 4%–5% points when making personalised predictions. These results collectively make for an empirical case that underscores the importance of considering users' psychological features in abusive language detection. Crucially, these findings also pave the way for developing personalised prediction systems.

1. Introduction

Cyberbullying is a complex issue on social media that urgently needs to be solved. Depending on the definition, the prevalence of cyberbullying rates ranges from 10% to 40% among both primary and high school students (Whittaker & Kowalski, 2015). Research has also found a correlation between cyberbullying victimisation and suicidal thoughts (Dorol-Beauroy-Eustache & Mishara, 2021; Extremera, Quintana-Orts, Mérida-López, & Rey, 2018).

To address the cyberbullying issue, the NLP community has been implementing abusive language detection systems (detection systems) to identify abusive language patterns, such as combinations of letters, keywords, or phrases. These systems have achieved reasonably positive results in identifying these textual patterns (Jahan & Oussalah, 2023; Madukwe, Gao, & Xue, 2020; Pamungkas, Basile, & Patti, 2023; Vidgen & Derczynski, 2020).

However, the majority of NLP solutions have overlooked a fundamental aspect that the same triggers, such as a phrase or tweet, do not always yield the same consequences (Davani, Atari, Kennedy, & Dehghani, 2023; Ding, You, Machulla, Jacobs, Sen et al., 2022; Ejaz, Razi, & Choudhury, 2024; Fortuna, Dominguez, Wanner, & Talat, 2022; Romberg, 2022; Sachdeva, Barreto, Vacano, & Kennedy, 2022; Sandri, Leonardelli, Tonelli, & Jezek, 2023). Empirical findings from psychology studies have proven that individuals can hold varying perceptions or feelings in response to the same triggers due to their distinct beliefs and attitudes (psychological features) (Ellis, 1991; Ellis & MacLaren, 1998; Goldfried & Sobocinski, 1975; Monks & Smith, 2006; Otte, 2011). For instance, "you are so smart!!" could be intended as a compliment; however, someone with a sensitive personality can also perceive it as sarcasm. Therefore, the current NLP solutions focusing solely on textual patterns must be revised, as they cannot handle this complexity among individuals.

* Corresponding author.
E-mail addresses: tsungcheng.yao@griffithuni.edu.au (T. Yao), s.binnewies@griffith.edu.au (S. Binnewies), e.foo@griffith.edu.au (E. Foo), m.alavi@griffith.edu.au (M. Alavi).

https://doi.org/10.1016/j.eswa.2025.127188
Received 18 June 2024; Received in revised form 27 February 2025; Accepted 5 March 2025
Available online 15 March 2025

To enhance the capacity of abusive language detection systems in handling individual complexity, we have incorporated insights from psychological theories. This incorporation involves considering the feasibility of applying a psychological framework and proposing psychology-based features designed explicitly for abusive language detection tasks. To be more specific, this research is based on the Antecedents, Behaviour, Consequences Model (ABC model) proposed by Ellis (Ellis, 1991; Ellis & MacLaren, 1998). This model posits that an individual's response to a trigger is impacted not only by the trigger itself but also by their attitudes and beliefs, leading to different emotional reactions. In the context of abusive language detection, such triggers are represented by messages, and responses refer to the class label of the messages. Therefore, as per the ABC model, people can react differently to the same messages, as their responses are shaped by their unique interpretations and personal traits. Recognising this connection between responses and personal traits, we suggest that personalised detection systems can be realised by incorporating psychology-based features, representing one of the approaches to assessing one's personal traits.

The contributions of the paper are as follows:

- This research created a new data preparation framework based on psychological, management, and NLP studies. Furthermore, with this framework, we built and shared a new abusive language dataset that contains psychological features from 505 online users (Sections 5 & 6). The new dataset is available on our GitHub page https://github.com/tsungcheng-yao-griffith/ALDIPF Link.
- Our statistical analysis validated the feasibility of psychological features in the context of abusive language detection. Specifically, we examined the correlation between psychological features and the *Harmful* class label. Utilising logistic regression analysis, we observed that *Irrationality* and *Self-down* scores positively correlate with the *Harmful* class, while *Rationality* score exhibited a negative correlation. These findings are not only supported by established psychological assumptions (Section 7.1.2) but also pave the way for developing personalised prediction systems.
- We are among the early adopters in integrating structured psychological frameworks and theories into personalised abusive language detection, showcasing a remarkable 4%–5% points enhancement in both Macro and Weighted F1 scores on a CNN model (Section 7.2).
- Our statistical analysis noted that a two-layer self-selection bias could lead to online users having significantly higher *Irrationality* scores than ordinary users. As a corollary of the second point, these elevated scores have the potential to disproportionate abusive class labels in a crowdsourcing-based dataset (Section 7.1).

This paper is constructed as follows: In Section 2, we elaborate on the research objectives that consider the rationale of incorporating psychological features to personalise detection systems. Following this, Section 3 discusses the underlying psychological theory (ABC model) and methodology employed. Subsequently, in Section 4, leveraging the insights from the ABC model, we propose two significant changes (See Table 2) that must be made. To evaluate our proposal, we built a new abusive language dataset based on a novel data preparation framework that accounts for psychological features, detailed in Sections 5 and 6. Utilising this new dataset, two studies are initiated in Section 7: Study One concentrates on the effectiveness of psychological theories in the context of abusive language detection studies, while Study Two examines the impact of incorporating psychological features on detection systems. Section 8 presents related work and comparative studies. Finally, in Section 9, based on the positive experiment results, we suggest several practical implications for future studies, including building personalised detection systems and considering the fairness of detection systems from psychological lenses.

2. Research objectives

Considering the potential contribution of psychological studies in personalising abusive language detection tasks, we aim to achieve three objectives in this research:

1. **Develop a dataset preparation framework supported by both NLP and psychological studies.** To the best of our knowledge, there is no existing dataset that contains both NLP and users' complete psychological features available for this study. Therefore, we aim to build the first abusive language dataset that includes and considers psychological factors.
2. **Assess the feasibility of psychological features in the context of abusive language detection.** Although the effectiveness of psychological theories has been widely reviewed and validated for decades, their effectiveness in the domain of abusive language detection has not yet been assessed. In particular, we concentrate on establishing the correlation between certain psychological features and the *Harmful* class label.
3. **Examine the impact of incorporating the psychological features on detection systems' performance.** The experiment method will be used to assess if detection models enhanced with psychological features outperform detection models without those features when making personalised detection. Given consistent improvement is observed across models, it can be inferred that psychological features improve the detection system and vice versa.

3. Theoretical framework and methodology

This section outlines the theoretical framework and methodology employed in the study.

3.1. Theoretical framework

This paper focuses on the ABC Model, which finds application in Rational Emotive Behaviour Therapy and Cognitive Behaviour Therapy. The ABC model, in particular, concentrates on the sequential events that lead to the manifestation of behaviours and emotions. While it is not a comprehensive therapeutic approach in itself, it serves as a vital component within broader psychological theories and therapeutic frameworks (Ciarrochi & Bailey, 2009; DiGiuseppe, Doyle, Dryden, & Backx, 2013; Dryden, 2005; Ellis, 1991).

Albert Ellis (Ellis, 1991; Ellis & MacLaren, 1998) proposed the initial ABC model in 1956, which has been validated and revised in the decades since (see Figure ??). The key principle of the ABC model is that adverse events (A) themselves do not directly lead to emotional disturbance (C). Instead, individuals have the power to choose their attitudes (B) towards these events. If they adopt rigid and extreme attitudes (B) in response to adversity (A), it can trigger emotional, behavioural, and cognitive disturbances (C). In essence, the ABC Model explains that attitudes and interpretations (B) play a significant role in shaping responses (C) to external events (A). This framework provides insight into how beliefs and reactions interact, shedding light on the processes underlying emotions and behaviours.

Beliefs and Attitudes (B) refer to positive, negative, or mixed evaluations of an object expressed at some level of intensity, and, in general, they can be broken down into three components: affective, behavioural, and cognitive attitudes (Eagly & Chaiken, 1993; Pratkanis, Breckler, & Greenwald, 1989). These components shed light on how individuals perceive things from diverse angles. Furthermore, these attitudes are closely intertwined with one's belief system, which can generally be categorised as either rational or irrational beliefs (Ellis, 1991; Ellis & MacLaren, 1998).

Table 1
Research Methods and Corresponding Sections.

Tasks	Related to	Methods	Section
Data acquisition	Research Objectives 1	Survey – Crowdsourcing via an online platform	5 & 6
Data analysis	Research Objectives 2	Statistical Analysis – ANOVA – Correlation Coefficient – Feature Importance by Random Forest – Correlation Analysis by Logistic Regression	6 & 7.1
Development	Research Objectives 3	Experiment – Build control and experimental detection systems with different language models - Compare the performances between the systems	6 & 7.2

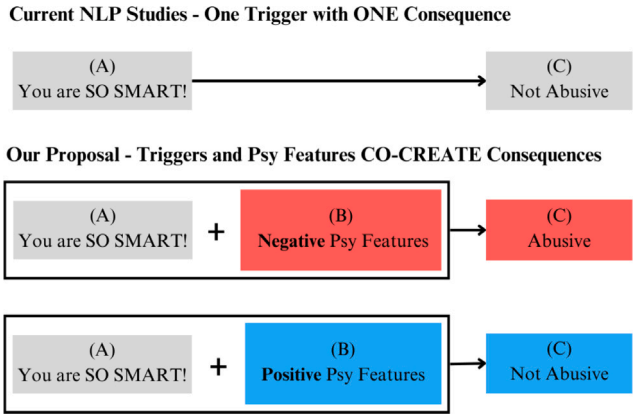


Fig. 1. Different Understandings of the Causation between Trigger and Consequences.

Table 2
The Differences between Current Work and Our Proposal. We suggest triggers and psychological features co-create the consequences. As a result, a single trigger can lead to more than one consequence.

	Conventional NLP models	Psychological studies (Our Proposal)
What factor(s) determine class labels?	The class label of the message is determined by the **message only**	The class labels are determined by **messages and users' psychological features**
How many class labels for a message?	A message can only be associated with **one class label**	As a corollary of the point; a message can be related to **more than one class label** because users have different psychological features

3.2. Methodology

To achieve the aforementioned research objectives, this research involves three major tasks - *data acquisition*, *data analysis*, and *development*, which require varying research methods (see Table 1). First, the data acquisition task was leveraged by a survey method. A new dataset was created through a crowdsourcing method that is believed to reach a wide range of users with diverse backgrounds. Second, regarding the *data analysis*, a few statistical analyses were conducted to establish the correlation between specific psychological features and abusive (positive) class labels with the data collected. We were particularly interested in whether these correlations are aligned with the assumption of prior psychological studies. Finally, an experimental approach was employed for the *development*. We constructed detection systems in pairs: experimental and control. The sole distinction between them is that the experimental system was enhanced by psychological features, while the other is not. Subsequently, if the experimental system outperforms its counterpart (control), we can infer that this improvement is attributed to the psychological features.

4. Theoretical implications and our proposal

NLP and psychology often have different understandings of the causation between the *Triggers* (Messages) and the *Consequences* (Class Labels). As shown in Fig. 1, current detection systems tend to assume that a *Trigger* (A) directly links to a *Consequence* (C); however, in light of insights from the ABC model, individuals possess diverse beliefs and attitudes (*Psychological Features*), leading them to interpret and perceive the same *Trigger* in varying *Consequences*. As a result, a single *Trigger* (A) can result in more than one *Consequence* (C) because of varying *Psychological Features* (B).

Given the insights from the psychological theory, personalised detection systems can be realised by implementing two major changes, as shown in Table 2. First, users' psychological features should be taken into account when detecting abusive language, as these can lead to varying interpretations of the same message, sometimes even in opposite ways. Second, due to these varied interpretations, a message should be associated with more than one class label, reflecting the different perceptions users can have towards the same message.

Finally, to avoid confusion, we employ the term **Psychological Features** to denote all components within the *Belief and Attitude layer (B)* collectively. Lastly, the measurement of (B) will be discussed in Section 5.3.

5. Preparation framework and data collection

To our knowledge, no existing dataset encompasses all the essential components: Triggers, Psychological Features, and Consequences necessary for evaluating our proposal. Therefore, our first objective is to construct a novel abusive language dataset that incorporates and considers psychological factors.

This section addresses the **first research objective** proposed in the Research Objectives. It outlines the overarching concept of a new dataset preparation framework aimed at acquiring all essential features of the ABC model via a crowdsourcing method. Subsequently, we discussed the details of constructing a novel abusive language dataset called ALDIPF: An Abusive Language Dataset Including Psychological Features. This new dataset was utilised for sequential evaluation.

Lastly, it is important to note that this data preparation framework can also be used for other psychological frameworks, such as decision-making theories and trait theories, by substituting the psychological tool in Step 1.

5.1. Interdisciplinary preparation framework

Fig. 2 illustrates the framework for collecting the features proposed in the ABC model. First, Step 0 evaluated the data collection methods

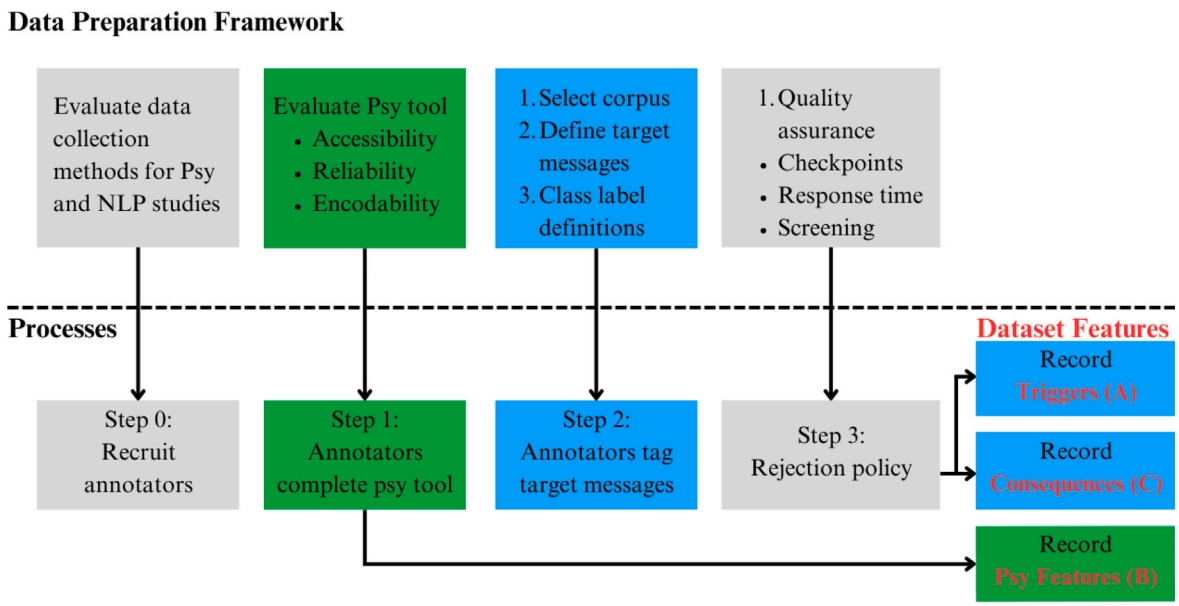


Fig. 2. The Framework and Processes for Data Preparation. By substituting the psychological tool in Step 1, this framework can be used for other psychological frameworks.

that suit psychological and NLP studies. After that, step 1 involved collecting annotators' psychological features by administering a psychological measurement scale. The results from this step contributed to the *Psychological Features* (B) of the ABC model. Subsequently, in Step 2, we selected appropriate messages from Founta (Founta et al., 2018) data. Notably, Steps 1 and 2 were presented in the same survey form during the implementation to ensure a timely correlation between the features. Step 3 adopted four indicators to evaluate the quality of work. These indicators were inspired by management and psychological studies (Aguinis, Villamor, & Ramani, 2021; Hunt & Scheetz, 2019; Kennedy, Clifford, Burleigh, Waggoner, Jewell et al., 2020) that suggested best practices for conducting research on crowdsourcing platforms.

Furthermore, having recognised the influence of psychological features (see Section 7.1), we made an effort to elaborate on the annotation processes that could potentially affect these features. The inclusion of specific implementation details, such as checkpoint questions and rejection policies, in our discussion may also be beneficial for subsequent studies.

5.2. Step 0: Recruit annotators

Crowdsourcing is used for data collection as it has been the method of choice and validated by both NLP and psychological studies. For NLP studies, crowdsourcing has been widely used in building language datasets (Jahan & Oussalah, 2023; Kocoń, Figas, Gruza, Puchalska, Kajdanowicz et al., 2021; Madukwe et al., 2020; Mishra, Yannakoudakis, & Shutova, 2020). Crowdsourcing is popular in psychological studies as well. According to Zhou and Fishbach (2016), between 20% and 45% of top psychological studies were based on an online crowdsourcing method in 2015. These psychological studies covered various topics, such as eating disorders, personal traits, and online bullying (Barlett, Kowalewski, Kramer, & Helmstetter, 2019; Burnette et al., 2022; Colman, Vineyard, & Letzring, 2018).

5.3. Step 1: Annotators complete psy tool

This step aims to collect annotators' *Psychological Features*, referring to the ABC model's psychological features (B). The General Attitude and Belief Scale (GABS) is a widely used and established measure of beliefs and attitudes (Bernard, 1998; DiGiuseppe, Leaf, Gorman, & Robin, 2018; Lindner, Kirkby, Wertheim, & Birch, 1999). However, due to its length of over 150 questions and taking around 12 min to complete, previous studies have shown that the GABS has a higher dropout rate compared to other scales (Aguinis et al., 2021; Hunt & Scheetz, 2019; Kennedy et al., 2020). Therefore, a quality alternative to GABS is required to complete the scale more efficiently.

Addressing the aforementioned concern, Step 1 employs the complete "Five-point Shortened General Attitude and Belief Scale (SGABS)" as an alternative due to its accessibility, reliability, and ease of encoding. First, regarding accessibility, the SGABS represents a condensed version of GABS comprising 26 questions, which can be completed within 4 min. This brevity can potentially enhance annotator attentiveness and reduce attrition rates. The SGABS can be accessed at here.

Secondly, SGABS has demonstrated its reliability in assessing psychological features, despite being a condensed scale. Studies have revealed significant associations between the SGABS indicators and the Irrational Belief Scale scores, corroborating its reliability as a measure of these attributes (Lindner et al., 1999; MacInnes, 2003). Furthermore, SGABS has been the basis for developing several other scales (Gavit, David, DiGiuseppe, & DelVecchio, 2011; Turner et al., 2018), adding to its credibility as a dependable tool for assessing psychological features.

Moreover, SGABS generates five clinically relevant indicators, which can assist in predicting an individual's tendencies or distinguishing specific groups of people from others. SGABS assesses seven items related to individuals, and these items are instrumental in deriving clinical-based indicators (see Table 3). These indicators have been proven to have strong correlations with other psychological measurement scales, such as the Irrational Belief Scale and Beck Depression Inventory, which are employed to evaluate negative emotions, including anger, anxiety, and depression (Lindner et al., 1999; MacInnes, 2003; Turner et al., 2018).

Furthermore, the SGABS results are encodable by machine learning models due to their quantitative measurement methods. Beliefs and Attitudes (B) can be measured in different ways; however, not all psychological measurement scales are appropriate for NLP studies as some psychological measurement scales adopt qualitative methods to measure psychological features, which could be challenging to transfer these qualitative results to NLP models. In contrast, SGABS uses a 5-point rating scale to measure and describe psychological features (see Fig. 3), which are then aggregated into seven SGABS items (see Table

Table 3
SGABS items: 1. Rationality 2. Need for Achievement 3. Other Down 4. Need for Comfort 5. Self-down 6. Need for Approval 7. Demand for Fairness. Note 1: Our dataset comprises complete SGABS results, including all the items required to generate the clinical indicators. Note 2: SGABS's statement is provided in Appendix A.

Clinical indicators	Description
Irrationality score	Sum of items 2 - 7
Rationality score	The score of item 1
Total score	Irrationality score deduct Rationality score
Downing score	The scores of items 5
The Highest Irrational Item	Select an item that has the highest average score from items 2 to 7

Example:	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
People should never break a promise	1	2	3	(4)	5

The person has shown that he/she agrees with the statement by circling number 4. If the person had strongly agreed with the statement he/she would have circled number 5.

Fig. 3. An Example of SGABS Questions.

3). Each item is presented in a numeric format that allows them to be encoded to NLP models. Therefore, SGABS is a suitable psychological measurement scale for NLP studies.

Finally, we acknowledge the value of other widely validated psychological tools, such as the Big Five model and the Myers-Briggs Type Indicator (MBTI), in understanding general life outcomes. However, these tools do not directly address the cognitive-emotional mediation processes involved in reactions to specific triggers, such as abusive language. As a result, we selected the ABC model over the Big Five and MBTI.

5.4. Step 2: Annotators tag target messages

This step collects annotators' emotional responses towards messages. In other words, this step focuses on the *Consequence* (C) and *Triggers* (A) parts of the ABC model. Step 2 is similar to conventional NLP annotation tasks, which involve collecting corpus and asking annotators to tag them according to class definitions.

5.4.1. Selection of corpus

We reuse the corpus from a current dataset for reliability since studies have proven that corpus creation correlates with a dataset's quality (Alkomah & Ma, 2022; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021; Vidgen & Derczynski, 2020; Xia, Field, & Tsvetkov, 2020). Reusing a corpus from a current dataset can be an alternative to quickly building a new dataset with a reputation, as some current datasets have been widely validated and made open for re-annotation.

To select the most suitable dataset, current datasets were evaluated from five aspects: (i) Reputation, (ii) Accessibility, (ii) Recency, (iv) Size, and (v) Genericity. Particularly, the genericity of abusive language is vital because abusive language can have different natures. For instance, sexism and racism are highly oriented and may provoke more severe consequences for particular demographic groups (Garg, Masud, Suresh, & Chakraborty, 2023; Parikh et al., 2019; Şahinüç, Yılmaz, Toraman, & Koç, 2023). As a result, we only nominate datasets focusing on generic abusive language. Sexism, racism, and other abusive language datasets are saved for future work.

After evaluation, Founta (Founta et al., 2018) abusive language dataset is selected (available at here). Since Founta published this dataset in 2018, it has been widely recognised and used in several abusive language studies (Antonakaki, Fragopoulou, & Ioannidis, 2021; Poletto, Basile, Sanguinetti et al., 2021; Vidgen & Derczynski, 2020). This dataset comprises around 100k tweets tagged either *Normal*, *Abusive*, *Hateful*, or *Spam*. All these class labels came from crowdsourcing and were built on a voting scheme. Each tweet has at least five votes, and the class label with the highest vote becomes the finalised class label.

Fig. 4. Majority's vs. Minority's Perception.

5.4.2. Selection of target messages

To further narrow the scope of the research, we focused on *Target Messages* (see Fig. 4) that can provoke moderate emotional responses from annotators. Therefore, we eliminate (i) messages that contain obvious abusive patterns and (ii) messages that are obviously not abusive. This elimination is based on the understanding that when messages are obviously abusive (or obviously not abusive), annotators tend to judge solely based on the messages rather than their *Psychological Features*. Therefore, we focused on *Target Messages* to utilise the ABC model, and *Target Messages* were defined and selected as follows.

Inspired by psychological studies, we suggest that abusive language can be categorised into four categories rather than two. According to the ABC model, the same trigger can provoke different emotional responses since individuals hold different psychological features. Therefore, the taxonomy of abusive language should be a spectrum instead of a dichotomy. In this instance, some messages are always abusive/ not abusive for almost everyone. Nevertheless, depending on context and other factors, other messages drift between abusive and not abusive. Therefore, the understanding of abusive language can be extended into four categories, as shown in Fig. 4. The minority refers to users who tend to frequently possess different opinions against ordinary users. These different opinions can stem from psychological features, culture, gender, religion or any other factors (Plank, 2022; Sandri et al., 2023; Uma, Fornaciari, Hovy, Paun, Plank et al., 2022).

We concentrate on the messages in the second quadrant for three reasons. First, the impact of *Psychological Features* may be more significant in the second quadrant. The messages from the first and the third quadrants are almost always abusive (or not abusive) for almost everyone. Therefore, we suggested these messages must be clear and sufficient for annotators to judge solely based on textual patterns rather than other factors, such as *Psychological Features* and context of the messages. As a result, the correlation between *Psychological Features* and *Consequences* could be weak in the first and the third quadrants. In contrast, the messages in the second quadrant have mixed *Consequences*, which may stem from individuals' various *Psychological Features*. Therefore, the messages from the second quadrant may provide more valuable information for studying the impact of individuals' *Psychological Features*.

Table 4
Ambiguity Levels of the *Normal* Class Label. In Founta's dataset, each tweet should be annotated by at least five annotators. However, the authors did not provide clarification regarding instances where a tweet received only one or two votes but was still tagged as the *Normal* class.

Votes received	Number of message
>=5	15 831
>=4	18 720
>=3	16 235
>=2	3009
>=1	49
Total	53 851

Additionally, the second quadrant is crucial to creating a more comprehensive countermeasure against abusive language. The current detection systems adopt voting schemes to aggregate different opinions among annotators at the cost of ignoring the voice of minorities (Romberg, 2022). In other words, even though a message could harm a small number of users, the message could still be tagged non-harmful if most people vote so. In this case, the current solutions cannot bring the benefit equally (Davani et al., 2023; Ding et al., 2022; van der Wal, Bachmann, Leidinger, van Maanen, Zuidema et al., 2024). Considering this drawback, studying the second quadrant may provide a better solution for bridging the majority and minority.

Third, the fourth quadrant was not included because of the research interest. We conjectured that the fourth quadrant is more relevant to freedom of speech as these messages can be seen as controversial. Freedom of speech is essential but has a different nature from abusive language. Considering this potential difference, the fourth quadrant was not selected for this research.

The *Target Messages* were selected for this research as follows. First, we collected messages initially tagged as *Normal* in the Founta dataset. After that, messages that did not achieve 100% inter-rater agreement, as indicated by the voting column, were selected. In essence, any message with a voting column below five was categorised as a *Target Message*. This classification was based on the guideline established by Founta et al. (2018), which specified that each tweet should be annotated by at least five annotators. According to these criteria, 38k messages have exhibited varying levels of ambiguity and were grouped into *Target Messages* for the annotation task (Table 4).

5.4.3. Class label definitions

The class labels refer to the *Consequences* (C) part of the ABC model. *Consequences* are determined by personal feelings or perceptions rather than literal definitions. In other words, *Consequences* are annotators' responses towards the *Triggers* (Messages), and annotators are allowed to have different reactions towards the same *Triggers*. Consequently, we developed a compatible class label system that (i) allows annotators to give different class labels to the same message and (ii) emphasises personal perceptions. The definitions of class labels are as follows.

- Harmful: If you experience any negative emotion or feeling, such as sadness, anger, or unease, after reading this message, this message is *Harmful*.
- Neutral: Any other messages that do not fall in the *Harmful* category

5.5. Step 3: Quality assurance and rejection policy

We adopted four indicators to evaluate the quality of Human Intelligence Tasks (HITs). These indicators were inspired by management and psychological studies (Aguinis et al., 2021; Hunt & Scheetz, 2019; Kennedy et al., 2020) that suggested best practices for conducting research on crowdsourcing platforms. With these indicators, a rejection policy was developed to assess whether HITs should be accepted or rejected.

5.5.1. Indicators of hits' quality

Regarding the first indicator, three checkpoint questions were included in the message annotation section to check whether annotators were working on a task attentively and carefully.

Checkpoint questions are artefact questions that (i) ask annotators to select a particular answer or (ii) enter specific keywords. We applied both in our HITs. In the message annotation section, two messages were replaced by two checkpoint questions that asked annotators to select a particular option. For instance, we inserted "*#NAME #NAME Im a happy MTurkers!!! I read through every line carefully and wont miss anything. Pls select harmful for this attentive check question*". In this instance, as we deliberately inserted "happy" at the beginning of the message, annotators may select the *Harmful* option only when they pay enough attention to the HITs.

Additionally, an open-ended feedback column was attached at the end of HITs. This column had two functions – (i) collecting feedback from annotators and (ii) asking annotators to enter (or not enter) messages according to a description. For instance, we asked annotators to enter any feedback other than "good" or "nice", which is common feedback from bots. With these checkpoint questions, we can evaluate whether annotators were attentive and genuine.

The second indicator is the response times. We recorded (i) when an annotator joined and completed a HIT and (ii) how much time they spent on each HIT sub-component. According to our pilot studies, most annotators spend 12–15 min completing the task, and attentive annotators should be able to complete the task within 30 min. We concluded three suspicious classes for submissions that do not fall into this reasonable response time.

- Long response time: When response time is longer than 30 min, annotators may be inattentive after accepting HITs.
- Hurried response time: When response time is between 8 – 12 min, annotators may put insufficient effort into current HITs.
- Short response time: annotators spend minimal effort on current work when response time is shorter than 8 min.

Regarding the third indicator, by default, once annotators participate in HITs, their working history is recorded and can be tracked on Amazon Mechanical Turk (MTurk). This information allows us to build a blocklist to ban inferior annotators. In addition, this information is also important when evaluating the quality of annotators' work with *Hurried response time*. For instance, some experienced annotators can complete the survey faster than others. One possible reason is that they skip description sections and complete the annotation task more effectively than others. In this case, this working history provides contextual information to evaluate the quality of annotators' work.

Finally, every HIT is manually screened to ensure that submissions are reasonable. This manual screening is based on two supplementary indicators: (i) Tagging Ratio: since all messages were initially tagged *Normal* in the Founta dataset, we expect that most messages should be tagged *Neutral* in our annotation task. Therefore, if a work has a high ratio of *Harmful*, this work may be suspicious, (ii) Response Patterns: Bots and inattentive annotators tend to complete the survey with static patterns, such as all messages being tagged *Harmful*. Human examiners can quickly identify these static patterns.

5.5.2. Rejection policy

Based on the aforementioned indicators, a rejection policy was developed to eliminate inferior submissions (see Table 5). All submissions must pass this policy to be accepted and incorporated into the final dataset. Within the policy, caution points were assigned to each inattentive or suspicious behaviour. These points were determined collaboratively by the authors to ensure reliability and diverse perspectives. The higher the points, the lower the quality of work. Submissions with three caution points (or more) were automatically disqualified unless exceptions were warranted.

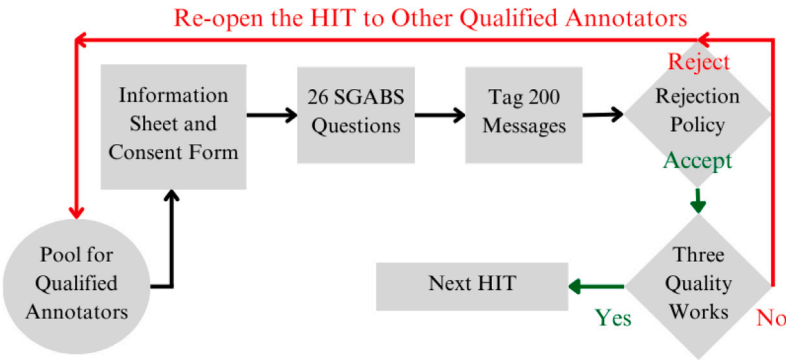


Fig. 5. HITs Workflow: A total of 576 HITs were completed by 505 unique annotators. Black arrows represent the annotation process, red arrows indicate conditions for re-opening the HIT, and green arrows signify the continuation of the process.

Table 5
Rejection Policy. Submissions that had three (or more) caution points were automatically disqualified unless exceptions were warranted.

Behaviour	Caution Point
Fail all checkpoint questions	1 point
Short response	4 points
Hurried response	2 points
Long response	2 points
Extreme tagging ratio	1 point
Suspicious patterns	4 points
Decision: If a submission has less than three caution points, then accept. Otherwise, reject. Exceptions: 1. One caution point may be deducted if an annotator can give meaningful feedback to explain their misconduct. 2. People with unique psychological features may have an extreme ratio of harmful and neutral tweets. In this case, this submission may be accepted. 3. If annotators re-do this task, a shorter response may be accepted.	

5.6. Crowdsourcing platform and task setting

The proposed data preparation framework was implemented on Amazon Mechanical Turk (MTurk). This section details the setting of the platform's configuration and task setup.

5.6.1. The crowdsourcing platform and qualifications

The crowdsourcing method was performed on Amazon Mechanical Turk (MTurk). In addition, because of the nature of the Founta dataset, annotators should be native English speakers to participate in the survey. Nevertheless, "native English speaker" was not on the MTurk qualifications list when we conducted the research. Thus, "annotators must be in the United States" is an alternative.

Furthermore, to gain more general *Psychological Features*, all qualified annotators were invited. Hiring master annotators who have a higher acceptance rate is a common strategy to enhance human intellectual tasks (HITs) quality (Aguinis et al., 2021; Kennedy et al., 2020). However, as a prior study suggested that master annotators may have specific psychological features and cannot represent the general *Psychological Features* of all annotators (Cheung, Burns, Sinclair, & Sliter, 2017), we did not adopt this strategy.

5.6.2. Human Intellectual Tasks (HITs) and task descriptions

The workflow of HITs is shown in Fig. 5, which contains three subcomponents — HITs description, psychological measurement scale, and annotating messages. We initially provided the annotators with the information sheet and consent form. Once they had read and agreed to our conditions and terms, they could proceed to the psychological measurement scale. Then, annotators were asked to tag 200 messages according to the class label definitions discussed in Section 5.4.3. Additionally, *Target Messages* were first normalised and cleaned, with usernames, names, and URLs replaced by tokens. Finally, the rejection policy evaluated annotators' submissions. If an annotator's work failed the policy, it was rejected, and the HIT reopened to other annotators until three quality submissions were obtained.

6. New dataset description

This section outlines essential features and quality evaluations of the new abusive language dataset named ALDIPF: An Abusive Language Dataset Including Psychological Features. This new dataset is created through the processes in Section 5.

6.1. Dataset overview and features

Through the process mentioned in Section 5, 38,009 unique messages (or 114,027 messages in total) were annotated by 505 annotators, and annotators' SGABS scores (*Psychological Features*) were aggregated with the *Triggers* (messages) and *Consequence* (class labels). The configuration of this dataset is illustrated in Table 6. This new dataset was named — ALDIPF: Abusive Language Dataset Including Psychological Features, and it is available on our GitHub page https://github.com/tsungcheng-yao-griffith/ALDIPF Link.

ALDIPF has three columns – *Trigger*, *Psychological Features*, and *Consequence*, and each column is Tab-delimited (see Table 7). The *Trigger* column contains messages, and they are not pre-processed. The *Psychological Features* column shows the SGABS scores of an annotator who tagged this message, and these scores are presented in the same sequence as they are in the original SGBAS. The *Consequence* column only has one of two class labels – *Harmful* or *Neutral*. 65.6% of these messages were tagged *Neutral*, and the rest were *Harmful*. Furthermore, if the class labels are aggregated with a voting scheme, the rate of *Neutral* is 73.5%, and the *Harmful* one is 26.5%.

Regarding the inter-rater agreement, 33.5% of messages reached full consensus (29.2% *Neutral*, 4.3% *Harmful*). Of the remaining messages, 44.3% were tagged *Neutral* with one disagreement, and the rest were tagged *Harmful* with one annotator disagreeing. Nevertheless, it is worth noting that the ABC model posits that individuals may react differently to the same messages, and these reactions can all be authentic and genuine. As a result, once annotators' submissions pass the rejection policy, they are deemed valid and reliable.

Table 6
ALDIPF Description. The ratio of class labels varies by folded/unfolded structure.

Corpus	*Target Messages* from Founta dataset (see Section 5.4)
Annotation Method	crowdsourcing methods
Annotators	505 anonymous users on MTurk
Class Labels	Harmful and Neutral
Outcomes	114k annotated messages and users' Psychological Features
Ratio of Unfolded Class Labels	Neutral 65.6% Harmful 34.4%
Ratio of Folded Class Labels	Neutral 73.5% Harmful 26.5%

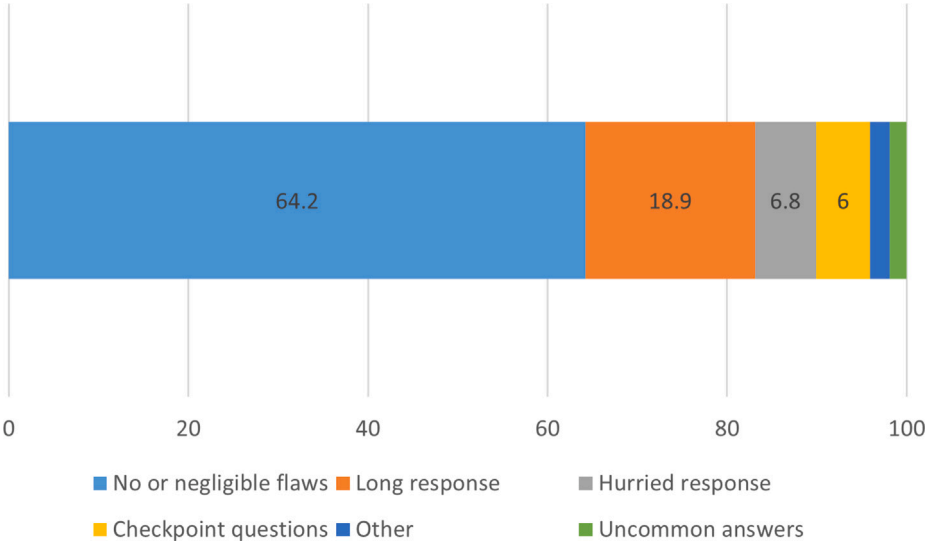


Fig. 6. Distribution of Minor Flaws. *Uncommon answer* counts 2.2%, and *Other* is 1.9%.

Table 7
An Illustration of ALDIPF. The same message can be associated with multiple class labels. That structure is rooted in the ABC model, where Consequences (Class Labels) are co-created by Triggers (Messages) and Personal Traits (Psychological Features). Note 1: *Annotated By* was not presented in the open-access dataset. Note 2: In this example, the *Psychological Features* are presented in SGABS items, which is the same format used for input in the Evaluation section.

Message	Annotated by	Psychological features	Class label
This is an example	Annotator 1	17,15,13,15,11,10,16,80	0
This is an example	Annotator 2	17,16,18,18,13,13,18,96	0
This is an example	Annotator 3	14,13,14,13,11,10,16,77	1

6.2. Dataset quality

We published 695 HITs on MTurk, and 576 were accepted. The acceptance rate is 82.8%, within the acceptance range suggested by previous studies (Barends & de Vries, 2019; Zhou & Fishbach, 2016). Furthermore, within these 576 accepted HITs, 64.2% of them passed our rejection policy with no or negligible flaws, such as missing one checking point question. The rest of the HITs passed the rejection policy with minor flaws. Fig. 6 indicates the distribution of these minor flaws. The *Other* category refers to HITs that either had a combination of flaws or were accepted as exceptions.

Long response time is the most common minor flaw among accepted HITs. While it can sometimes indicate inattentiveness, this statement is only sometimes valid. **Long response time only implies that annotators are distracted after accepting HITs instead of when doing HITs**. Annotators might return to the task later when motivated. To differentiate between these scenarios, checkpoint questions serve as a detector. According to our rejection policy, submissions are rejected if annotators take longer to complete the task and fail all checkpoint questions. Since the HITs in question passed the checkpoint questions, we suggest that most with longer response times are as reliable as those with no or negligible flaws.

7. Evaluation

This section assesses the **Research Objectives Two and Three** outlined in two studies. Study One will examine the general suitability of psychological features in the context of abusive language detection. Study Two will employ an experimental method to assess the effectiveness of including psychological features on the performance of detection systems. The studies were conducted on the newly introduced dataset — ALDIPF.

7.1. Study one: Feasibility of psychological features in the context of abusive language detection

In this section, several statistical analyses were performed to (i) investigate the difference in psychological features between ALDIPF and prior studies and (ii) assess whether specific psychological features are associated with the class labels. Specifically, we concentrated on whether particular psychological features demonstrate the same attributes, such as being positively and negatively associated with perceiving harm, as psychological theories suggested. The analysis will help clarify to what extent the ABC model and psychological theories can be useful for abusive language detection studies.

Furthermore, it is important to note that the following analytical results will be evaluated and validated by prior psychological studies rather than computer science studies for two reasons. First, this section aims to assess the feasibility of psychological features in abusive language detection, specifically exploring the applicability of the ABC model and related psychological theories. Second, no prior computer science studies offer a direct comparison due to the fundamental differences in underlying theories.

Table 8
Mean values (Standard Deviation) on GABS/SGABS items. DF = Demands for Fairness; NAch = Need for Achievement; NComf = Need for Comfort; OD = Other-Downing; NApp = Need for Approval; and SD = Self-downing; Ra = rationality; IR = Irrationality Note 1: These values were normalised by dividing the item score by the number of questions. Note 2: Irrationality scores are not directly comparable due to differences in the number of questions between GABS/SGABS.

Item	Linder	MacInnes	Owings	Ours
DF	3.04(0.9)	3.30(0.6)	3.23(0.7)	3.51(0.9)
NAch	2.39(0.8)	2.77(1.0)	2.94(0.7)	3.45(0.8)
NComf	2.78(0.9)	3.17(0.7)	2.55(0.7)	3.44(0.9)
OD	2.15(0.8)	2.27(0.6)	2.52(0.8)	3.45(0.9)
NApp	2.30(0.7)	2.53(0.7)	2.70(0.8)	3.38(0.9)
SD	2.47(0.5)	2.45(0.9)	2.02(0.6)	3.13(1.0)
Ra	4.11(0.5)	3.62(0.7)	3.82(0.5)	3.71(0.7)
IR	–	–	–	–

Lastly, as a pioneer study, we concentrate on general tendencies rather than specific thresholds. For instance, almost all prior psychological studies suggest that the Rationality score should be negatively associated with perceiving harm; however, the *"correct"* confidence level and strength of this correlation are usually not specified because they can vary by studies and experiment setting. As a result, most of the analysis in Study One will focus on evaluating the general attributes that particular psychological features should possess rather than matching *"correct"* numbers.

7.1.1. Difference in psychological features between ALDIPF and prior studies

ALDIPF was compared with previous psychological studies to clarify whether it is representative of the ordinary population. We surveyed previous psychological studies that adopted either SGABS or GABS, as these items can be interchangeable. After removing studies concentrated on specific settings, such as education and workplaces, three studies were selected for comparison. These studies are:

1. Lindner et al. (1999): based on 90 college students.
2. MacInnes (2003): based on 14 mental health outpatients and 27 nursing students.
3. Owings, Thorpe, McMillan, Burrows, Sigmon et al. (2013): based on 544 University students.

According to the ANOVA test, ALDIPF is significantly different from all selected studies on almost all items except Rationality (Ra). To be more specific, as shown in Table 8, **users from ALDIPF possess (i) higher scores on all irrational items (DF to SD)** and (ii) slightly lower scores on Rationality.

One possible explanation for this disparity lies in a two-layer self-selection bias, where annotators make decisions before undertaking annotation tasks, decisions influenced by various factors (Burnham, Le, & Piedmont, 2018; Chandler & Paolacci, 2017). The first decision involves registering on crowdsourcing platforms, often driven by financial motives. The second decision is whether annotators prefer to participate in particular types of annotation tasks. Consequently, these self-selection processes may attract crowdsourcing users from particular socioeconomic backgrounds, potentially endowing them with distinct personal psychological features compared to the ordinary population (Burnham et al., 2018). For instance, previous psychological research has noted that annotators on crowdsourcing platforms tend to possess particular traits, such as heightened sensitivity, openness, and introversion (Colman et al., 2018). We conjecture that these sensitive and introverted psychological features could also influence our SGABS results, potentially contributing to the observed higher *Irrationality* scores ($\sum$ DF to SD) compared to the broader population.

Table 9
Correlation Coefficient Between Critical Psychological Features (all p values are less than 0.05). Rationality and Irrationality scores are offset rather than bipolar.

Psychological features	Correlation coefficient
Ra and IR	0.481
SD and IR	0.855
Ra and SD	0.272

7.1.2. Feature importance analysis

Psychological studies suggest that clinical indicators (See Table 3) can assist in predicting an individual's tendencies or distinguishing specific groups of people from others (Lindner et al., 1999; MacInnes, 2003; Turner et al., 2018). To assess this attribute in the context of abusive language detection tasks, a feature importance analysis was performed using a Random Forest classification model with an accuracy of 70% in predicting binary class labels - *Neutral* and *Harmful*. Notably, this analysis aimed to compare feature importance against prior psychological studies rather than establish a new correlation between psychological features and class labels. Therefore, accuracy is not the only criterion we considered. We also prioritised the generalisability of the model's results.

As shown in Fig. 7, the Self-down score emerged as the most significant feature, showing a mean decrease in impurity greater than 0.2. The Total score is another crucial feature derived from subtracting the Rationality score from the Irrationality score. Additionally, both Rationality and Irrationality scores played crucial roles in predicting class labels. These features, which are also clinical indicators (see Table 3), significantly affect an individual's general well-being and perception of triggers (Ciarrochi & Bailey, 2009; David, DiGiuseppe, Dobrean, Păsărelu, & Balazsi, 2019; DiGiuseppe et al., 2013, 2018; Dryden, 2005; Ellis, 1991; Owings et al., 2013). Therefore, our findings are consistent with prior psychological studies.

7.1.3. Correlation between psychological features

We also performed a correlation coefficient to examine the relationship between critical psychological features (see Table 9). Self-down and Irrationality demonstrate a robust positive linear correlation. Notably, Rationality and Irrationality scores illustrated a moderate positive linear relationship. In other words, Rationality and Irrationality scores are not bipolar. Despite the absence of explicit psychological assumption, we suggest this result does not contradict prior psychological studies suggesting that Rationality and Irrationality scores are offset rather than bipolar (Ellis, 1991; Ellis & MacLaren, 1998).

7.1.4. Psychological features and perceiving harm

According to psychological studies, Total score, Self-down, and Irrationality are usually positively correlated with the perception of harm (David et al., 2019; DiGiuseppe et al., 2018; Owings et al., 2013). In other words, these scores should demonstrate a positive correlation with the *Harmful* class label. By contrast, Rationality is believed to negatively correlate to the perception of harm. Nevertheless, as discussed, the *"correct"* confidence level and strength of this correlation are usually not specified.

We leverage a Logistic Regression model to analyse the correlation between psychological features and the *Harmful* class label. Particularly, we focused on four psychological features (i) *Rationality*, (ii) *Irrationality*, (iii) *Self-down*, and (iv) *Total score* because of their significance in clinical settings and predicting class labels (see Section 7.1.2).

Analysis results (see Fig. 8) suggested that Irrationality, Self-down, and Total scores positively correlate with the *Harmful* class label, while Rationality scores exhibited negatively correlate. These findings are supported by established psychological assumptions. In addition, these results also indicate that the impact of Irrationality, Self-down, and Total on the *Harmful* class label may be more pronounced with higher score values.

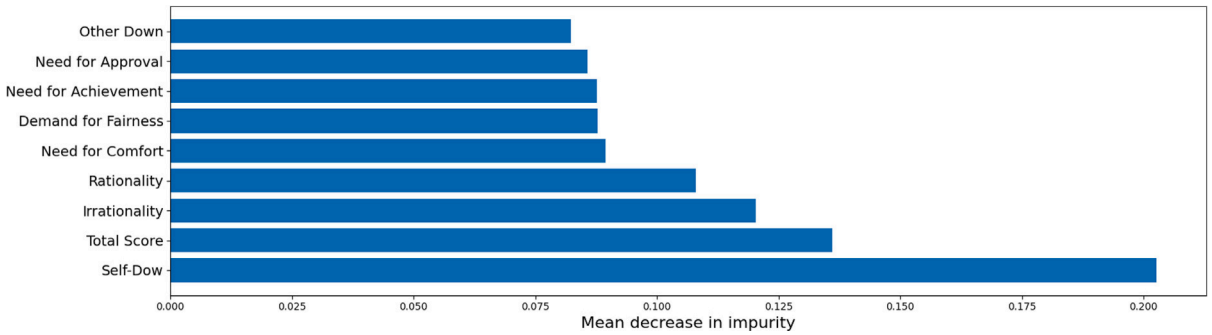


Fig. 7. Feature Importance Analysis. The X-axis denotes the mean decrease in impurity of each feature.

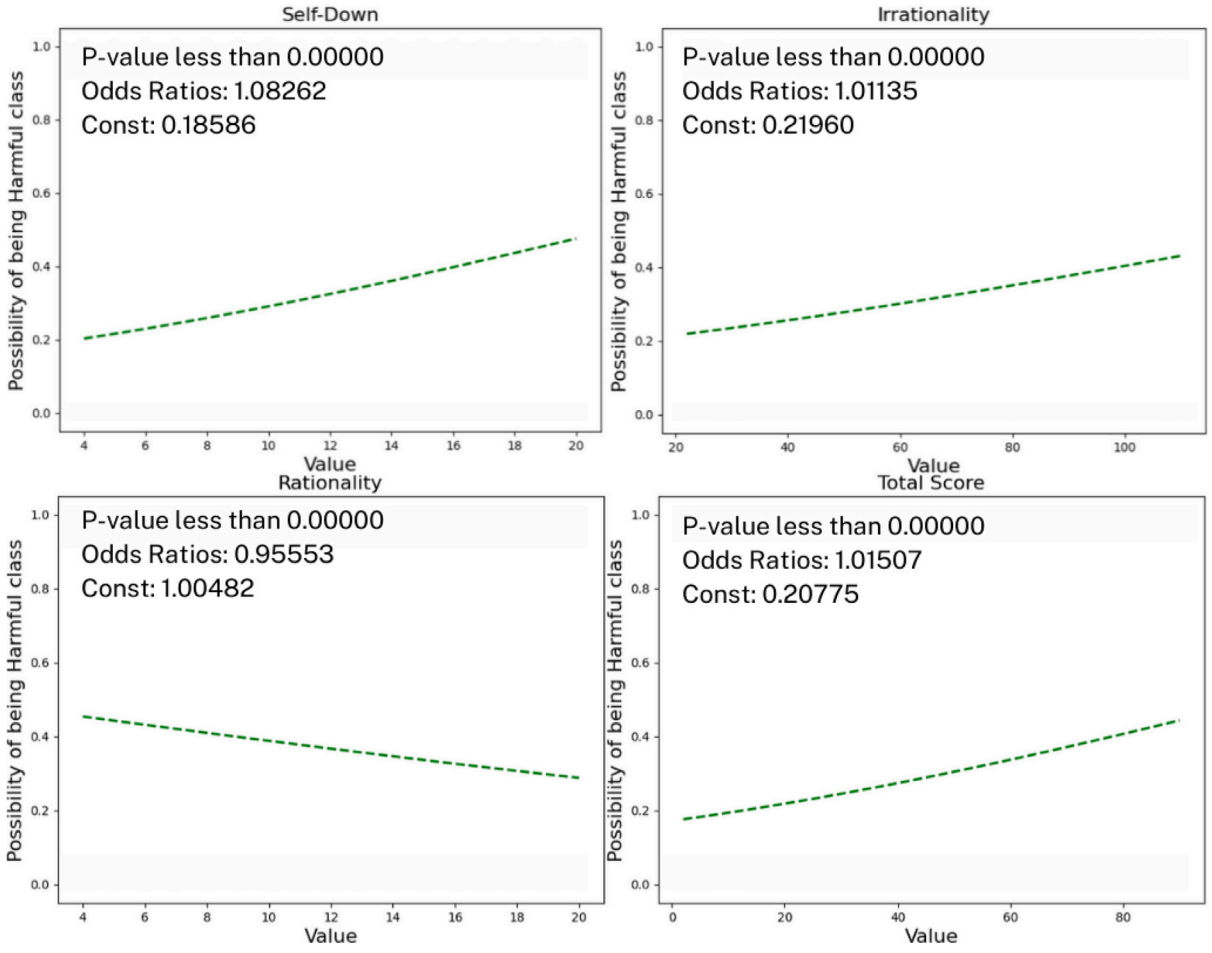


Fig. 8. One Factor Analysis. The y-axis denotes the possibility of being *Harmful* class.

7.1.5. Discussion for study one

Collectively, our analysis results provided strong evidence that psychological features are applicable in the context of abusive language detection. Despite the divergence between ALDIPF and prior psychology datasets, all other analysis results are supported by psychological studies. Particularly, we noticed the significance of four psychological features (i) Rationality, (ii) Irrationality, (iii) Self-down, and (iv) Total score. These features exhibit higher importance scores and demonstrate psychologically supported correlations with the *Harmful* class label. Consequently, we propose that these psychological features can serve as the foundation for developing personalised detection systems.

In addition, we noticed two important findings that may make future studies reconsider the competency and the limitation of abusive language datasets based on crowdsourcing methods, as these datasets may exhibit potential biases. These findings are:

- Finding 1: Users from crowdsourcing platforms might display a higher sensitivity level than ordinary users. In Section 7.1.1, our analysis demonstrated that crowdsourcing users exhibit higher Irrationality and Self-down scores than the border population. This observation aligns with the findings of a previous psychological study.
- Finding 2: The sensitivity level can subsequently impact the proportion of the *Harmful* class label. Section 7.1.4 established a positive association between Irrationality and Self-down scores and the *Harmful* class label.

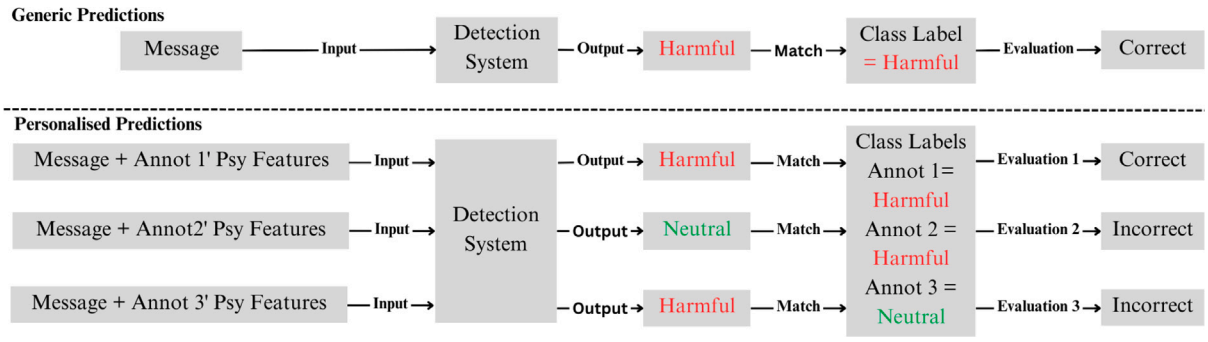


Fig. 9. Difference in Evaluation. Our evaluation cannot be directly compared to prior studies because ours concentrates on personalised predictions that provide **tailored predictions for the same message** across users based on their psychological features.

Considering these two interconnected findings, we conjecture that crowdsourcing-based datasets should contain a higher proportion of the *Harmful* class label. As a result of the potentially elevated proportion of the *Harmful* class label, we argue that these datasets may not be capable of developing detection systems for border user groups that do not possess the same sensitivity level as crowdsourcing users. Therefore, we argue further studies must be undertaken to (i) assess the impact of two-layer self-section bias in the context of abusive language studies and (ii) examine users' psychological features across various platforms and topics.

Additionally,

7.2. Study two: Impact of incorporating psychological features on making personalised predictions

An experimental approach was employed to assess the impact of psychological features on detection systems. Also, it is worth emphasising that **the following evaluation result cannot be directly compared to prior studies** because our detection systems were built for a distinct downstream task - **personalised predictions** (see Fig. 9). Most prior studies focus on one-size-fits-all predictions, where one message is always associated with one class label. In contrast, the following evaluation focused on the capability of detection systems to make personalised predictions, meaning that they can provide tailored predictions for the same message across users based on their psychological features. In other words, due to different psychological features among users, one message could be deemed harmful to one user while being perceived as neutral by others.

7.2.1. Experiment settings

We constructed two detection systems — experimental and control systems (end with _C). Both experimental and control systems generate class labels using two inputs: messages and psychological features. The sole distinction is that the control systems were built on fabricated psychological features where all items were replaced with zero. Subsequently, we assess the performance of the detection systems using a testing dataset comprising messages and authentic psychological features. Consequently, if an experimental system outperforms its counterpart (_C), we can infer that this improvement is attributed to proposed psychological features.

7.2.2. Dataset and preprocessing

ALDIPF was used for this evaluation. First, the messages were normalised and cleaned. Usernames, URLs, and Names were replaced by tokens. In addition, all letters were converted to lowercase. After that, the SGABS scores were transferred to 7 items (psychological features) in numeric format. Then, the dataset was randomly split into training (80%) and testing (20%) datasets. All models were trained on the training dataset and evaluated by the testing dataset.

7.2.3. Language models

We employed the following six language models:

Logistic Regression Models: LR models combined messages and numeric SGABS data to make predictions.

SVM Models: SVM models had similar settings to LR and were trained using the scikit-learn library with a radial basis function kernel.

CNN Models: we adopted the architecture proposed by Kim (2014). These models had two separate inputs — one for the message data and one for the psychological features. The messages were processed through two convolutional layers, followed by global max pooling. Respecting the numeric psychological features, it was fed into three dense layers and concatenated with text data. A dropout layer was added after a fully connected layer to prevent overfitting by randomly deactivating a fraction of the neurons during training. Additionally, we implemented an early stopping mechanism, which halts training when the model's performance on the validation set ceases to improve for 3 iterations.

LSTM Models: LSTM models had settings similar to the CNN models. The same pre-trained FastText embeddings were used. The messages were processed through an LSTM layer, while the numeric data was passed through three dense layers. The same dropout layer and early stopping mechanism, as mentioned above, were added after an LSTM layer to prevent overfitting.

FastText Model: Wiegand, Ruppenhofer, and Kleinbauer (2019) leveraged a FastText to achieve the highest F1 score on the original Founta dataset. Considering the origin of ALDIPF (see Section 5.4), a FastText model was constructed and tested; however, by the nature of Fasttext, it cannot process numeric SGABS data. Consequently, the FastText model was trained solely on message and class labels.

DistilBERT Model: This distilled BERT model has demonstrated efficacy across various abusive language detection tasks (Sanh, Debut, Chaumond, & Wolf, 2020). We built a detection model using HuggingFace's framework (distilbert-base-uncased). The model was optimised with techniques such as learning rate scheduling, and early stopping to enhance performance. However, similar to FastText, by design, DistilBERT cannot process numeric SGABS data either. As a result, the DistilBERT model was also trained solely on message and class labels.

RoBERTa Model: RoBERTa, an improved BERT-based model by training the model longer, with bigger batches over more data (Liu et al., 2019). We implemented a detection model using Hugging Face's framework (roberta-base) with the same optimisation processes and limitations as the DistilBERT model.

DeBERTa Model: DeBERTa enhances RoBERTa with a disentangled attention mechanism that separately encodes word content and position, along with an enhanced mask decoder that incorporates absolute positions for masked token prediction (He, Liu, Gao, & Chen, 2021). It is implemented using Hugging Face's framework (deberta-v3-base) with the same optimisation processes and limitations as the DistilBERT model.

Table 10
Experiment Results. Figures are highlighted in bold if an experimental model outperforms its counterpart.

Model	Weighted F1	Macro F1
Control Group		
FastText	0.529	0.400
DistilBERT	0.551	0.436
RoBERTa	0.543	0.421
DeBERTa	0.529	0.399
SVM_C	0.563	0.503
LR_C	0.535	0.400
CNN_C	0.572	0.498
LSTM_C	0.568	0.499
Experimental Group		
SVM	0.500	**0.511**
LR	**0.539**	**0.410**
CNN	**0.610**	**0.556**
LSTM	**0.574**	**0.511**

7.2.4. Experiment results

Psychological features demonstrated notable enhancement in neural network-based detection models (see Table 10). **The experimental CNN model surpassed the control model by 4%–5% points on Weighted and Macro F1 scores.** We suggest that the psychological features contribute to this improvement, as it is the only difference between the two models. **Comparable improvement was also observed in LSTM and LR pairs**, although the extent of improvement was not as pronounced as observed in CNN models. However, SVM models demonstrated mixed results; the Weighted F1 score significantly declined, while the Macro F1 score slightly improved.

Given the potential impact of the CNN model on our experimental results, we discuss its strengths and limitations in this specific application. CNNs excel in efficient pattern recognition, fast training, and seamless integration of numeric psychological features, making them suitable for lightweight applications. In our experiment, these advantages enable the evaluation of psychological feature integration with minimal control variables, such as data normalisation processes and the architectural complexities of advanced language models, which potentially enhance performance but compromise explainability.

However, we also recognised that CNNs struggle to capture long-range dependencies and contextual semantics, as their fixed receptive field, determined by kernel size, limits their ability to interpret varying contexts and longer sentences. Consequently, CNN models may over-rely on surface-level features while ignoring cross-sentence dependencies, which could negatively impact their performance when processing longer or more context-dependent messages.

7.2.5. Discussion for study two

Considering the consistent improvement across almost all language models, we suggest that psychological features demonstrate the substantial potential to enhance the effectiveness of detection systems in making personalised detections. In addition, recognising the findings in Study One, we posit that this performance improvement is not arbitrary; rather, it stems from the established correlations between psychological features and the *Harmful* class label. In other words, this improvement can be further assessed and explained by psychological theories and assumptions.

Furthermore, we noticed that while Fasttext, DistilBERT, RoBERTa, and DeBERTa detection systems have previously demonstrated exceptional performance in other studies, they did not yield the highest performance in our experiment setting. This discrepancy may be attributed to two underlying factors. First, explicit abusive languages were removed from ALDIPF. As mentioned in Section 5.4, we have eliminated (1) messages that contain obvious abusive patterns and (2) messages that are obviously not abusive. In other words, by design, ALDIPF focuses on handling ambiguous messages, which are challenging to classify by nature (Sandri et al., 2023). Secondly, ALDIPF recognised the different *Consequences* that a *Trigger* could cause. ALDIPF allows one message (*Triggers*) to have more than one class label (*Consequences*) since *Consequences* are correlated with *Psychological Features*. As a result, a third of the messages were tagged with two class labels, and they could only be differentiated through psychological features. This multifaceted nature of class labels adds complexity to the classification task.

8. Related work

This section first discusses several prior studies related to disagreement among annotators. After that, we compare our research with two similar studies.

8.1. Disagreement among annotators

Many detection systems have been developed to identify abusive textual patterns, such as combinations of letters, keywords, or phrases (Chhabra & Vishwakarma, 2023; Jahan & Oussalah, 2023; Nejadgholi, Fraser, & Kiritchenko, 2022), because, according to statistical analysis, these patterns strongly correlate with abusive language. Therefore, NLP studies have developed different taxonomies and models to identify these patterns.

Extensive research has been dedicated to comprehending the linguistic disparities among various forms of abusive language. Waseem, Davidson, Warmsley, and Weber (2017) suggest a typology of abusive language based on two features (1) target and (2) the degree of explicitness. This typology helps researchers better identify the correlation between lexical feature sets and abusive languages. For instance, swear words and sarcasm show different lexical features. Similarly, Balayn, Yang, Szlavik, and Bozzon (2021) have proposed a reconciled taxonomy highlighting the affiliation of different types of abusive language. This taxonomy clarifies relationships and definitions of different types of abusive language, such as aggression, harassment and cyberbullying. Other researchers have created definitions and taxonomies via experiments and statistical inferences (Founta et al., 2018; Lewandowska-Tomaszczyk, Bączkowska, Liebeskind, Oleskeviciene, & Žitnik, 2023). Finally, some researchers have focused on specific types of abusive language that target certain demographic groups, such as gender and race (Garg et al., 2023; Parikh et al., 2019; Şahinüç et al., 2023).

However, despite the positive outcome of leveraging textual patterns of abusive language, the definition of abusive language remains ambiguous (Balayn et al., 2021; Chhabra & Vishwakarma, 2023; Ejaz et al., 2024; Jahan & Oussalah, 2023; Vidgen & Derczynski, 2020), and none can capture the entire spectrum of abusive language. As a result, researchers may employ varying taxonomies and definitions that do not necessarily align with each other (Mishra et al., 2020; Vidgen & Derczynski, 2020), thereby causing further confusion in abusive language studies.

In addition, previous research has demonstrated that differences among annotators can significantly impact the accuracy of abusive language detection systems. Waseem (2016) found that amateur annotators tended to tag more messages as hate speech than expert annotators, highlighting the need for further investigation into sociolinguistic features. Al Kuwatly, Wich, and Groh (2020) also suggested that annotators' demographic features play a notable role in identifying abusive language. Similarly, Larimore, Kennedy, Haskett, and Arseniev-Koehler (2021) noted that annotators' racial identity (White and non-White) impacts their understanding of tweets with a high prevalence of certain racially charged topics. Sap et al. (2022) found that some beliefs and identities are strongly associated with tagging of abusive language, emphasising the need for additional features that can improve the accuracy and fairness of detection systems. Lastly, Balayn et al. (2021) argued that the lack of features relevant to individual

Table 11
Comparison between Ours and Previous Studies by Attributes. *Note:* (Srinivas et al., 2024) includes additional features beyond those listed here; however, we have included only the relevant ones in this table.

Research	Proposed features	Feature source	Underpinned theory	Personalised prediction
Ours	SGABS	Survey Participants	The ABC Model	Yes
Balakrishnan, Khan, and Arabnia (2020)	Big Five & Dark Triad	IBM Watson	Unaddressed	No
Kocoń et al. (2021)	Demographic Features	Metadata	Unaddressed	Yes
Beck, Schuff, Lauscher, and Gurevych (2024)	Sociodemographic Features	Survey Participants	Unaddressed	Yes
Srinivas et al. (2024)	Big Five, Dark Triad & Schwartz	Combination of Synthesis, Metadata, & Survey	Partially addressed	Yes

differences is a pressing issue related to the fairness and explainability of detection systems.

Regarding the root cause of difference among annotators, some studies have suggested that one of the key factors is their subjectivity (Uma et al., 2022). For instance, Plank (2022) proposes the term *human label variation* to capture the root causes of disagreement among annotators, with annotators' subjectivity being one of these causes. Sandri et al. (2023) also suggested that subjectivity plays a substantial role in disagreement among annotators on subjective tasks, such as abusive language detection. Srinivas, Das, and Pulabaigari (2024) found that users' inner personalities impact hate speech diffusion. Wan, Kim, and Kang (2023) propose a demographic-based disagreement predictor for estimating inter-annotator disagreement in subjective tasks. Similarly, considering the impact of subjectivity among annotators, Cabitza, Campagner, and Basile (2023) presented a perspectivist stance with multi-channel to process both subjective and objective tasks.

8.2. Comparative studies

To highlight the differences and contributions of this research, we compared our research with two studies that aim to enhance detection systems by either integrating psychological features or recognising the different perceptions among individuals (See Table 11). We concentrated on comparing two attributes: (1) whether a study is grounded in any theory and (2) whether a study supports personalised prediction.

Balakrishnan et al. (2020) enhanced detection systems by introducing psychological features — Big Five and Dark Triad psychological features. However, a primary limitation of this research is that these features were generated by IBM Watson Service based on users' Twitter portfolios rather than being collected from authentic users. In addition, the theoretical framework underpinning these psychological features remains unaddressed. Lastly, as this research concentrated on the personal traits of the perpetrator, these features do not support personalised predictions.

On the other hand, Kocoń et al. (2021) recognised the different perceptions among groups and users. By incorporating user demographic features, such as gender, English as first language, age group, and education, their detection systems can make adjusted predictions based on personal or community profiles. However, despite the positive experiment outcomes, the absence of a robust theoretical foundation to explain this enhancement limits the impact of their research. In addition, Hung, Lauscher, Hovy, Ponzetto, and Glavaš (2023) suggested that the improvement of the downstream performance gains from demographic features does not stem from demographic knowledge.

Beck et al. (2024) evaluated the effectiveness of incorporating users' sociodemographic attributes, including gender, race, age range, education level, and political affiliation, into Large Language Models (LLMs) based detection systems. This evaluation was in response to concerns about introducing bias and stereotypes through sociodemographic attributes (Cheng, Durmus, & Jurafsky, 2023; Deshpande, Murahari, Rajpurohit, Kalyan, & Narasimhan, 2023). The authors concluded that while there are potential benefits to using these attributes, they must be applied cautiously, as outcomes can significantly vary depending on the settings.

Srinivas et al. (2024) developed a hate speech diffusion model incorporating psycho-sociological features from sources like the Dark Triad, Big Five personality traits, Schwartz values, and diffusion data. The study found strong associations between Machiavellianism and Narcissism with sexism and racism. However, it does not address the underlying theory linking abusive language to the selected features, particularly the effectiveness of the Big Five and Schwartz values. Consequently, the lack of an underpinning theory limits the generalisability of the experimental results.

9. Practical implications and conclusion

This paper was the first and pivotal step in integrating psychological theories and features to personalise abusive language detection systems. We leveraged psychological theories to conceptualise and transfer annotators' subjectivity into a series of numeric features that are encodable in detection systems. In addition, our findings demonstrate compelling evidence underscoring the significance of users' psychological features in abusive language detection studies. The practical implications for future studies are as follows:

- **Incorporate psychological features to personalise abusive language detection systems.** Collectively, our research results make for a strong empirical case that users' psychological features can improve detection systems in generating personalised predictions. In addition, recognising the findings in Study One, we posit that this improvement is not arbitrary; rather, it is undermined by the established psychological theories. Future studies may conduct a series of experiments to assess the confounding factor.
- **Redefine the factors contributing to individuals experiencing negative feelings.** Considering our research results, psychological features constitute another crucial yet overlooked factor in abusive language studies. As discussed in Section 7.1.4, particular psychological features exhibited either positive/negative correlations with the *Harmful* class label, which are consistent with the psychological theories. In other words, *Consequences* (Class labels) are co-creations of *Triggers* (Messages) and *Psychological Features*. This finding is particularly critical for understanding and addressing the disagreement among annotators in subjective tasks.
- **Facilitate further studies to investigate the role of personal traits in abusive language detection studies.** As discussed in Section 5.3, personal traits can be measured by various tools that concentrate on different aspects. By substituting the corpus and psychological tool, the proposed data preparation framework enables future studies to investigate the role of personal traits in different experiment settings.

Table A.12
Statement for SGABS's Sub-scale.

Sub-scale	Statement
Need for achievement	It's unbearable to fail at important things, and I can't stand not succeeding at them.
Need for approval	When people who I want to like me disapprove of me or reject me, I can't bear their disliking me
Need for comfort	It's unbearable being uncomfortable, tense, or nervous, and I can't stand it when I am.
Demand for fairness	It is awful and terrible to be treated unfairly by people in my life.
Self-downing	If important people dislike me, it is because I am an unlikable, bad person
Other downing	If people treat me without respect, it goes to show how bad they really are.
Rationality	I have worth as a person even if I do not perform well at tasks that are important to me

- **Reconsider the fairness and bias of detection systems from psychological lenses.** Future studies should be attentive to potential biases and fairness concerns within user groups distinguished by diverse personal traits. These concerns can arise from at least two aspects — data collection methods and a voting scheme. First, as discussed in Section 7.1.1, two-layer self-selection may introduce bias to a dataset. Second, a voting scheme can exclude the opinions of users with sensitive personal traits. With a voting scheme, class labels are usually determined by the majority's opinions; however, minority opinions should not be automatically dismissed as inattentive responses; instead, they can also be genuine and valuable.

Lastly, it is essential to acknowledge the limitations of the research. First, this pilot study solely focused on generic abusive language and excluded explicitly abusive language (see Section 5.4), limiting the generalisability of our findings. Additionally, the context of messages was not considered, potentially impacting the correlation between *Consequences* and *Triggers*.

Second, the findings in this research may not generalise across different cultural and linguistic contexts, as the ABC model was primarily developed and validated in Western communities.

Third, despite countermeasures, this study relies on self-reported surveys from annotators, which may introduce bias due to distractions or misunderstandings of the measurement scale, affecting the accuracy of psychological profiles.

Fourth, demographic factors information could provide additional layers of insight into how psychological features and language use vary across different groups. Future studies may further explore the interaction between demographic diversity and psychological features, along with its implications.

These limitations highlight the need for future studies to address a broader range of abusive language types and consider contextual factors for more accurate classification.

In subsequent work, we will conduct a series of experiments to assess the confounding factors — primarily psychological features and the structure of language models. Furthermore, we plan to leverage feature engineering and advanced modelling techniques to further augment the performance of personalised detection systems.

10. Ethical consideration

Our survey was reviewed and approved by the Griffith University Human Research Ethics Committee (GU Ref No: 2022/170). We took several steps to protect annotators' confidentiality and privacy. No personally identifiable information was collected, and the annotators' IDs were anonymised to ensure that responses could not be traced back to individual participants. Original survey data was handled with strict confidentiality and securely stored in compliance with university guidelines.

To mitigate potential psychological harm, explicitly abusive language was removed from our experiment at our best effort. Before the survey, an information sheet and a consent form were presented. These documents detailed the risks, financial benefits, and details of taking our survey. Annotators were free to withdraw before submission. Notably, researchers' contacts were provided with the information sheet in case of any complaints or concerns.

Lastly, we recognise the ethical implications of using psychological profiling in abusive language detection, particularly the risk of misclassification. Misclassifying individuals based on their psychological profiles could lead to inappropriate conclusions or interventions. As a result, we emphasise that psychological features are only one of many factors influencing an individual's perception of abusive language. Relying solely on psychological profiling risks oversimplifying and stereotyping the nuanced process of interpreting harmful messages. We will continue to explore ways to mitigate these risks by analysing potential interactions between different psychological features in future research.

CRediT authorship contribution statement

Tsungcheng Yao: Conceptualization, Methodology, Software, Formal analysis, Investigation, & Writing - Original Draft. **Sebastian Binnewies:** Validation, Writing - Review & Editing, Supervision. **Ernest Foo:** Validation, Supervision, Writing - Review & Editing, Funding acquisition. **Masoumeh Alavi:** Validation, Writing - Review & Editing.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Grammarly in order to enhance the readability of the work. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Statement for SGABS's sub-scale

Table A.12, adapted from Lindner et al. (1999), provides an overview of the psychological features for clarity. It is important to note that these sub-scales, while simplified in our table for the purposes of this study, are more nuanced and complex in clinical settings.

Data availability

The link for data is provided in the manuscript.

References

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). Mturk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837.

Al Kuwatly, H., Wich, M., & Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms* (pp. 184–190). Association for Computational Linguistics, Online.

Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. *Information*, *13*(6).

Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, *164*, Article 114006.

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using twitter users' psychological features and machine learning. *Computers & Security*, *90*, Article 101710.

Balayn, A., Yang, J., Szlavik, Z., & Bozzon, A. (2021). Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *Transactions of the Society for Computer*, *4*(3).

Barends, A. J., & de Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in mturk personality research to improve data quality. *Personality and Individual Differences*, *143*, 84–89.

Barlett, C. P., Kowalewski, D. A., Kramer, S. S., & Helmstetter, K. M. (2019). Testing the relationship between media violence exposure and cyberbullying perpetration. *Psychology of Popular Media Culture*, *8*(3), 280–286, PsycInfo Database Record (c) 2021 APA, all rights reserved.

Beck, T., Schuff, H., Lauscher, A., & Gurevych, I. (2024). Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics (volume 1: long papers)* (pp. 2589–2615).

Bernard, M. E. (1998). Validation of the general attitude and belief scale. *Journal of Rational-Emotive and Cognitive-Behavior Therapy*, *16*(3), 183–196.

Burnette, C. B., Luzier, J. L., Bennett, B. L., Weisenmuller, C. M., Kerr, P., Martin, S., Keener, J., & Calderwood, L. (2022). Concerns and recommendations for using amazon mturk for eating disorder research. *The International Journal of Eating Disorders*, *55*(2), 263–272.

Burnham, M. J., Le, Y. K., & Piedmont, R. L. (2018). Who is mturk? Personal characteristics and sample consistency of these online workers. *Mental Health, Religion & Culture*, *21*(9–10), 934–944.

Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(6), 6860–6868.

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, *8*(5), 500–508.

Cheng, M., Durmus, E., & Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. In J. Rogers, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), pages* (pp. 1504–1532). Toronto, Canada: Association for Computational Linguistics.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon mechanical turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*(4), 347–361.

Chhabra, A., & Vishwakarma, D. K. (2023). A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, *29*(3), 1203–1230.

Ciarrochi, J., & Bailey, A. (2009). *A CBT-practitioner's guide to ACT: How to bridge the gap between cognitive behavioral therapy and acceptance and commitment therapy (Vol. 50)*.

Colman, D. E., Vineyard, J., & Letzring, T. D. (2018). Exploring beyond simple demographic variables: Differences between traditional laboratory samples and crowdsourced online samples on the big five personality traits. *Personality and Individual Differences*, *133*, 41–46, Examining Personality and Individual Differences in Cyberspace.

Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, *11*, 300–319.

David, D. O., DiGiuseppe, R., Dobrean, A., Păsărelu, C. R., & Balazsi, R. (2019). *The measurement of irrationality and rationality* (pp. 79–100). Cham: Springer International Publishing.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 1236–1270). Singapore: Association for Computational Linguistics.

DiGiuseppe, R. A., Doyle, K. A., Dryden, W., & Backx, W. (2013). *A practitioner's guide to rational-emotive behavior therapy*. Oxford University Press.

DiGiuseppe, R., Leaf, R., Gorman, B., & Robin, M. W. (2018). The development of a measure of irrational/rational beliefs. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *36*(1), 47–79.

Ding, Y., You, J., Machulla, T.-K., Jacobs, J., Sen, P., & Höllerer, T. (2022). Impact of annotator demographics on sentiment dataset labeling. In *Proc. ACM hum.-comput. interact. 6(CSCW2)*.

Dorol-Beauroy-Eustache, O., & Mishara, B. L. (2021). Systematic review of risk and protective factors for suicidal and self-harm behaviors among children and adolescents involved with cyberbullying. *Preventive Medicine*, *152*, Article 106684, Suicide Prevention.

Dryden, W. (2005). *Rational emotive behavior therapy* (pp. 321–324). Boston, MA: Springer US.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL, US: Harcourt Brace Jovanovich College Publishers.

Ejaz, N., Razi, F., & Choudhury, S. (2024). Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm. *Computers in Human Behavior*, *153*, Article 108123.

Ellis, A. (1991). The revised abc's of rational-emotive therapy (ret). *Journal of Rational-Emotive and Cognitive-Behavior Therapy*, *9*(3), 139–172.

Ellis, A., & MacLaren, C. (1998). *Rational emotive behavior therapy: A therapist's guide. The practical therapist series*. Atascadero, CA, US: Impact Publishers.

Extremera, N., Quintana-Orts, C., Mérida-López, S., & Rey, L. (2018). Cyberbullying victimization, self-esteem and suicidal ideation in adolescence: Does emotional intelligence play a buffering role? *Frontiers in Psychology*, *9*, 367.

Fortuna, P., Dominguez, M., Wanner, L., & Talat, Z. (2022). Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 11794–11805). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media (Vol. 12, No. 1)*.

Garg, T., Masud, S., Suresh, T., & Chakraborty, T. (2023). Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, *55*(13s).

Gavit, O. A., David, D., DiGiuseppe, R., & DelVecchio, T. (2011). The development and validation of the parent rational and irrational beliefs scale. *Procedia - Social and Behavioral Sciences*, *30*, 2305–2311, 2nd World Conference on Psychology, Counselling and Guidance - 2011.

Goldfried, M. R., & Sobocinski, D. (1975). Effect of irrational beliefs on emotional arousal. *Journal of Consulting and Clinical Psychology*, *43*(4), 504–510.

He, P., Liu, X., Gao, J., & Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention.

Hung, C.-C., Lauscher, A., Hovy, D., Ponzetto, S. P., & Glavaš, G. (2023). Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. In A. Vlachos, & I. Augenstein (Eds.), *Findings of the association for computational linguistics: EACL 2023* (pp. 1565–1580). Dubrovnik, Croatia: Association for Computational Linguistics.

Hunt, N. C., & Scheetz, A. M. (2019). Using mturk to distribute a survey or experiment: Methodological considerations. *Journal of Information Systems*, *33*(1), 43–65.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, Article 126232.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, *58*(5), Article 102643.

Larimore, S., Kennedy, I., Haskett, B., & Arseniev-Koehler, A. (2021). Reconsidering annotator disagreement about racist language: Noise or signal? In L.-W. Ku, & C.-T. Li (Eds.), *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media* (pp. 81–90). Association for Computational Linguistics, Online.

Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Oleskeviciene, G. V., & Žitnik, S. (2023). An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics*, *19*(1), 7–48.

Lindner, H., Kirkby, R., Wertheim, E., & Birch, P. (1999). A brief assessment of irrational thinking: The shortened general attitude and belief scale. *Cognitive Therapy and Research*, *23*(6), 651–663.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

MacInnes, D. (2003). Evaluating an assessment scale of irrational beliefs for people with mental health problems. *Nurse Researcher*, *10*(4), 53–67.

Madukwe, K., Gao, X., & Xue, B. (2020). In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the fourth workshop on online abuse and harms* (pp. 150–161). Association for Computational Linguistics, Online.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6).

Mishra, P., Yannakoudakis, H., & Shutova, E. (2020). Tackling online abuse: A survey of automated abuse detection methods.

Monks, C. P., & Smith, P. K. (2006). Definitions of bullying: Age differences in understanding of the term, and the role of experience. *British Journal of Developmental Psychology*, *24*(4), 801–821.

Nejadgholi, I., Fraser, K., & Kiritchenko, S. (2022). Improving generalizability in implicitly abusive language detection with concept activation vectors. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 5517–5529). Dublin, Ireland: Association for Computational Linguistics.

Otte, C. (2011). Cognitive behavioral therapy in anxiety disorders: Current state of the evidence. *Dialogues in Clinical Neuroscience*, *13*(4), 413–421.

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

Pamungkas, E. W., Basile, V., & Patti, V. (2023). Towards multidomain and multilingual abusive language detection: A survey. *Personal and Ubiquitous Computing*, *27*, 17–43.

Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., & Varma, V. (2019). Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 1642–1652). Hong Kong, China: Association for Computational Linguistics.

Plank, B. (2022). The problem of human label variation: On ground truth in data, modeling and evaluation. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10671–10682). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Poletto, F., Basile, V., & Sanguinetti, M. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, *55*, 477–523.

Pratkanis, A. R., Breckler, S. J., & Greenwald, A. G. (1989). *Attitude structure and function* (1st ed.). Psychology Press.

Romberg, J. (2022). Is your perspective also my perspective? Enriching prediction with subjectivity. In *Proceedings of the 9th workshop on argument mining* (pp. 115–125). Gyeongju, Republic of Korea: International Conference on Computational Linguistics, Online.

Sachdeva, P. S., Barreto, R., Vacano, C. von., & Kennedy, C. J. (2022). Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 1585–1603). New York, NY, USA: Association for Computing Machinery.

Şahinüç, F., Yılmaz, E. H., Toraman, C., & Koç, A. (2023). The effect of gender bias on hate speech detection. *Signal, Image and Video Processing*, *17*(4), 1591–1597.

Sandri, M., Leonardelli, E., Tonelli, S., & Jezek, E. (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* (pp. 2428–2441). Dubrovnik, Croatia: Association for Computational Linguistics.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., & Smith, N. A. (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* (pp. 5884–5906). Seattle, United States: Association for Computational Linguistics.

Srinivas, P. Y. K. L., Das, A., & Pulabaigari, V. (2024). Racists spreader is narcissistic; sexists is machiavellian influence of psycho-sociological facets in hate-speech diffusion prediction. *Expert Systems with Applications*, *247*, Article 123211.

Turner, M. J., Allen, M. S., Slater, M. J., Barker, J. B., Woodcock, C., Harwood, C. G., & McFayden, K. (2018). The development and initial validation of the irrational performance beliefs inventory (ipbi). *European Journal of Psychological Assessment*, *34*(3), 174–180.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2022). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, *72*, 1385–1470.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data. A systematic review: Garbage in, garbage out. *PLoS One*, *15*(12), Article e0243300.

van der Wal, O., Bachmann, D., Leidinger, A., van Maanen, L., Zuidema, W., & Schulz, K. (2024). Undesirable biases in nlp: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, *79*.

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

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138–142). Austin, Texas: Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the first workshop on abusive language online* (pp. 78–84). Vancouver, BC, Canada: Association for Computational Linguistics.

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of School Violence*, *14*(1), 11–29.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. 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)* (pp. 602–608). Minneapolis, Minnesota: Association for Computational Linguistics.

Xia, M., Field, A., & Tsvetkov, Y. (2020). Demoting racial bias in hate speech detection. In *Proceedings of the eighth international workshop on natural language processing for social media* (pp. 7–14). Association for Computational Linguistics, Online.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504.