Cognitive Computation (2025) 17:65
https://doi.org/10.1007/s12559-025-10423-3

RESEARCH

Discovering the Cognitive Bias of Toxic Language Through Metaphorical Concept Mappings

Mengshi Ge[1] · Rui Mao[1] · Erik Cambria[1]

Received: 27 March 2024 / Accepted: 28 January 2025

Abstract
With the prosperity of social media, toxic language spreading over social media has become an unignorable challenge for individual mental health and social harmony. Many researchers have studied toxic language identification to control or mitigate it. However, it still leaves a blank in the cognitive patterns of toxic language. Metaphors as a common feature in natural language connect literal and metaphorical meanings, which could be a useful tool to study the underlying cognitive patterns of the text. In this paper, we utilize a metaphor processing tool, MetaPro, to process a public toxic language dataset and analyze the cognitive biases between toxic and non-toxic language, multiple levels and subtypes of toxic language as well as toxic language mentioning different genders, sexual orientations, and races. Our study demonstrates that significant differences exist in cognitive patterns of the above-mentioned categories and analyzes the differences with machine learning methods.

Keywords Conceptual metaphor processing · Cognitive biases · Toxic language · MetaPro

Introduction

Language is the primary means by which humans communicate information, navigate social interactions, and express emotions. It is inherently complex, reflecting not only the content of what is said but also the underlying cognitive and cultural frameworks that shape human thought [1]. This complexity becomes even more pronounced in digital communication, where linguistic nuances can amplify misunderstandings or emotional responses [2, 3]. Spending increasingly more time on social media, people are more likely to get exposed to toxic language [4, 5]. Toxic language often involves expression that is harmful, offensive, or intended to provoke negative emotions and discomfort. The negative impacts of toxic language have been widely discussed and acknowledged, including causing psychological harm to individuals, exacerbating social divisions, and leading to a hostile environment [6, 7]. Toxic language is a significant concern in online platforms, where anonymity and distance can sometimes give rise to a higher possibility of such behavior [8, 9]. However, the cognitive patterns behind toxic language have yet to be discussed in the research community. Cognitive patterns provide insights into the underlying psychological mechanisms that lead individuals to use toxic language and contribute to a better understanding of human behavior in social interactions [10, 11]. Therefore, studying the cognitive patterns of toxic language is essential for preventing toxic behaviors, thereby promoting positive communication, and fostering a healthy society.

As a pervasive language phenomenon, metaphors exist commonly in daily communication, conveying cognitive patterns unconsciously. Many of our ways of thinking and acting are virtually represented metaphorically. Ottati et al. [12] demonstrated that a sports metaphor increases message-relevant elaboration and sensitivity to argument strength among individuals who enjoy sports. Conversely, the sports metaphor can reduce message-relevant elaboration and sensitivity to argument strength among individuals who dislike sports. Lakoff [13] analyzed how liberals and conservatives in the USA use different metaphorical frameworks to understand and discuss politics. Ang and Lim [14] found that

✉ Erik Cambria
cambria@ntu.edu.sg

Mengshi Ge
mengshi001@e.ntu.edu.sg

Rui Mao
rui.mao@ntu.edu.sg

[1] College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore

Published online: 14 February 2025

metaphors used in advertising could significantly impact consumers' perceptions of a brand's personality, showing the power of metaphors in shaping consumer attitudes and behaviors. Mao et al. [15] found that there is a moderate correlation between metaphorical cognition and voting behaviors, demonstrating how metaphors can reflect underlying cognitive and decision-making processes. Therefore, a metaphor is a suitable medium for researching cognitive patterns.

Lakoff and Johnson [16] proposed a perspective to understand metaphors, termed Conceptual Metaphor Theory, where metaphors are associated with concept mappings from target to source concepts. A target concept represents the actual object that the speaker aims to express, which is usually more abstract. A source concept represents the object that the speaker compares the actual object to, which is usually more concrete. For example, "she *attacks*[1] his argument." The word *attacks* indicates that the speaker compares ARGUMENT to WAR. The target concept is ARGUMENT, and the source concept is WAR. The concept mapping is (ARGUMENT, WAR).[2] The source concept WAR and the target concept ARGUMENT are from different domains. This metaphor frames ARGUMENT in the WAR shape, associating with aggression and weapon attributes. Concept mappings can reveal the implicit meanings of metaphors. Besides, the conceptualization of concept mappings enables to explain multiple metaphors with a general description, e.g., (ARGUMENT, WAR) also explains the metaphors in "Your claims are *indefensible*" and "I have never *won* an argument with him" [16]. Previous studies also have demonstrated that concept mapping can be employed for analyzing cognition. Crawford [17] from experimental social psychology revealed that the associations between affect and physical domains in conceptual metaphors could influence cognition, such as performance on attention, memory, and judgment tasks. Experimental studies showed that conceptual metaphors shape context-sensitive judgments of individuals about the meanings of idioms [18], coherent connections during text processing [19], and response to temporal events [20].

Conceptual metaphor processing focuses on identifying metaphors and then generating the target and source concepts for understanding the concept mappings of the identified metaphors [21]. We utilize MetaPro [22], a conceptual metaphor processing tool, to identify metaphors and generate corresponding concept mappings. According to the survey of [23], MetaPro is the only end-to-end system capable of generating concept mappings with state-of-the-art performance while effectively handling a broad spectrum of metaphor processing tasks. In this work, we processed a publicly available English dataset for toxicity classification sourced from a comment platform from 2015 to 2017. This dataset has diverse labels about toxicity, such as toxicity expressed by numeric values on a scale of 0 and 1, toxicity subtypes, and mentioning identities. Hence, we apply the Chi-square test for homogeneity and association rule mining (ARM) to analyze and compare the cognitive biases between different groups about toxicity.

In this work, we aim to address the following research questions:

(1) Is there any significant difference in cognitive patterns between toxic and non-toxic language, subdivided levels and subtypes of toxic language as well as toxic language mentioning different genders, sexual orientations, and races?
(2) What are the preferences of target concepts, source concepts, and concept mappings between the categories mentioned above?
(3) What are the dependency and associations of concept mappings between the above-mentioned categories? How are concept mappings associated with different categories?

Based on our experimental results and analysis, we have derived the following key findings:

(1) There are statistically significant differences in the target concepts, source concepts, and concept mappings between toxic and non-toxic language, subdivided levels and subtypes of toxic language as well as toxic language mentioning different genders, sexual orientations, and races. The findings indicate cognitive distinctions among speakers who speak various types of toxic language.
(2) Compared with non-toxic language, toxic language tends to deliver more negative sentiments and intentions. Metaphorical concepts in the obscene subtype show aggressive and strong emotions, while those in the sexual_explicit subtype involve direct physical or sexual comparison. Metaphors in toxic language mentioning males concentrate on strength and control, while those mentioning females express negative sentiments and undermine the competence of women. Toxic language mentioning heterosexuals emphasizes naturalness with direct metaphors, reflecting conventional views, while toxic language mentioning homosexuality often uses more complex concepts to reflect societal perceptions and actions that deviate from traditional norms.
(3) Some concept mappings contain explicit attributes of the corresponding categories, such as (UNPLEASANT_PERSON,

[1] Metaphors are shown in italics.

[2] We show concept mappings in (target, source) format, where the target and source concepts are separated by a comma. To distinguish, we enumerate concept examples in (concept1; concept2;...;) format, where the concepts are separated by semicolons.

DIFFICULTY) in toxic language, (PART, BODY_PART) in the obscene subtype, (ORGAN, REPRODUCTIVE_ORGAN) in the sexual_explicit subtype, and (CATTLE; CHICKEN) in toxic language mentioning Black people (*Black* is a label in the original dataset). The associations of concept mappings in toxic language mentioning males and Whites show significant similarity, which reflects a similar narrative of dominant groups in societal structures.

The contributions of this work are shown as follows: (1) We analyzed the cognitive patterns of large-scale toxic language based on concept mappings generated from metaphors. To the best of our knowledge, this is the first attempt to explore the cognitive patterns of toxic language and some subdivided categories from the perspective of conceptual metaphors. (2) We utilized certain machine learning methods to compare and summarize the experiment results of toxic language. Some of our findings are consistent with previous linguistic research, which can support the accuracy of our metaphor processing system and analyzing methods. Some findings could provide hypotheses for further linguistic or psycho-social studies.

Related Work

Toxic Language Studies

Computational methods have been widely used in other social science studies [24, 25]. The studies on biases related to toxic language cover various aspects. Some previous research has studied the biases in toxic language detection system [26, 27]. This sort of bias usually appears in the scenario that some text containing particular surface markers, such as African American English, is more likely to be classified as toxic, even though it does not contain toxicity. This false positive error makes the online platforms with the toxic language detection system more likely to remove content about minorities and exacerbate the discrimination against them in real life [28]. The social bias studied in the work of [29] is a preconceived belief toward or against specific social identities. The authors proposed a Transformer-based model to identify and mitigate social bias.

The analysis of toxic language usually uses simple statistics. [30] investigate several linguistic features of online Dutch toxic comments and non-toxic comments, focusing on the differences in average length, lexical diversity, and linguistic standardness of comments. Sharma et al. [31] analyzed female-related themes of item songs in Bollywood movies in a limited number, based on women's activist theory and cultural traditions. The frequent themes of the Top Item Songs include the glorification of criminal activities, dismemberment, materialism, high libidinal drive in women, the sexual objectification of women, etc. However, the biases of cognitive patterns in toxic language have yet to be studied. In this paper, we compare and analyze concepts generated from metaphors in different categories, which reflect the cognitive biases of people when they comment on various content.

Cognitive Analysis of Metaphors

As a bridge to connect target and source concepts, metaphors can help us deliver vivid expressions, understand complicated concepts, and enhance communication. Many researchers have worked on cognitive studies related to target and source concepts of metaphors. Hu and Wang [32] analyzed the target and source concepts in political documents from two countries to show the underlying cognitive patterns behind metaphors. Chen [33] analyzed the connotations and influence of the *greenhouse* metaphor on the climate system and concluded that a new conceptual metaphor for climate change needed to be presented, because the study found the *greenhouse* metaphor ineffective in raising people's positive attitude towards climate change. Wang et al. [34] explored the structural attributes of idioms with "ru" (meaning "similar to") in Chinese and summarized the selection restriction and metaphorical mappings between target and source concepts. Dodge [35] evaluated target and source concepts by pre-defined syntactic patterns from linguistic features and external knowledge bases. The used dataset includes text related to specific conceptual domains, e.g., GOVERNMENT, BUREAUCRACY, DEMOCRACY, POVERTY, TAXATION, AND WEALTH. Lachaud [36] employed electroencephalogram coherence to demonstrate brain activity of conceptual metaphors during comprehension. Fu et al. [37] constructed a non-directional reference graph to seek the most compatible target concepts, based on the source and contextual concepts detected from images. Li et al. [38] proposed a data-driven and unsupervised concept generation method, utilizing a web corpus with "like-a" and "is-a" syntactic patterns. This method has limited application to other syntactic patterns. Rosen [39] extracted features, based on dependency relationships between a target word and its context to present a source domain mapping model. However, the source domains were processed as one-hot vectors, which limits the output to one of the known source domains. Han et al. [40] integrated metaphorical concept mappings into an explainable neural network for depression detection, showing the common concept mappings that are likely to result in depression. Mao et al. [41] uncovered the cognitive patterns of financial analysts under different market environments from metaphors with MetaPro, showing the common concept mappings during bull and bear markets, respectively. MetaPro has been widely used in diverse cognitive analysis domains, including the cognitive analysis of political speech [15] and politicians [42], public perception towards

weather disasters [43], the analysis of the CEO's cognitive states [44], and the comparative analysis of the concept mapping patterns between humans and ChatGPT [45].

Previous works on metaphors have made significant contributions to understanding the cognitive connections between target and source concepts. However, several limitations remain. Many studies have focused on specific domains, such as political speech or climate change, limiting their generalizability to broader contexts of biased language and cognition. Furthermore, methodologies relying on predefined syntactic patterns or one-hot vector representations often constrain the flexibility and depth of metaphorical concept mappings, restricting the ability to uncover nuanced cognitive patterns. This paper seeks to examine cognitive biases present in a substantial volume of toxic language through the lens of metaphors, using a conceptual metaphor processing system.

Dataset Statistics

We use the dataset from the work of [46] to analyze the cognitive bias from toxic language. The dataset was collected for toxicity classification, sourced from the Civil Comments platform, a commenting plugin for independent news sites that was discontinued at the end of 2017. These public comments were created from 2015 to 2017. Each record contains seven toxicity-related labels (toxicity, severe_toxicity, obscene, sexual_explicit, identity_attack, insult, and threat). Part of the records are also annotated with 24 identity labels (male, female, Asian, Hindu, etc). The seven toxicity-related labels are values between 0 and 1, indicating the fraction of human annotators who believed the given comment belonged to the attribute. The last six labels are additional toxicity subtypes. The 24 identity labels are values between 0 and 1, indicating the fraction of human annotators who believed the given comment mentioned the identity group.

The dataset size is 1,999,515, of which 450,000 were annotated with identities indicating whether the comment mentions the corresponding identity. We cleaned the dataset by removing URLs, mentions, and hashtags in the text. MetaPro can process at most 512 tokens after Byte-Pair Encoding [47]. The very short text cannot provide sufficient context for accurately identifying metaphors. Thus, we selected the text with 30∼200 words. After MetaPro processed the dataset, we screened out 1,036,106 records with concept mapping outputs. We found some records where $toxicity = 0$ but additional toxicity subtypes $\neq 0$. After checking the background information of the annotation process, we regarded these records as ambiguous and removed them. For 10,616 records with the same cleaned text but different toxicity-related and identity labels, we calculated the averages as the values of their numerical labels, removed the duplicates, and retained one record for each clean text. The final size of the analyzing dataset is 1,015,290. The numeric toxicity-related labels can provide more refined information about toxicity. The 24 identity labels give various perspectives to analyze the text with metaphorical concept mappings. The large size of the dataset supports us in using statistical machine learning methods to find potential cognitive patterns in the text.

Figure 1 shows the distribution of 31 labels in this dataset. Considering each label has numerous records with 0 values, we display them with the y-axis on a logarithmic scale. Half of the labels show a similar trend of skewed to the right in the distribution, such as seven toxicity subtypes, bisexual, physical_disability, and intellectual_or_learning_disability. These categories have a concentration of data points toward the lower values on the scale, suggesting a prevalence of lower toxicity or identity mentioned in this dataset. Labels of female, homosexual_gay_or_lesbian, Muslim, Black, and White show left skewness. This suggests that a significant portion of the data for these labels is clustered at the higher end of the scale, indicating that this dataset contains large numbers of comments highly likely mentioning these groups.

Methods

MetaPro

We employed MetaPro [22], a metaphor processing tool to detect metaphors for each sentence and generate concept mappings for metaphorical ones. MetaPro contains metaphor identification [48], metaphor interpretation [49], and concept mapping generation [50] modules. The latest version of MetaPro is improved by a metaphor processing-tailored pretrained task, termed anomalous language modeling [51]. The metaphor identification detects metaphors on the token level. Next, the metaphor interpretation module paraphrases the metaphors into their literal counterparts. Finally, the concept mapping generation module abstracts the target and source concepts from the paraphrases and the original metaphors, respectively. The final output of concept mappings is formulated as "a target concept is a source concept." For example, given "my car drinks gasoline," MetaPro identifies *drinks* as a metaphor. Next, *drinks* is paraphrased as "consumes" in the context. Finally, UTILITY IS BODILY_PROCESS is generated, where UTILITY is abstracted from "consumes"; BODILY_PROCESS is abstracted from *drinks*.

Chi-square Test for Homogeneity

The Chi-square test for homogeneity is used to determine whether the frequency distribution of certain events observed in different groups is the same. In this paper, the test has the

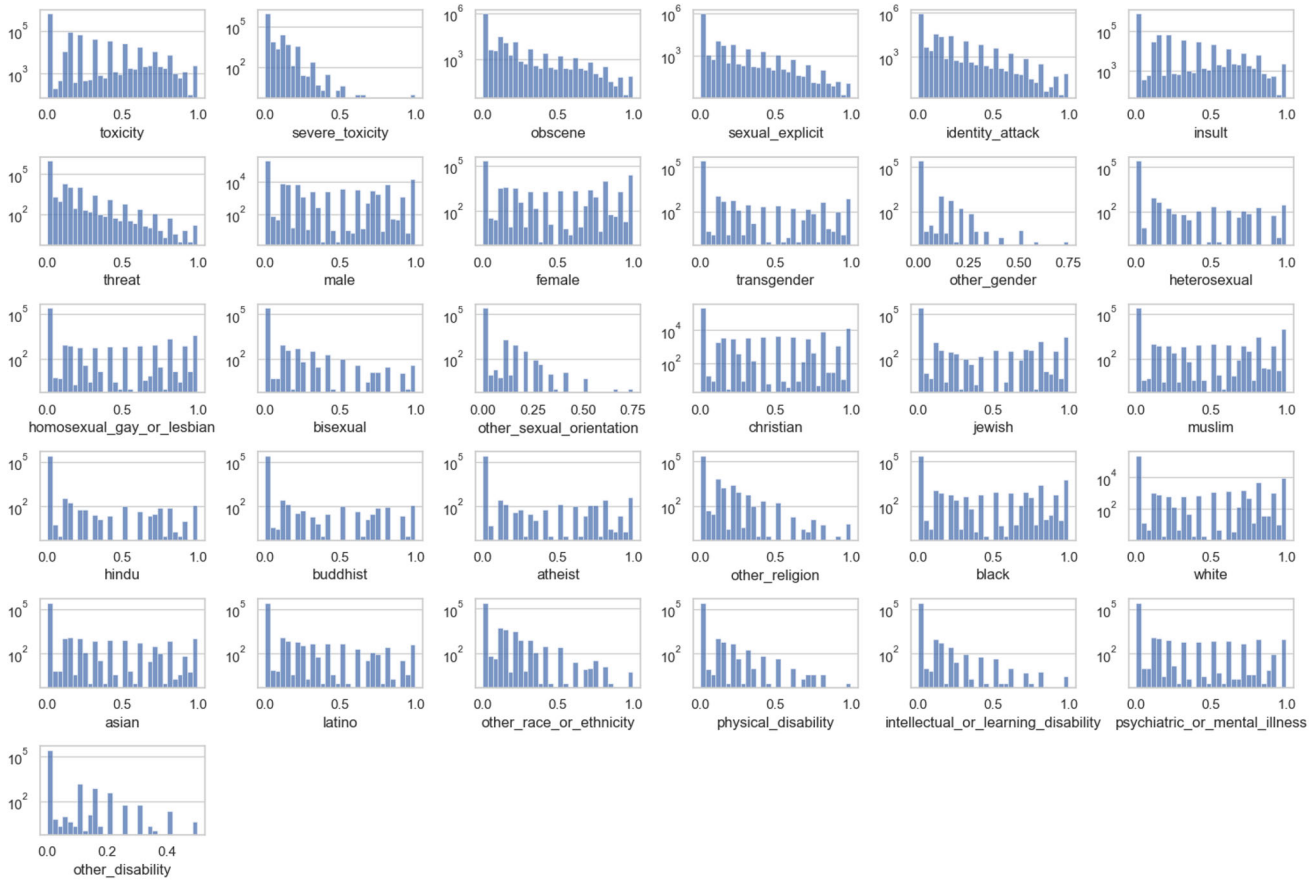


Fig. 1 Distribution plot on toxicity and identity labels. The y-axis is on a logarithmic scale

null hypotheses that the two groups have the same target concept, source concept, or concept mapping preferences.

The Chi-square statistic is calculated as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{1}$$

where O_i is an observed value, and E_i is an expected value. In our experiments, O_i and E_i represent the percentage of each concept or concept mapping in two categories (XX and non-XX, e.g., toxic and non-toxic). Then, we can obtain the p-value based on χ^2 and degrees of freedom (length of the corresponding categories). If the p-value is less than 0.05, we can reject the null hypotheses that the two groups have the same target concept, source concept, or concept mapping preferences.

The Chi-square test for homogeneity has several assumptions that need to be satisfied to ensure valid results. Each observation in the dataset is independent. The data for test is in the form of frequencies for categorical variables. All expected frequencies are 5 or more. The samples are drawn randomly from the populations to avoid bias. The categories within the variable are mutually exclusive, meaning each observation falls into only one category. The numbers of comparing variables and categories are fixed. Our dataset satisfies all the assumptions of the Chi-square test for homogeneity.

Concept Frequency Comparison

The frequency of target concepts, source concepts, and concept mappings gives us a general perspective of each category. We calculated the frequency of target concepts, source concepts, and concept mappings in two categories (XX and non-XX, e.g., toxic and non-toxic). We concentrate on the concepts and mappings with the most different frequencies between the two categories. Therefore, for concepts and mappings that exist in both categories, we selected concepts or mappings with the largest and smallest frequency ratios in the two categories for subsequent data analysis. For concepts and mappings unique in one category, we selected those with the largest frequency in the corresponding category for subsequent data analysis. We remove concepts that occur in both categories ≤ 2 times to focus more on frequent concepts. It is worth noting that we only select the concepts and mappings with the most significant differences for comparison and analysis.

Association Rule Mining

ARM, first proposed by [52], is an unsupervised learning technique utilizing a rule-based approach to discover interesting relationships between valuable features from a large dataset. It quantitatively describes the influence of object A occurrence on object B occurrence. Our study aims to use ARM to reflect the dependency and connection among concept mappings under different categories. ARM provides us with creative hypotheses and intuitive deduction of the cognitive patterns at a more micro level in the subsequent sections.

Some related terms need to be introduced before our analysis. Let T_i be a set consisting of all the concept mappings in one record, where $i = 1, 2, ..., n$. n is the total number of the records in the dataset. An association rule is in the form of $X \Rightarrow Y$, which means X implies Y. X, Y are sets of concept mappings, and $X \cap Y = \varnothing$. X is called antecedent, while Y is called consequent.

Support of an association rule is the probability of X and Y simultaneously shown in the concept mapping set of one record. $Support$ shows the frequency of an association rule in the dataset. We calculate $Support$ in the following formula:

$$Support = P(XY) = \frac{\sum_{i=1}^{n} I\{(X \cup Y) \subseteq T_i\}}{n}, \tag{2}$$

where I is an indicator equaling 1 when the corresponding condition is satisfied, or else it equals 0.

Confidence of an association rule is defined as the probability of Y given X. It reflects the strength of an association rule. We calculate $Confidence$ in the following formula:

$$Confidence = P(Y|X) = \frac{\sum_{i=1}^{n} I\{(X \cup Y) \subseteq T_i\}}{\sum_{i=1}^{n} I\{X \subseteq T_i\}}. \tag{3}$$

Lift of an association rule is defined as the fraction of the probability of Y given X to the probability of Y. It reflects the enhanced influence of the association rule on the consequent. We calculate $Lift$ in the following formula:

$$Lift = \frac{P(Y|X)}{P(Y)} = \frac{\sum_{i=1}^{n} I\{(X \cup Y) \subseteq T_i\}}{\sum_{i=1}^{n} I\{X \subseteq T_i\} \sum_{i=1}^{n} I\{Y \subseteq T_i\}}. \tag{4}$$

Cognitive Biases Between Toxic and Non-Toxic Language

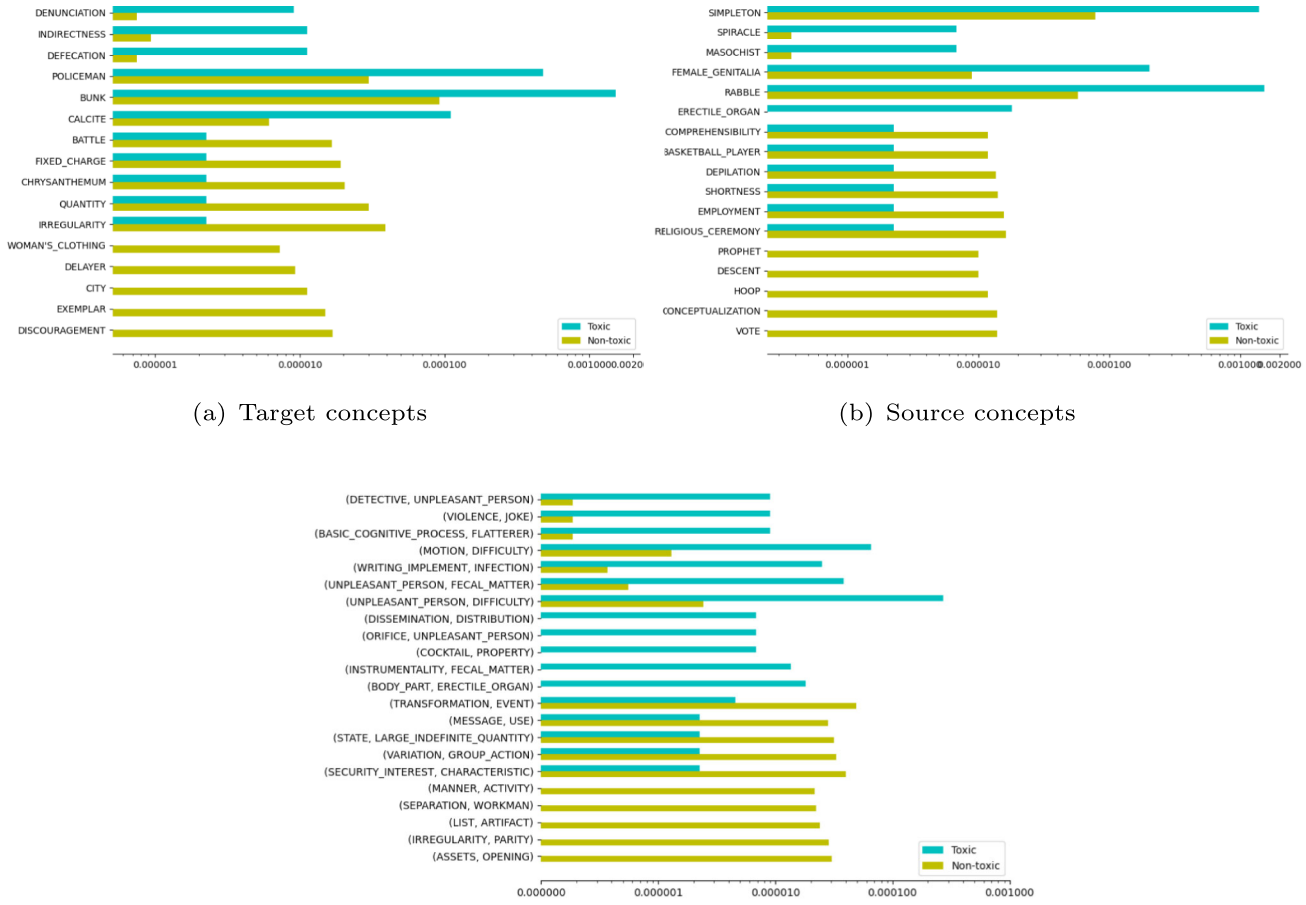
Following the instructions of [53], we consider records with $toxicity \geq 0.5$ as positive (toxic), the rest of which is negative (non-toxic). Basic statistics are shown in Table 1. The definitions of low-toxic and high-toxic categories will be introduced in "Cognitive Biases Between Level of Toxicity" section.

We run Chi-square tests for homogeneity on the target concepts, source concepts, and concept mappings in toxic and non-toxic language. The three p-values are all less than 0.05. Thus, we reject the null hypotheses that the toxic and non-toxic languages have the same target concept, source concept, or concept mapping preferences. The findings indicate cognitive distinctions among speakers who use varying types of toxic language.

Figure 2 shows the frequency comparison of target concepts, source concepts, and concept mappings between toxic and non-toxic language. We select concepts or concept mappings that show significant differences between two categories. When multiple concepts or mappings exhibit the same difference, all are included in the figures, resulting in varying numbers of concepts or mappings across the frequency comparison figures. To demonstrate the comparison, we use a logarithmic scale on the x-axis. Thus, the actual differences between bars are much more significant than they look. The target concepts much more frequent in toxic language are negative or related to contentious topics (POLICEMAN; DEFECATION; DENUNCIATION). Those in non-toxic language are more neutral in sentiment (EXEMPLAR), relating to everyday objects (CITY; WOMANS_CLOTHING; FIXED_CHARGE). The source concepts more frequent in toxic language are provocative (SIMPLETON; RABBLE) to demean or convey explicit sexual connotations (MASOCHIST; FEMALE_GENITALIA; ERECTILE_ORGAN). The concept mappings much more frequent in toxic language are more offensive with belittling or mocking tendencies, such as (BODY_PART, ERECTILE_ORGAN); (UNPLEASANT_PERSON, DIFFICULTY); (UNPLEASANT_PERSON, FECAL_MATTER); (BASIC _COGNITIVE_PROCESS, FLATTERER). Those in non-toxic language are more descriptive, aiming to explain or enhance understanding, such as (IRREGULARITY, PARITY); (STATE, LARGE_INDEFINITE_QUANTITY); (TRANSFORMATION, EVENT). For example, "I included this with my testimony to city council on /3/17 when they were dis-

Table 1 Statistics of basic toxicity categories

Categories	Toxicity range	No. of records	% in dataset	% in toxic language
Non-toxic	(0, 0.5)	933833	91.97697%	–
Toxic	[0.5, 1]	81457	8.02303%	–
Low-toxic	[0.5, 0.75)	64972	6.39935%	79.76233%
High-toxic	[0.75, 1]	16485	1.62367%	20.23767%



(a) Target concepts

(b) Source concepts

(c) Concept mappings

Fig. 2 Frequency comparison between toxic and non-toxic language. The x-axis is on a logarithmic scale

cussing the *recovery* [improvement] plan." The generated concept mapping is (TRANSFORMATION, EVENT). The source concept TRANSFORMATION and paraphrased word [improvement] describe the change of the conditions, while *recovery* emphasizes the plan is an essential EVENT.

Tables 2 and 3 show the top 10 association rules generated from toxic and non-toxic language, respectively. (PART, LEFTFIELDER) ⇒ (FORMATION, SOCIAL_GROUP) and (PARITY, LEFTFIELDER) ⇒ (FORMATION, SOCIAL_GROUP) frequently exist in toxic language. They suggest a connec-

Table 2 Top 10 association rules of concept mappings in toxic language ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(IMPROVEMENT, ABSTRACTION)	(FORMATION, SOCIAL_GROUP)	0.00266	0.71854	47.24000
(PART, LEFTFIELDER)	**(FORMATION, SOCIAL_GROUP)**	**0.00204**	**0.50303**	**33.07130**
(SENSING, APPEARANCE)	(POSSESSION, ACTION)	0.00160	0.66327	35.49777
(PARITY, LEFTFIELDER)	**(FORMATION, SOCIAL_GROUP)**	**0.00114**	**0.55030**	**36.17873**
(BLOOD_SPORT, ACTIVITY)	(NONRELIGIOUS_PERSON, OCCULTIST)	0.00101	0.87234	507.55881
(NONRELIGIOUS_PERSON, OCCULTIST)	(BLOOD_SPORT, ACTIVITY)	0.00101	0.58571	507.55881
(DOCUMENT, ACCOMPLISHMENT)	**(PRAISE, PEOPLE)**	**0.00075**	**0.63542**	**136.92893**
(EXTREMITY, LINE)	**(BASIC COGNITIVE PROCESS, HIGHER COGNITIVE PROCESS)**	**0.00053**	**0.62319**	**159.13184**
(LARGE_INDEFINITE_QUANTITY, TRANSACTION)	(SIZE, IMPORTANCE)	0.00050	0.71930	28.87722
(MALE, MALE_OFFSPRING)	**(UNPLEASANT_PERSON, DIFFICULTY)**	**0.00050**	**0.51250**	**363.01489**

We highlight the key rules of our analysis in bold. Conf. is short for Confidence

Table 3 Top 10 association rules of concept mappings in non-toxic language ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(IMPROVEMENT, ABSTRACTION)	(FORMATION, SOCIAL_GROUP)	0.00223	0.70375	90.98554
(SENSING, APPEARANCE)	(POSSESSION, ACTION)	0.00189	0.62057	30.74636
(LARGE_INDEFINITE_QUANTITY, TRANSACTION)	(SIZE, IMPORTANCE)	0.00106	0.77769	26.46333
(BLOOD_SPORT, ACTIVITY)	(NONRELIGIOUS_PERSON, OCCULTIST)	0.00080	0.86012	901.46390
(NONRELIGIOUS_PERSON, OCCULTIST)	(BLOOD_SPORT, ACTIVITY)	0.00080	0.83502	901.46390
(SYSTEM, COMPUTER_NETWORK)	**(STRONGBOX, DEVICE)**	**0.00049**	**0.74637**	**1113.38716**
(STRONGBOX, DEVICE)	**(SYSTEM, COMPUTER_NETWORK)**	**0.00049**	**0.73802**	**1113.38716**
(ACT, ATTRIBUTE)	(MAGNITUDE_RELATION, RELATION)	0.00044	0.50433	474.75461
(ACTIVITY, MUSICAL_PERFORMANCE)	(REGION, GEOGRAPHICAL_AREA)	0.00043	0.75233	337.43923
(CHROMATIC_COLOR, WOOD)	(DECISION_MAKING, ACTION)	0.00042	0.60815	87.66764

tion between individual roles (LEFTFIELDER) and broader social context (SOCIAL_GROUP), reflecting a cognitive pattern where toxic language tends to link personal roles with social structures. (EXTREMITY, LINE) $\Rightarrow$ (BASIC_COGNITIVE_PROCESS, HIGHER_COGNITIVE_PROCESS) indicates a toxic mindset where extremity or linearity narrows basic and higher cognitive processes. It suggests an inflexible approach to cognitive activities, potentially hindering creative thinking or open-mindedness within the toxic language context. The (MALE, MALE_OFFSPRING) $\Rightarrow$ (UNPLEASANT_PERSON, DIFFICULTY) rule might imply a negative impression where males and their male offspring are associated with being unpleasant individuals and causing difficulties in toxic language. This interpretation also suggests sexist bias and negative perception of males and their male offspring in the toxic context.

(SYSTEM, COMPUTER_NETWORK) and (STRONGBOX, DEVICE) are mutually enhanced in non-toxic language. Their $Lift$ reaches up to 1113.39, indicating that the two concept mappings are strongly associated. The occurrence of one concept mapping provides substantial information to boost the likelihood of the other concept mapping occurring. This association may imply a non-toxic context where systems, computer networks, strongboxes, and devices are involved in technological security measures to ensure data without toxic contaminants.

In summary, the critical differences between toxic and non-toxic language in using metaphors lie in their intended focuses and effects. Toxic language possibly tends to harm with explicitly hostile or insulting metaphors. In contrast, non-toxic language aims to inform, explain, or enhance understanding with neutral or constructive metaphors. Some previous natural language processing research [54, 55] also demonstrated the correlation between toxicity and negative sentiment by multiple experiments. Zhang et al. [56] prevented large language models from generating harmful information by intention analysis and chain-of-thoughts, which supports our findings that intentions of toxic and non-toxic language are distinct.

Cognitive Biases Between Level of Toxicity

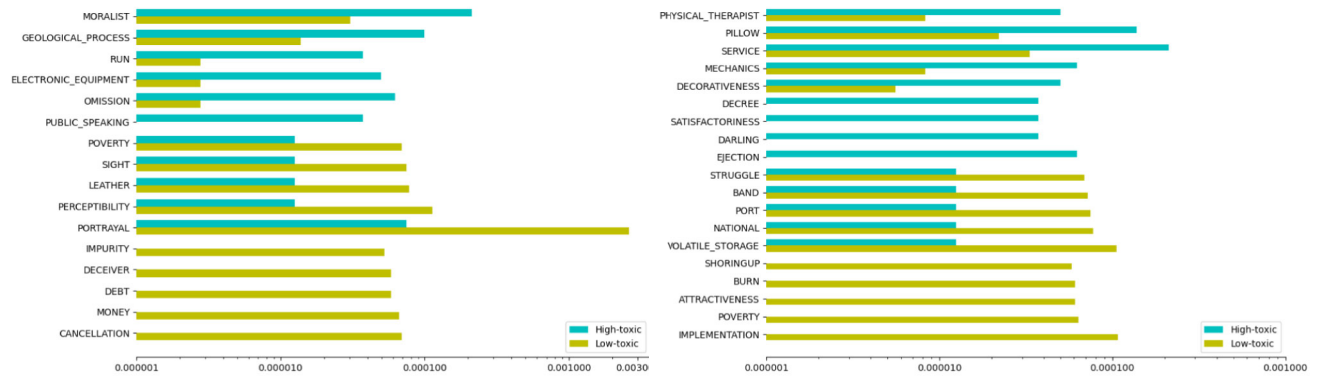
To refine the cognitive details of toxic language, we separate toxic language into low-toxic and high-toxic categories, defined by toxicity $\in [0.5, 0.75)$ and toxicity $\in [0.75, 1]$. The basic statistics of the two categories are shown in Table 1.

We run Chi-square tests for homogeneity on the target concepts, source concepts, and concept mappings in low-toxic and high-toxic language. The three p-values are all less than 0.05. Thus, we reject the null hypotheses that the low-toxic and high-toxic languages have the same target concept, source concept, or concept mapping preferences. The findings indicate cognitive distinctions between different levels of toxic language.

Figure 3 shows the frequency comparison of target and source concepts between low-toxic and high-toxic language. The target and source concepts in low-toxic language reflect everyday concerns and experiences, focusing on socio-economic issues (MONEY; DEBT; POVERTY) and physical perceptions (PORTRAYAL; PERCEPTIBILITY; SIGHT). In high-toxic language, the metaphors related to strong sentiments (DARLING; SATISFACTORINESS)[3] or technical and professional contexts (EJECTION; MECHANICS; PHYSICAL THERAPIST; ELECTRONIC_EQUIPMENT; GEOLOGICAL_PROCESS). They may trigger strong emotional responses or be used to attack someone's capabilities or actions in a particular area.

Tables 4 and 5 show the top 10 association rules of concept mappings generated from low-toxic and high-toxic language, respectively. It is reasonable to observe that the rules in Table 4 are similar to those in Table 2 since low-toxic

[3] For example, let us just *coddle* [treat] these idiotic standoff perps and let them have their way. The source concept for *coddle* is DARLING.



(a) Target concepts (b) Source concepts

Fig. 3 Frequency comparison between low-toxic and high-toxic language. The x-axis is on a logarithmic scale

language accounts for 80% of toxic language. (COSMETIC, MAKEUP) ⇒ (LOCATION, UNPLEASANT_PERSON) suggests a judgmental perspective where using cosmetics or makeup is unfairly linked to the presence of unpleasant people involved in inappropriate activities in dubious locations.

(MALE, MALE_OFFSPRING) ⇒ (UNPLEASANT_PERSON, DIFFICULTY) and (EXTREMITY, LINE) ⇒ (BASIC_COGNITIVE_PROCESS, HIGHER_COGNITIVE_PROCESS) are top rules in toxic and high-toxic language, but not in low-toxic language. These two rules are relatively intuitive for negative explanations. Besides, more concept mappings with negative sentiment tendencies are shown in Table 5, indicating that high-toxic language contains more negative cognition patterns.

(UNPLEASANT_PERSON, SEED) maps the impact of an unpleasant person to the growth of a seed. (UNPLEASANT_PERSON, SEED) ⇒ (ACTIVITY, WORK) implies that the presence of an unpleasant person might affect the productivity of work-related activities. The rule (ASSETS, POUCH) ⇒ (ENOUGH, BOUNDARY) possibly suggests a misguided approach to resource exploitation, connoting a toxic intent of crossing moral or ethical boundaries in managing financial assets.

To sum up, high-toxic language often focuses on contentious or societal critique themes, while low-toxic language revolves around more daily, socio-economic, or abstract concepts. Metaphors in high-toxic language serve to express strong sentiments, often damaging or confrontational, while those in low-toxic language aim to engage in relatively rational or constructive discourse.

Cognitive Biases Among Subtypes of Toxicity

We extract records with $subtypes \geq 0.5$ and $toxicity \geq 0.5$ as positive records for corresponding subtypes of toxicity to study their cognitive biases. Table 6 shows the size of each subtype and their percentage in toxic language. The subtypes are not mutually exclusive. One record of toxic language may belong to multiple or zero subtypes. Therefore, the sum of the percentage is not 100%. Here, we show the analysis of the two most characteristic subtypes, the obscene and sexual_explicit subtypes.

Table 4 Top 10 association rules of concept mappings in low-toxic language ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(IMPROVEMENT, ABSTRACTION)	(FORMATION, SOCIAL_GROUP)	0.00291	0.71863	49.09664
(PART, LEFTFIELDER)	(FORMATION, SOCIAL_GROUP)	0.00205	0.50763	34.68136
(SENSING, APPEARANCE)	(POSSESSION, ACTION)	0.00166	0.66258	34.30194
(BLOOD_SPORT, ACTIVITY)	(NONRELIGIOUS_PERSON, OCCULTIST)	0.00114	0.86047	462.03421
(NONRELIGIOUS_PERSON, OCCULTIST)	(BLOOD_SPORT, ACTIVITY)	0.00114	0.61157	462.03421
(PARITY, LEFTFIELDER)	(FORMATION, SOCIAL_GROUP)	0.00111	0.53333	36.43715
(DOCUMENT, ACCOMPLISHMENT)	(PRAISE, PEOPLE)	0.00072	0.61039	134.43469
(LARGE_INDEFINITE_QUANTITY, TRANSACTION)	(SIZE, IMPORTANCE)	0.00057	0.72549	29.00711
(COSMETIC, MAKEUP)	**(LOCATION, UNPLEASANT_PERSON)**	**0.00049**	**0.76190**	**213.37274**
(DEVICE, LIST)	(UTILITY, ACTION)	0.00048	0.67391	67.77938

Table 5 Top 10 association rules of concept mappings in high-toxic language ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(IMPROVEMENT, ABSTRACTION)	(FORMATION, SOCIAL_GROUP)	0.00170	0.71795	41.09509
(SENSING, APPEARANCE)	(POSSESSION, ACTION)	0.00133	0.66667	41.16105
(PARITY, LEFTFIELDER)	(FORMATION, SOCIAL_GROUP)	0.00127	0.61765	35.35386
(UNPLEASANT_PERSON, SEED)	**(ACTIVITY, WORK)**	**0.00103**	**0.50000**	**70.44872**
(DOCUMENT, ACCOMPLISHMENT)	(PRAISE, PEOPLE)	0.00085	0.73684	146.34750
(EXTREMITY, LINE)	**(BASIC COGNITIVE PROCESS, HIGHER COGNITIVE PROCESS)**	**0.00079**	**0.72222**	**212.60417**
(COMMUNICATION, PERSON)	(OBJECT, SOIL)	0.00073	0.57143	376.80000
(PURPOSE, INDICATION)	(DEFINITE_QUANTITY, DIGIT)	0.00067	0.61111	279.83796
(MALE, MALE_OFFSPRING)	**(UNPLEASANT_PERSON, DIFFICULTY)**	**0.00067**	**0.57895**	**212.08772**
(ASSETS, POUCH)	**(ENOUGH, BOUNDARY)**	**0.00067**	**0.50000**	**329.70000**

The Obscene Subtype in Toxic Language

For Chi-square tests and concept frequency comparison in subtypes of toxicity, we compare one subtype with the complement set of that subtype with respect to the whole toxic language. We run Chi-square tests for homogeneity on the target concepts, source concepts, and concept mappings in obscene and other toxic language. The three p-values are all less than 0.05. Thus, we reject the null hypotheses that the obscene and other toxic languages have the same target concept, source concept, or concept mapping preferences. The findings indicate cognitive distinctions between the obscene and other toxic languages.

Figure 4 shows the frequency comparison of source concepts and concept mappings between obscene and other toxic language. Compared with other toxic language, the source concepts in obscene language are more direct, provocative, and carry stronger emotive connotations (DIFFICULTY; FEMALE_GENITALIA; DENUNCIATION). Concept mappings in obscene language, e.g., (COMMUNICATION, DIFFICULTY); (WORTHLESSNESS, FECAL_MATTER) reflect a more confrontational and explicit style of expression. The choice of metaphors reflects the norms and dynamics of the context. Metaphors in obscene language are typically used in more informal and aggressive contexts to express strong emotions, while metaphors in other toxic language might be effective in vividly illustrating points and provoking thought.

Table 6 Statistics of toxicity subtypes

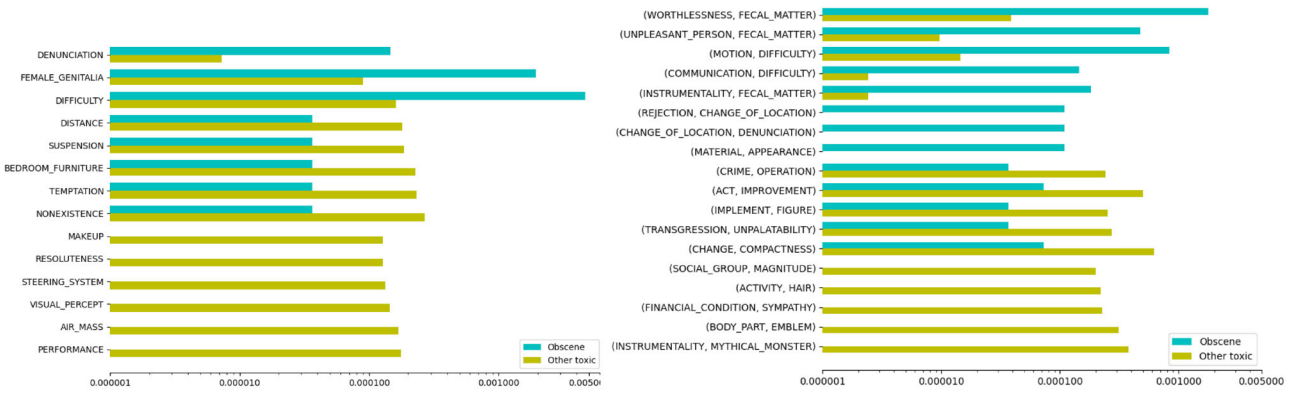
Subtypes of toxicity	No. of records	% in toxic language
Severe_toxicity	7	0.00859%
Obscene	5107	6.26957%
Sexual_explicit	2038	2.50193%
Identity_attack	6953	8.53579%
Insult	59247	72.73408%
Threat	1083	1.32954%

Table 7 shows the top 10 association rules of obscene subtype in toxic language. Different from the above-mentioned categories, rules involving three concept mappings appear frequently in subtypes of toxic language. The $Confidence$ of some rules reaches up to 1.0, indicating that the consequent concept mapping always occurs given the presence of the antecedent one. These phenomena imply a stronger association among concept mappings when we narrow the analyzing domain. (COCKTAIL, ACTION) is a frequent antecedent, which may imply a scenario where the consumption of cocktails leads to harmful or aggressive actions. One consequent (PART, BODY_PART) suggests specific body parts involved during these actions. Another consequent (DIRECTION, POSITION) might imply a chaotic correlation intended to create confusion about direction or position in an obscene environment. (FOLLOWER, DEVICE) $\Rightarrow$ (ACTIVITY, FIGHT) implies an obscene event where individuals such as FOLLOWER use a DEVICE to engage in aggressive ACTIVITY, such as online or physical fights. This rule suggests a negative influence, where vulnerable individuals might be pressured into malicious events.

The Sexual_explicit Subtype in Toxic Language

We run Chi-square tests for homogeneity on the target concepts, source concepts, and concept mappings in sexual_explicit and other toxic language. The three p-values are all less than 0.05. Thus, we reject the null hypotheses that sexual_explicit and other toxic language have the same target concept, source concept, or concept mapping preferences. The findings indicate cognitive distinctions between sexual_explicit and other toxic language.

Figure 5 shows the frequency comparison of target concepts, source concepts, and concept mappings between sexual_explicit and other toxic language. In sexual_explicit language, the target concepts (ERECTILE_ORGAN; INFLAMMA-



(a) Source concepts (b) Concept mappings

Fig. 4 Frequency comparison between obscene and other toxic language. The x-axis is on a logarithmic scale

TORY_DISEASE), source concepts (GENDER; ACNE), and concept mappings ((BODY_PART, ERECTILE_ORGAN); (INFLAMMATORY_DISEASE, ACNE); (UNPLEASANT_PERSON, FEMALE_ GENITALIA)) focus on more direct and crude physical or sexual comparisons. In contrast, other toxic language employs metaphors to emphasize, critique, or negatively portray various abstract concepts.

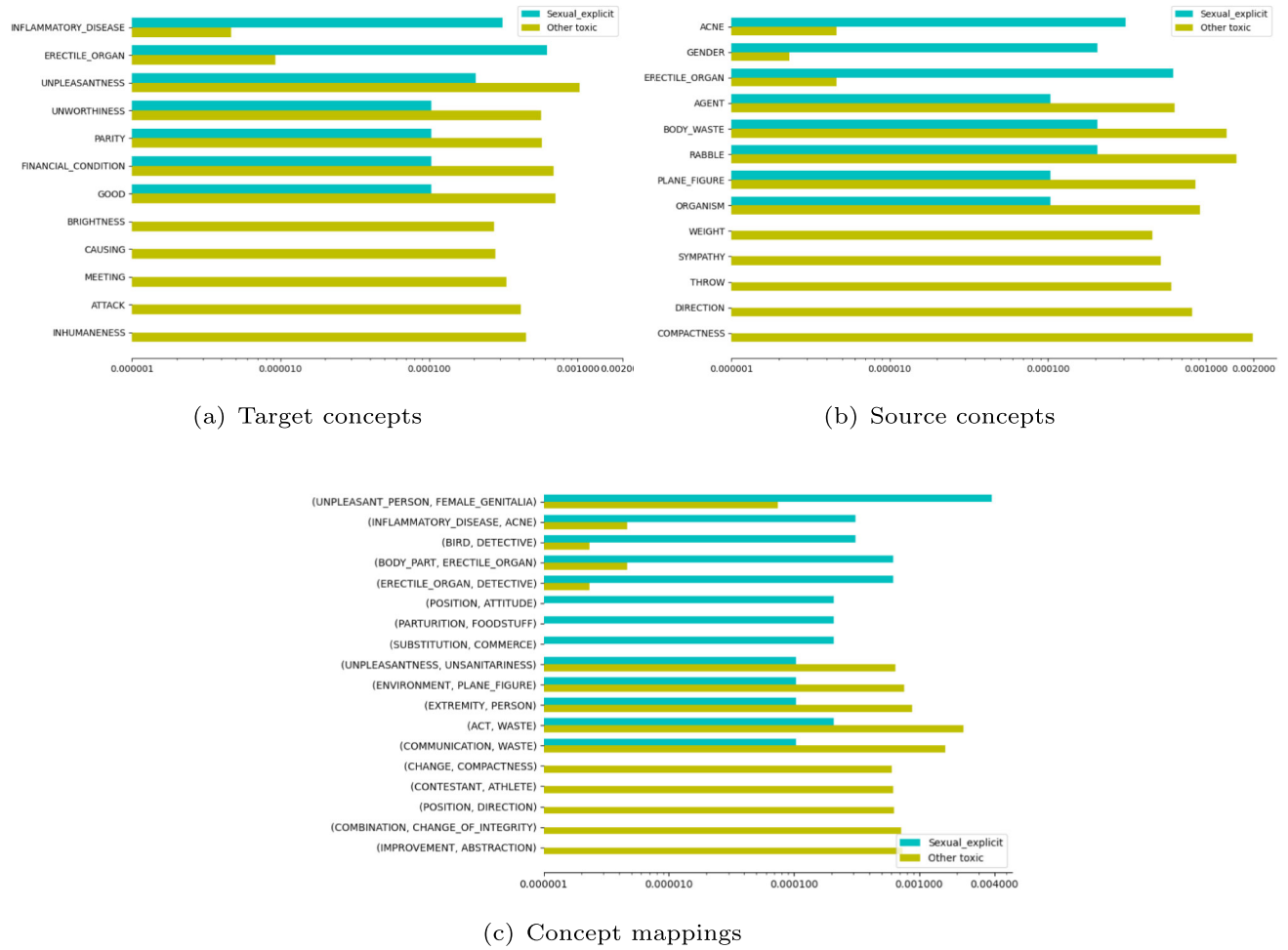
Table 8 displays the top 10 association rules of concept mappings generated from sexual_explicit in toxic language. One frequent association rule is (ORGAN, REPRODUCTIVE_ORGAN) $\Rightarrow$ (BODILY_PROCESS, CONSUMPTION). This rule implies a toxic scenario where organs mapped to reproductive organs are linked with mere bodily processes and consumption. It suggests an objectifying mindset stripping humanity of individuals, reducing individuals to physiological functions, and reinforcing harmful stereotypes.

Similar to the obscene subtype, (COCKTAIL, ACTION) is a frequent antecedent of (PART, BODY_PART) because obscene language and sexual_explicit language might have some cognitive patterns in common. Additionally, the concept mapping (PART, BODY_PART) is also evoked frequently by (COGNITION, PERSON), (BURNING, ACTION), and (QUALITY, ACKNOWLEDGMENT). The associations might involve sexualization, where body parts are used inappropriately for demeaning purposes, contributing to a toxic atmosphere. It can also be interpreted as an emphasis on physical appearance and associated judgments, such as body shaming. Body parts are used to criticize and ridicule individuals based on their actions and cognition. These rules suggest a toxic culture of body-related criticism and negativity.

The strong association among (PHYSICS, AUTOMATIC_FIREARM), (EXPERT, CELESTIAL_BODY), and (FORCE, DEVICE)

Table 7 Top 10 association rules of concept mappings generated from obscene in toxic language ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(MALE, MALE_OFFSPRING)	(UNPLEASANT_PERSON, DIFFICULTY)	0.00509	0.76471	45.94533
(COCKTAIL, ACTION)	**(PART, BODY_PART)**	**0.00196**	**0.62500**	**14.77720**
(IMPROVEMENT, ABSTRACTION)	(FORMATION, SOCIAL_GROUP)	0.00157	0.88889	122.69069
(COCKTAIL, ACTION)	**(DIRECTION, POSITION)**	**0.00157**	**0.50000**	**8.89721**
(SENSING, APPEARANCE)	(POSSESSION, ACTION)	0.00137	0.63636	33.16234
(FORCE, DEVICE)	(EXPERT, CELESTIAL_BODY)	0.00117	1.00000	464.27273
(FOLLOWER, DEVICE)	**(ACTIVITY, FIGHT)**	**0.00117**	**1.00000**	**268.78947**
(EXPERT, CELESTIAL_BODY)	(FORCE, DEVICE)	0.00117	0.54545	464.27273
(TERRESTRIAL_PLANET, ARTIFACT)	(HIGH_STATUS, DEGREE)	0.00117	0.54545	32.39112
(PHYSICS, AUTOMATIC_FIREARM) (EXPERT, CELESTIAL_BODY)	(FORCE, DEVICE)	0.00098	1.00000	851.16667



(a) Target concepts

(b) Source concepts

(c) Concept mappings

Fig. 5 Frequency comparison between sexual_explicit and other toxic language. The x-axis is on a logarithmic scale

might indicate a power relationship within a sexually explicit context. The expertise (EXPERT) and technological elements (PHYSICS, DEVICE) could represent sources of power, while FORCE and AUTOMATIC_FIREARM represent aggression and control. These interpretations suggest that wielding power aggressively potentially leads to harmful consequences.

Cognitive Biases Among Toxic Language Mentioning Genders

Similar to subtype labels, we extract records with $identity\ groups \geq 0.5$ and $toxicity \geq 0.5$ as positive records for corresponding identity groups. The basic statistics for labels

Table 8 Top 10 association rules of concept mappings generated from sexual_explicit in toxic language ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(ORGAN, REPRODUCTIVE_ORGAN)	**(BODILY_PROCESS, CONSUMPTION)**	**0.00442**	**0.69231**	**37.12955**
(COCKTAIL, ACTION)	(PART, BODY_PART)	0.00442	0.56250	6.94773
(COGNITION, PERSON)	**(PART, BODY_PART)**	**0.00343**	**0.70000**	**8.64606**
(BURNING, ACTION)	**(PART, BODY_PART)**	**0.00343**	**0.50000**	**6.17576**
(FORCE, DEVICE)	(EXPERT, CELESTIAL_BODY)	0.00294	0.75000	152.85000
(QUALITY, ACKNOWLEDGMENT)	**(PART, BODY_PART)**	**0.00294**	**0.75000**	**9.26364**
(EXPERT, CELESTIAL_BODY)	(FORCE, DEVICE)	0.00294	0.60000	152.85000
(PHYSICS, AUTOMATIC_FIREARM), (EXPERT, CELESTIAL_BODY)	**(FORCE, DEVICE)**	**0.00245**	**1.00000**	**254.75000**
(PHYSICS, AUTOMATIC_FIREARM), (FORCE, DEVICE)	**(EXPERT, CELESTIAL_BODY)**	**0.00245**	**1.00000**	**203.80000**
(FORCE, DEVICE), (EXPERT, CELESTIAL_BODY)	**(PHYSICS, AUTOMATIC_FIREARM)**	**0.00245**	**0.83333**	**212.29167**

mentioning gender, sexual orientation, and race are shown in Table 9. Our research emphasis is highlighted in bold.

Toxic Language Mentioning Males and Females

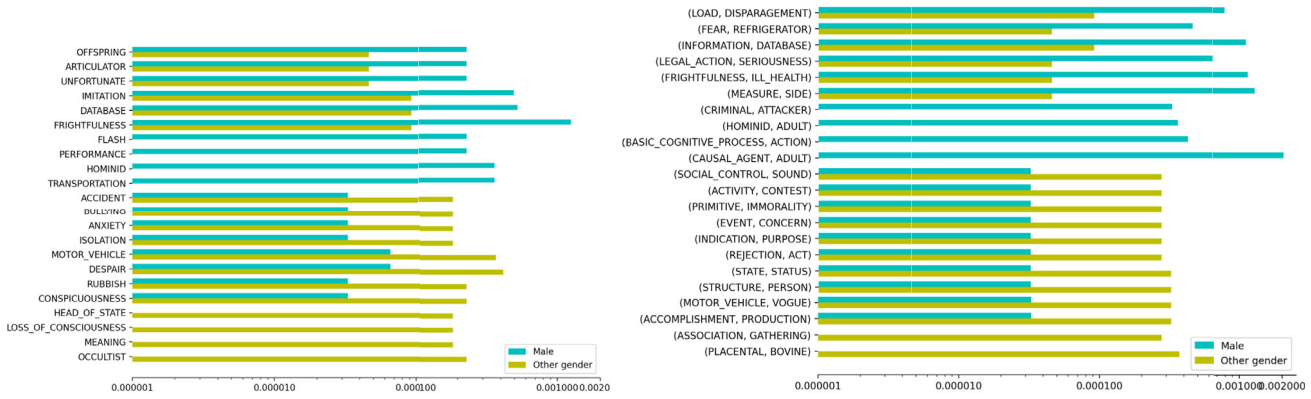
We run Chi-square tests for homogeneity on the target concepts, source concepts, and concept mappings in toxic language mentioning males/females and other genders. Here, "other gender" (without a dash) has different meanings from the label "other_gender" (with a dash) in Table 9. In Table 9, "other_gender" means other genders excluding males, females, and transgender, while "other gender" in the former context means non-XX genders, including all the other gender-related labels in the original dataset. We also use similar expressions in the following sexual orientation and race groups. The three p-values are all less than 0.05. Thus, we reject the null hypotheses that the toxic language mentioning males/females and other genders have the same target concept, source concept, or concept mapping preferences. The findings indicate cognitive distinctions between toxic language mentioning males/females and other genders.

Figure 6 shows the frequency comparison of target concepts and concept mappings between toxic language mentioning males and other genders. The target concepts in toxic language mentioning males cover physical (PERFORMANCE; FLASH; DATABASE) and biological (HOMINID; ARTICULATOR; OFFSPRING) domains, while those in toxic language mentioning other genders focus on psychological states (OCCULTIST; LOSS_OF_CONSCIOUSNESS; DESPAIR; ISOLATION; ANXIETY). Concept mappings in toxic language mentioning males ((CRIMINAL, ATTACKER); (FRIGHTFULNESS, ILL_HEALTH); (LEGAL_ACTION, SERIOUSNESS); (FEAR, REFRIGERATOR); (LOAD, DISPARAGEMENT)) may link males to roles, actions, or states that involve power and control. Concept mappings in toxic language mentioning other genders ((PLACENTAL, BOVINE); (MOTOR_VEHICLE, VOGUE); (STRUCTURE, PERSON); (PRIMITIVE, IMMORALITY); (SOCIAL_ CONTROL, SOUND)) indicate a focus on societal roles, appearances, and behaviors with connotations of judgment or stereotyping. Metaphors in toxic language mentioning males may aim to challenge notions of strength and control. In contrast, those mentioning other genders concentrate more on societal stereotypes or emotional judgments, potentially reinforcing traditional biases.

Figure 7 shows the frequency comparison of target concepts, source concepts, and concept mappings between toxic language mentioning females and other genders. Some target concepts in toxic language mentioning females cover professional roles (EXPERT; INTELLECTUAL; PHYSICIST) and emotional states (OFFENSIVENESS; DESPAIR). The source concepts in toxic language mentioning females contain more negative sentiment (DISLIKE; DISCOURTESY) and sexual relation (SEXUAL_DESIRE; WILD). The concept mappings in toxic language mentioning females, such as (CHANGE, CREATING_BY_REMOVAL); (CHANGE_OF_STATE, ADJUSTMENT); (WAR, ACTIVITY); (OFFENSIVENESS, AVERAGE), tend to link females to abstract concepts. The toxic language mentioning females focuses on expressing negative sentiments and undermining the competence of women. Metaphors in toxic language mentioning females are more abstract or subtly derogatory, potentially making the toxic expression require more interpretation but still harmful. In contrast, metaphors mentioning other genders tend to use more concrete metaphors, with direct belittlement or stereotype.

Table 9 Subset statistics of identity groups

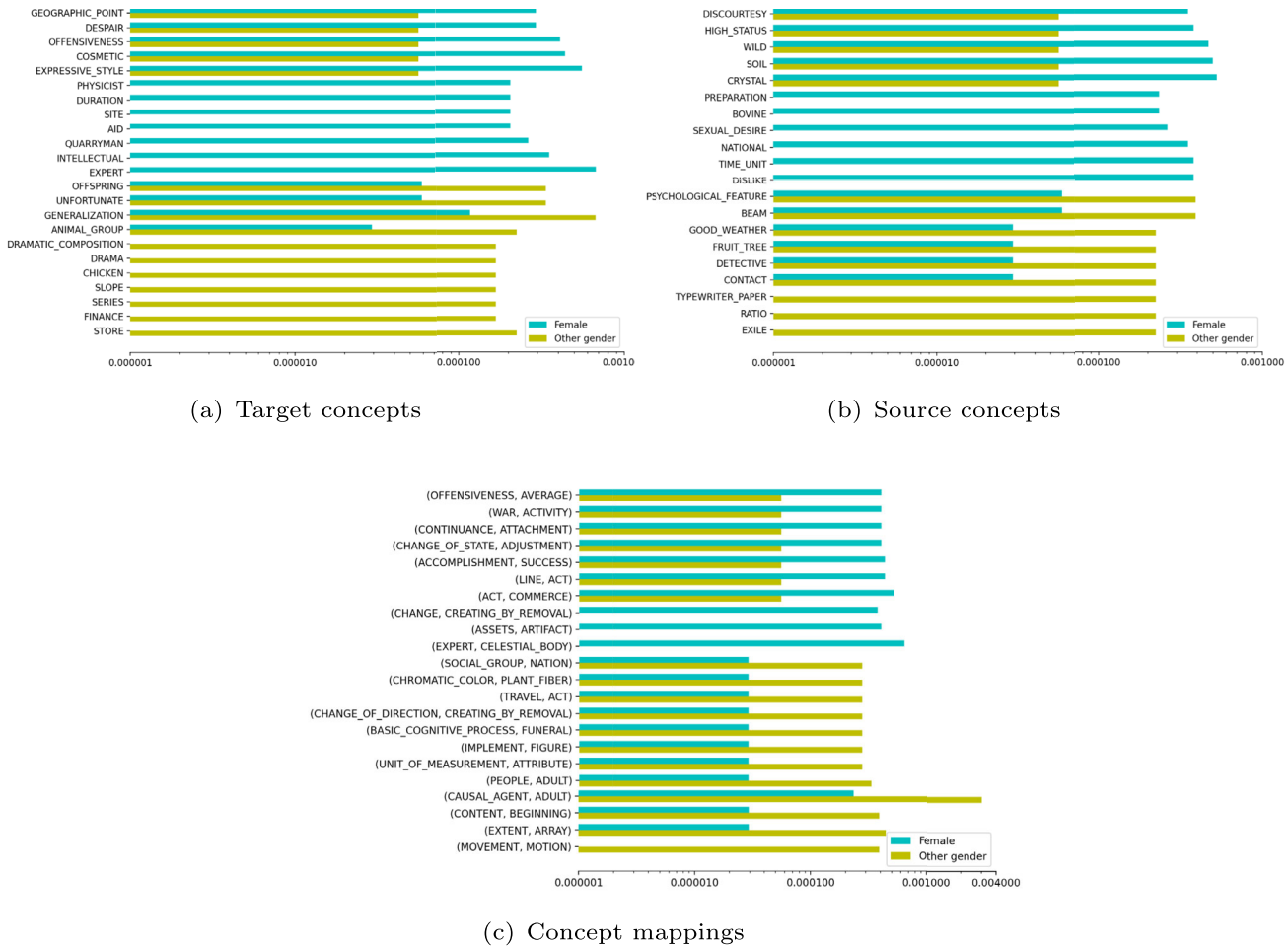
Identity groups	No. of positive records	% in toxic language
Male	**5122**	**6.28798%**
Female	**5778**	**7.09331%**
Transgender	**382**	**0.46896%**
Other_gender	3	0.00368%
Heterosexual	**263**	**0.32287%**
Homosexual_gay_or_lesbian	**2292**	**2.81375%**
Bisexual	50	0.06138%
Other_sexual_orientation	1	0.00123%
Black	**3485**	**4.27833%**
White	**5232**	**6.42302%**
Asian	429	0.52666%
Latino	302	0.37075%
Other_race_or_ethnicity	68	0.08348%



(a) Target concepts (b) Concept mappings

Fig. 6 Frequency comparison between toxic language mentioning males and other genders. The x-axis is on a logarithmic scale

Tables 10 and 11 show the top 10 association rules of concept mappings generated from toxic language mentioning males and females, respectively. All their top 10 rules involve (MEASURE, SIDE), (INFORMATION, DATABASE), and (FRIGHTFULNESS, ILL_HEALTH). These three concept mappings describe scenarios of misrepresenting data, based on specific sides, making skewed evaluations, and harnessing fear to manipulate individuals mentally. The language might



(a) Target concepts (b) Source concepts

(c) Concept mappings

Fig. 7 Frequency comparison between toxic language mentioning females and other genders. The x-axis is on a logarithmic scale

Table 10 Top 10 association rules of concept mappings in toxic language mentioning males ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(INFORMATION, DATABASE)	(MEASURE, SIDE)	0.00625	0.94118	133.90850
(MEASURE, SIDE)	(INFORMATION, DATABASE)	0.00625	0.88889	133.90850
(FRIGHTFULNESS, ILL_HEALTH), (MEASURE, SIDE)	(INFORMATION, DATABASE)	0.00605	1.00000	150.64706
(FRIGHTFULNESS, ILL_HEALTH), (INFORMATION, DATABASE)	(MEASURE, SIDE)	0.00605	1.00000	142.27778
(MEASURE, SIDE), (INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH)	0.00605	0.96875	141.76964
(INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH), (MEASURE, SIDE)	0.00605	0.91176	150.64706
(INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH)	0.00605	0.91176	133.43025
(FRIGHTFULNESS, ILL_HEALTH)	(MEASURE, SIDE), (INFORMATION, DATABASE)	0.00605	0.88571	141.76964
(FRIGHTFULNESS, ILL_HEALTH)	(INFORMATION, DATABASE)	0.00605	0.88571	133.43025
(FRIGHTFULNESS, ILL_HEALTH)	(MEASURE, SIDE)	0.00605	0.88571	126.01746

lead to unfair discrimination, misinformation, and negative health consequences.

Furthermore, the top 10 rules of toxic language mentioning females also include (COMMUNICATION, COMPONENT). The distinctive association rules indicate a gender-specific bias and stereotype reinforcement within the toxic language. (COMMUNICATION, COMPONENT) might suggest a toxic scenario where communication plays a significant role in concerns related to females.

Toxic Language Mentioning Transgender

We run Chi-square tests for homogeneity on the target concepts, source concepts, and concept mappings in toxic language mentioning transgender and other genders. The three p-values are all less than 0.05. Thus, we reject the null hypotheses that the toxic language mentioning transgender and other genders have the same target concept, source concept, or concept mapping preferences. The findings indicate cognitive distinctions between toxic language mentioning transgenders and other genders.

Table 12 shows the top 10 association rules of concept mappings generated from toxic language mentioning transgender. Most rules are combination of (INVESTIGATION, COGNITION), (ACT, SEPARATION), and (LAKE, BODY_OF_WATER). (INVESTIGATION, COGNITION) shows a mindset to compare an investigation activity to a cognitive activity. It might suggest that when mentioning transgender topics in a toxic context, a strong emphasis exists on intellectual cognition, possibly overshadowing empathy or

Table 11 Top 10 association rules of concept mappings in toxic language mentioning females ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(FRIGHTFULNESS, ILL_HEALTH)	(COMMUNICATION, COMPONENT), (INFORMATION, DATABASE)	0.00450	1.00000	222.23077
(COMMUNICATION, COMPONENT), (INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH)	0.00450	1.00000	222.23077
(FRIGHTFULNESS, ILL_HEALTH)	(COMMUNICATION, COMPONENT), (MEASURE, SIDE)	0.00450	1.00000	222.23077
(COMMUNICATION, COMPONENT), (MEASURE, SIDE)	(FRIGHTFULNESS, ILL_HEALTH)	0.00450	1.00000	222.23077
(FRIGHTFULNESS, ILL_HEALTH)	(MEASURE, SIDE), (INFORMATION, DATABASE)	0.00450	1.00000	222.23077
(MEASURE, SIDE), (INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH)	0.00450	1.00000	222.23077
(FRIGHTFULNESS, ILL_HEALTH)	(COMMUNICATION, COMPONENT), (MEASURE, SIDE), (INFORMATION, DATABASE)	0.00450	1.00000	222.23077
(COMMUNICATION, COMPONENT), (FRIGHTFULNESS, ILL_HEALTH)	(MEASURE, SIDE), (INFORMATION, DATABASE)	0.00450	1.00000	222.23077
(COMMUNICATION, COMPONENT), (INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH), (MEASURE, SIDE)	0.00450	1.00000	222.23077
(COMMUNICATION, COMPONENT), (MEASURE, SIDE)	(FRIGHTFULNESS, ILL_HEALTH), (INFORMATION, DATABASE)	0.00450	1.00000	222.23077

Table 12 Top 10 association rules of concept mappings in toxic language mentioning transgender ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(INVESTIGATION, COGNITION)	(ACT, SEPARATION), (LAKE, BODY_OF_WATER)	0.00785	1.00000	127.33333
(ACT, SEPARATION), (LAKE, BODY_OF_WATER)	(INVESTIGATION, COGNITION)	0.00785	1.00000	127.33333
(INVESTIGATION, COGNITION)	(LAKE, BODY_OF_WATER)	0.00785	1.00000	95.50000
(ACT, SEPARATION), (INVESTIGATION, COGNITION)	(LAKE, BODY_OF_WATER)	0.00785	1.00000	95.50000
(INVESTIGATION, COGNITION)	(ACT, SEPARATION)	0.00785	1.00000	76.40000
(INVESTIGATION, COGNITION), (LAKE, BODY_OF_WATER)	(ACT, SEPARATION)	0.00785	1.00000	76.40000
(MILITARY_ACTION, GROUP_ACTION)	(ACTION, DISAPPEARANCE)	0.00785	1.00000	42.44444
(LAKE, BODY_OF_WATER)	(INVESTIGATION, COGNITION)	0.00785	0.75000	95.50000
(LAKE, BODY_OF_WATER)	(ACT, SEPARATION), (INVESTIGATION, COGNITION)	0.00785	0.75000	95.50000
(LAKE, BODY_OF_WATER)	(ACT, SEPARATION)	0.00785	0.75000	57.30000

emotional understanding. The concept mapping (ACT, SEPARATION) suggests an opinion perceiving actions related to transgender individuals as separative, emphasizing differences rather than inclusion. (LAKE, BODY_OF_WATER) is a relative straightforward concept mapping. It might imply oversimplifying complex identities, stereotyping transgender experiences, and overlooking the diversity of transgender narratives.

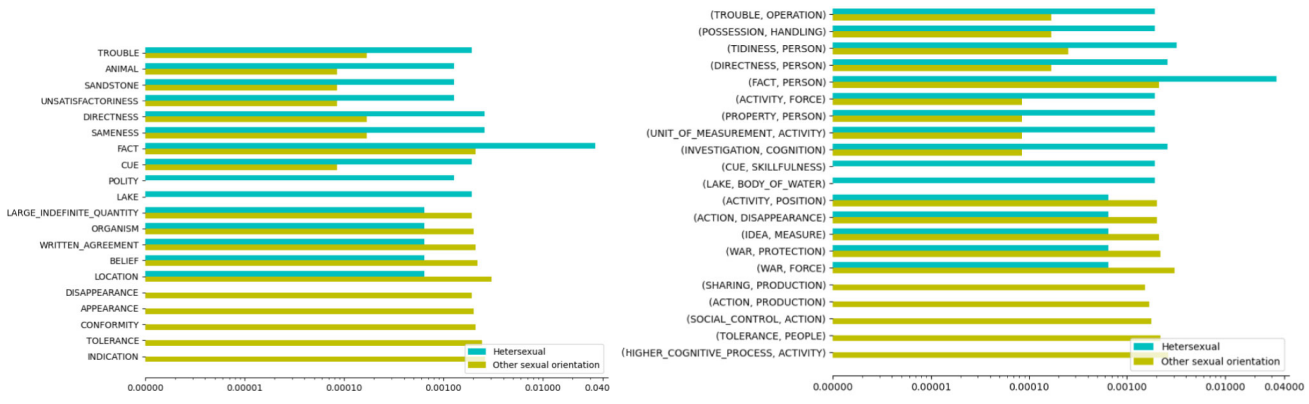
(MILITARY_ACTION, GROUP_ACTION) implies a perception that actions related to transgender topics are seen as coordinated with collective efforts. (ACTION, DISAPPEARANCE) might signify an interpretation where certain actions lead to the erasure of specific identities or narratives within the discourse. (MILITARY_ACTION, GROUP_ACTION) $\Rightarrow$ (ACTION, DISAPPEARANCE) might indicate a correlation between military or group-related actions and the disappearance of identity and its recognition.

The association rule (MILITARY_ACTION, GROUP_ACTION) $\Rightarrow$ (ACTION, DISAPPEARANCE) reflects a troubling scenario where toxic language is not just about expressing hostility but involves organized efforts to actively erase their visibility. This underscores the need for robust measures to protect and support transgender individuals in various aspects of society.

Cognitive Biases Among Toxic Language Mentioning Sexual Orientation

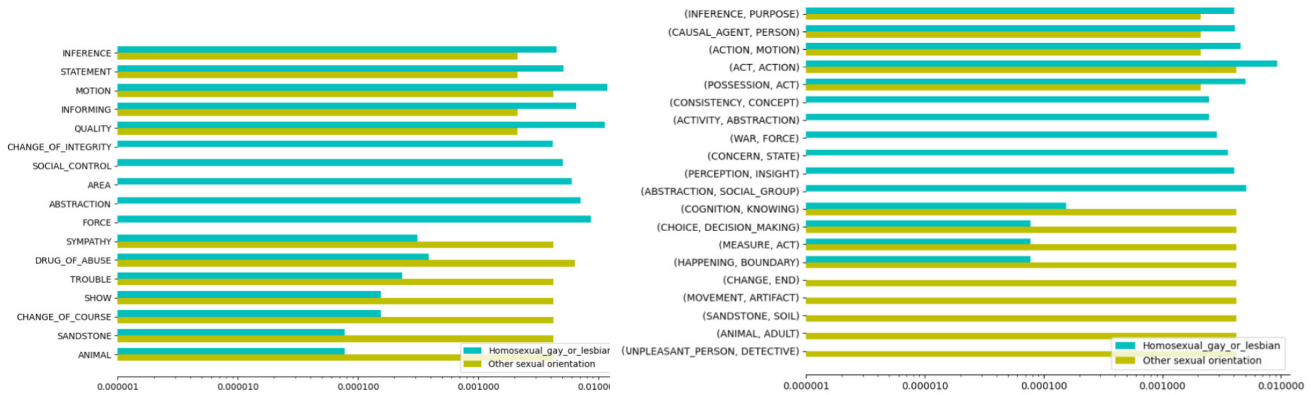
We run Chi-square tests for homogeneity on the target concepts, source concepts, and concept mappings in toxic language mentioning heterosexual/homosexual_gay_or_lesbian and other sexual orientations. The three p-values are all less than 0.05. Thus, we reject the null hypotheses that the toxic language mentioning heterosexual/homosexual_gay_or_lesbian and other sexual orientations have the same target concept, source concept, or concept mapping preferences. The findings indicate cognitive distinctions between toxic language mentioning heterosexual/homosexual_gay_or_lesbian and other sexual orientations.

Figure 8 shows the frequency comparison of target concepts and concept mappings between toxic language mentioning heterosexual and other sexual orientations. The



(a) Target concepts (b) Concept mappings

Fig. 8 Frequency comparison between toxic language mentioning heterosexual and other sexual orientations. The x-axis is on a logarithmic scale



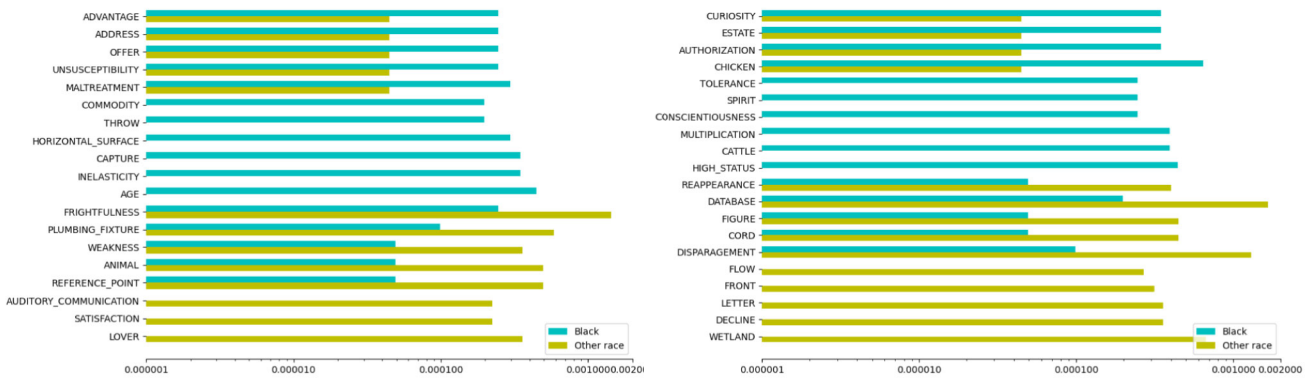
(a) Target concepts (b) Concept mappings

Fig. 9 Frequency comparison between toxic language mentioning homosexual_gay_or_lesbian and other sexual orientations. The x-axis is on a logarithmic scale

target concepts in toxic language mentioning heterosexual convey more negative sentiment (UNSATISFACTORINESS; TROUBLE). The concept mappings in toxic language mentioning heterosexual tend to link to natural entities ((LAKE, BODY_OF_WATER); (DIRECTNESS, PERSON); (TIDINESS, PERSON)), while the concept mappings in toxic language mentioning other sexual orientation focus more on social interactions, e.g., (SOCIAL_CONTROL, ACTION); (SHARING, PRODUCTION); (WAR, FORCE); (WAR, PROTECTION); (ACTION, DISAPPEARANCE).

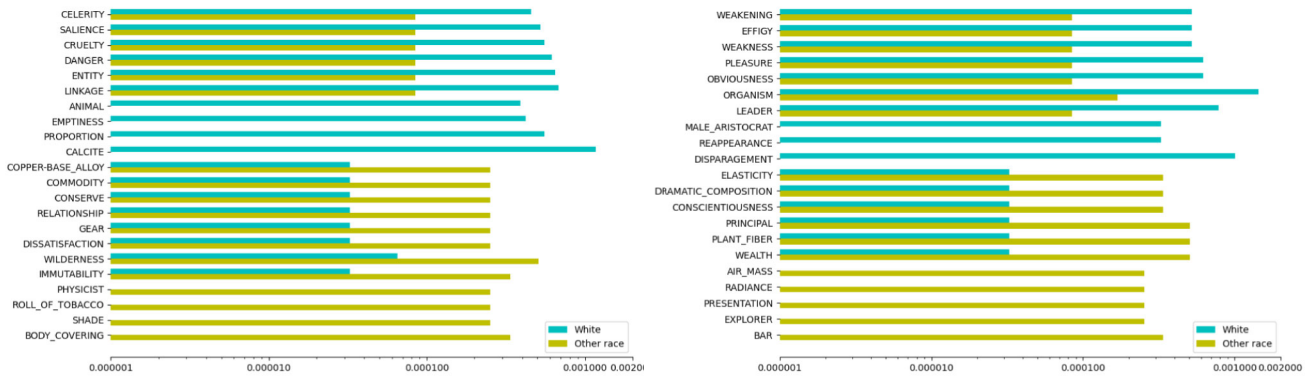
Figure 9 shows the frequency comparison of target concepts and concept mappings between toxic language mentioning homosexual_gay_or_lesbian and other sexual orientations. The target concepts in toxic language mentioning homosexual_gay_or_lesbian are more abstract than those mentioning other sexual orientations. The concept mappings in toxic language mentioning homosexual_gay_or_lesbian that usually reflect complex societal perceptions or attitudes, such as (PERCEPTION, INSIGHT); (CONCERN, STATE); (CONSISTENCY, CONCEPT); (INFERENCE, PURPOSE). Concept mappings in toxic language mentioning other sexual orientations employ more direct metaphors with tangible comparisons.

In summary, toxic language mentioning heterosexuals employs more direct and concrete metaphors that emphasize naturalness or straightforwardness, reflecting conventional views and focusing on specific behaviors or traits. Metaphors mentioning homosexual individuals often use more abstract and complex source concepts to present societal perceptions, roles, and actions that deviate from traditional norms. These differences highlight how societal attitudes and biases are reflected and reinforced through language.



(a) Target concepts (b) Source concepts

Fig. 10 Frequency comparison between toxic language mentioning Black and other races. The x-axis is on a logarithmic scale



(a) Target concepts (b) Source concepts

Fig. 11 Frequency comparison between toxic language mentioning White and other races. The x-axis is on a logarithmic scale

Cognitive Biases Among Toxic Language Mentioning Races

We run Chi-square tests for homogeneity on the target concepts, source concepts, and concept mappings in toxic language mentioning Black/White and other races. The three p-values are all less than 0.05. Thus, we reject the null hypotheses that the toxic language mentioning Black/White and other races have the same target concept, source concept, or concept mapping preferences. The findings indicate cognitive distinctions between toxic language mentioning Black/White and other sexual orientations.

Figure 10 shows the frequency comparison of target and source concepts between toxic language mentioning Black and other races. The target concepts in toxic language mentioning Black include actions (CAPTURE; THROW; MALTREATMENT). The source concepts relate to societal status (HIGH_STATUS; ESTATE), stereotypes (CATTLE; CHICKEN), and personal traits (CONSCIENTIOUSNESS; TOLERANCE; CURIOSITY), while those mentioning other races carry negative attributes (DECLINE; DISPARAGEMENT).

Figure 11 shows the frequency comparison of target and source concepts between toxic language mentioning White and other races. The target concepts in toxic language mentioning White are more abstract. The source concepts relate to societal status (MALE_ARISTOCRAT; LEADER) and negative concepts (DISPARAGEMENT; WEAKNESS; WEAKENING).

Toxic language mentioning Black and White individuals both involve societal roles. However, metaphors in toxic language mentioning Black focus on stereotypes, personal traits, and behaviors, while those mentioning White focus on negative attributes and judgments.

Table 13 shows the top 10 association rules of concept mappings generated from toxic language mentioning White. Compared with Table 10, we can observe that the top 10 association rules of toxic language mentioning males and White people are quite similar. This similarity might reflect a reinforcement of power structures and dominant narratives within toxic language. Both males and White people are portrayed as majorities, based on existing power dynamics and societal hierarchies, implying that speakers share the same narratives and cognitive patterns toward these two groups.

Table 13 Top 10 association rules of concept mappings in toxic language mentioning White ranked by support, confidence, and lift

Antecedent	Consequent	Support	Conf.	Lift
(INFORMATION, DATABASE)	(MEASURE, SIDE)	0.00612	0.94118	136.78431
(MEASURE, SIDE)	(INFORMATION, DATABASE)	0.00612	0.88889	136.78431
(FRIGHTFULNESS, ILL_HEALTH), (MEASURE, SIDE)	(INFORMATION, DATABASE)	0.00593	1.00000	153.88235
(FRIGHTFULNESS, ILL_HEALTH), (INFORMATION, DATABASE)	(MEASURE, SIDE)	0.00593	1.00000	145.33333
(MEASURE, SIDE), (INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH)	0.00593	0.96875	153.59091
(FRIGHTFULNESS, ILL_HEALTH)	(MEASURE, SIDE), (INFORMATION, DATABASE)	0.00593	0.93939	153.59091
(FRIGHTFULNESS, ILL_HEALTH)	(INFORMATION, DATABASE)	0.00593	0.93939	144.55615
(FRIGHTFULNESS, ILL_HEALTH)	(MEASURE, SIDE)	0.00593	0.93939	136.52525
(INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH), (MEASURE, SIDE)	0.00593	0.91176	153.88235
(INFORMATION, DATABASE)	(FRIGHTFULNESS, ILL_HEALTH)	0.00593	0.91176	144.55615

Recognizing these associations in toxic language is crucial for promoting inclusive and respectful dialogue. It is essential to foster understanding, empathy, and a recognition of the diversity within different communities, avoiding the reinforcement of harmful stereotypes or oversimplified associations in discussions surrounding gender, sexual orientation, and race. Encouraging open-mindedness and respecting individuality are vital in creating a more inclusive and supportive environment.

Conclusion

This study explores the cognitive patterns of toxic language, compares the cognitive biases between toxic and non-toxic language, and finds interesting insights into identity categories of toxic language. We employ conceptual metaphor processing on 1,015,290 data records sourced from a civil commenting platform from 2015 to 2017.

Our experiment results indicate cognitive distinctions of the target concepts, source concepts, and concept mappings between toxic and non-toxic language, subdivided levels and subtypes of toxic language as well as toxic language mentioning different genders, sexual orientations, and races. Toxic language contains more negative sentiments and intentions than non-toxic language. The obscene subtype shows aggressive and strong emotions, while the sexual_explicit subtype involves direct physical or sexual comparison.

Metaphors in toxic language mentioning females, including concepts, such as (OFFENSIVENESS; DESPAIR; DISLIKE; DISCOURTESY); and the concept mapping, such as (CHANGE, CREATING_BY_REMOVAL)), express negative sentiments and undermine the competence of women. Efforts should be made to address the bias against female competence in the job market. Roth et al. [57] conducted a meta-analysis of job performance measures from field studies. They found that females generally scored slightly higher than males. Other analyses suggested that, although job performance ratings favored females, ratings of promotion potential were higher for males.

The frequent concept mappings in toxic language mentioning transgender (ACT, SEPARATION); (LAKE, BODY_OF_WATER) suggest an oversimplified perception regarding transgender individuals as separative or equal. Diamond et al. [58] presented more flexible and broader models of gender identity development among transgender individuals, accepting that identity development can have a linear trajectory leading to a singular outcome or a recursive process that accommodates multiple and shifting identity states over time.

Toxic language mentioning heterosexuals emphasizes naturalness and conventional views with direct metaphors, while toxic language mentioning homosexual individuals often uses more abstract metaphors ((PERCEPTION, INSIGHT); (CONCERN, STATE); (INFERENCE, PURPOSE)) to comment on phenomena that deviate from traditional norms. Simon and Gagnon [59] offered a sociological view on homosexuality, arguing against treating it as merely deviant behavior. Instead, it proposed understanding homosexuality within the broader context of social and sexual roles, emphasizing the diversity and fluidity of homosexual as well as non-sexual roles.

Our findings offer valuable hypotheses for developing toxic language research using data mining techniques. Researchers can further verify the cognitive findings with human-involved laboratory tests. By leveraging concept mappings and analysis, platform operators and developers can deploy automated systems to detect and flag toxic content, enabling more efficient performance at scale. This can help create safer and more inclusive online environments for users, particularly those from marginalized or vulnerable communities.

Author Contributions Mengshi Ge: methodology, formal analysis, data curation, visualization, writing—original draft; Rui Mao: validation, formal analysis, data curation, investigation, writing—review and editing; Erik Cambria: resources, writing - review and editing, supervision, project administration.

Funding This research/project is supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005)

Data Availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Cambria E. Understanding natural language understanding. Springer; 2024. ISBN 978-3-031-73973-6
2. Cambria E, Zhang X, Mao R, Chen M, Kwok K. SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. In: Proceedings of international conference on human-computer interaction (HCII), Washington DC, USA; 2024. p. 197–216.
3. Mao R, Liu Q, He K, Li W, Cambria E. The biases of pre-trained language models: an empirical study on prompt-based sentiment analysis and emotion detection. IEEE Trans Affect Comput. 2023;14(3):1743–53.
4. Ekiciler A, Ahioğlu İ, Yıldırım N, Ajas II, Kaya T. The bullying game: sexism based toxic language analysis on online games chat logs by text mining. J Int Women's Stud. 2022;24(3):1–16.
5. Kumar A, Abirami S, Trueman TE, Cambria E. Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit. Neurocomputing. 2021;441:272–8.
6. Pelzer B, Kaati L, Cohen K, Fernquist J. Toxic language in online incel communities. SN Soc Sci. 2021;1:1–22.

7. Bhat MM, Hosseini S, Hassan A, Bennett P, Li W. Say 'yes'to positivity: detecting toxic language in workplace communications. In: Findings of the association for computational linguistics: EMNLP 2021; 2021. p. 2017–29.
8. Kwak H, Blackburn J. Linguistic analysis of toxic behavior in an online video game. In: Social informatics: SocInfo 2014 international workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers 6. Springer; 2015. p. 209–17.
9. Nexø LA, Kristiansen S. Players don't die, they respawn: a situational analysis of toxic encounters arising from death events in league of legends. Eur J Crim Pol Res. 2023;29(3):457–76.
10. Kanna RK, Mutheeswaran U, Jouda AJ, Hussein MA, Hussain A, Al-Tahee M. Computational cognitive analysis of ADHD patients using Matlab applications. In: 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE; 2023. p. 1344–8.
11. Montuori C, Gambarota F, Altoé G, Arfé B. The cognitive effects of computational thinking: a systematic review and meta-analytic study. Comput Educ. 2024;210:104961.
12. Ottati V, Rhoads S, Graesser AC. The effect of metaphor on processing style in a persuasion task: a motivational resonance model. J Person Soc Psychol. 1999;77(4):688.
13. Lakoff G. Moral politics: how liberals and conservatives think. Chicago, IL: University of Chicago Press; 2002.
14. Ang SH, Lim EAC. The influence of metaphors and product type on brand personality perceptions and attitudes. J Advert. 2006;35(2):39–53.
15. Mao R, Zhang T, Liu Q, Hussain A, Cambria E. Unveiling diplomatic narratives: analyzing United Nations Security Council debates through metaphorical cognition. In: Proceedings of the annual meeting of the cognitive science society (CogSci), vol. 46. Rotterdam, the Netherlands; 2024. p. 1709–16.
16. Lakoff G, Johnson M. Metaphors we live by. Chicago, IL: University of Chicago; 1980.
17. Crawford LE. Conceptual metaphors of affect. Emotion Rev. 2009;1(2):129–39.
18. Nayak NP, Gibbs RW. Conceptual knowledge in the interpretation of idioms. J Exp Psychol Gen. 1990;119(3):315.
19. Allbritton DW, McKoon G, Gerrig RJ. Metaphor-based schemas and text representations: making connections through conceptual metaphors. J Exp Psychol Learn Mem Cogn. 1995;21(3):612.
20. Boroditsky L, Ramscar M. The roles of body and mind in abstract thought. Psychol Sci. 2002;13(2):185–9.
21. Ge M, Mao R, Cambria E. A survey on computational metaphor processing techniques: from identification, interpretation, generation to application. Artif Intell Rev. 2023;56(Suppl 2):1829–95.
22. Mao R, Li X, He K, Ge M, Cambria E. MetaPro online: a computational metaphor processing online system. In: Proceedings of the 61st annual meeting of the association for computational linguistics (Volume 3: System Demonstrations);2023. vol. 3, p. 127–35.
23. Mao R, Ge M, Han S, Li W, He K, Zhu L, Cambria E. A survey on pragmatic processing techniques. Inf Fusion. 2025;114:102712.
24. Mu Y, Bontcheva K, Aletras N. It's about time: rethinking evaluation on rumor detection benchmarks using chronological splits. In: Findings of the association for computational linguistics: EACL 2023; 2023. p. 736–743.
25. Mu Y, Song X, Bontcheva K, Aletras N. Examining the limitations of computational rumor detection models trained on static datasets. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024); 2024. p. 6739–51.
26. Xia M, Field A, Tsvetkov Y. Demoting racial bias in hate speech detection. In: Proceedings of the eighth international workshop on natural language processing for social media; 2020. p. 7–14.
27. Halevy M, Harris C, Bruckman A, Yang D, Howard A. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In: Equity and access in algorithms, mechanisms, and optimization; 2021. p. 1–11.
28. Chuang Y-S, Gao M, Luo H, Glass J, Lee H-Y, Chen Y-N, Li S-W. Mitigating biases in toxic language detection through invariant rationalization. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021); 2021. p. 114–20.
29. Sahoo N, Gupta H, Bhattacharyya P. Detecting unintended social bias in toxic language datasets. In: Proceedings of the 26th conference on computational natural language learning (CoNLL); 2022. p. 132–43.
30. Gevers I, Markov I, Daelemans W. Linguistic analysis of toxic language on social media. Comput Linguist Netherlands J. 2022;12:33–48.
31. Sharma S, et al. Content analysis of item songs: reflections of a toxic socio-cultural milieu. Turk J Comput Math Educ (TURCOMAT). 2021;12(10):3856–61.
32. Hu R, Wang X. A cognitive pragmatic analysis of conceptual metaphor in political discourse based on text data mining. In: 2021 4th International conference on information systems and computer aided education; 2021. p. 235–8.
33. Chen, X.: The greenhouse metaphor and the greenhouse effect: a case study of a flawed analogous model. In: Philosophy and cognitive science: western & eastern studies, Springer; 2012. p. 105–14.
34. Wang Z, Wang L, Yu S. A metaphorical and cognitive study on idioms with "Ru". In: Chinese lexical semantics: 17th workshop, CLSW 2016, Singapore, Singapore, May 20–22, 2016, Revised Selected Papers 17. Springer; 2016. p. 534–45.
35. Dodge EK, Hong J, Stickles E. MetaNet: deep semantic automatic metaphor analysis. In: Proceedings of the third workshop on metaphor in NLP; 2015. p. 40–9.
36. Lachaud CM. Conceptual metaphors and embodied cognition: EEG coherence reveals brain activity differences between primary and complex conceptual metaphors during comprehension. Cognit Syst Res. 2013;22:12–26.
37. Fu C, Wang J, Sang J, Yu J, Xu C. Beyond literal visual modeling: understanding image metaphor based on literal-implied concept mapping. In: Multimedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part I 26. Springer; 2020. p. 111–23
38. Li H, Zhu KQ, Wang H. Data-driven metaphor recognition and explanation. Trans Assoc Comput Linguist. 2013;1:379–90.
39. Rosen Z. Computationally constructed concepts: a machine learning approach to metaphor interpretation using usage-based construction grammatical cues. In: Proceedings of the workshop on figurative language processing; 2018. pp. 102–9.
40. Han S, Mao R, Cambria E. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In: Proceedings of the 29th international conference on computational linguistics (COLING) 2022; p. 94–104.
41. Mao R, Du K, Ma Y, Zhu L, Cambria E. Discovering the cognition behind language: financial metaphor analysis with MetaPro. In: 2023 IEEE International Conference on Data Mining (ICDM). IEEE; 2023. p. 1211–6.
42. Jia M, Mao R, Xie Y, Ren S, Cambria E. Analyzing the cognitive impact of trauma from a metaphorical perspective: a case study on the attempted assassination of Donald Trump. In: 2025 IEEE Symposium Series on Computational Intelligence (SSCI), Trondheim, Norway; 2025.
43. Mao R, Lin Q, Liu Q, Mengaldo G, Cambria E. Understanding public perception towards weather disasters through the lens of metaphor. In: Proceedings of the thirty-third international joint Conference on Artificial Intelligence, (IJCAI-24); 2024. p. 7394–402.

44. Manro R, Mao R, Dahiya L, Ma Y, Cambria E. A cognitive analysis of CEO speeches and their effects on stock markets. In: Proceedings of the 5th International Conference on Financial Technology (ICFT), Singapore; 2024.
45. Mao R, Chen G, Li X, Ge M, Cambria E. A comparative analysis of metaphorical cognition in ChatGPT and human minds. Cognit Comput. 2025;17.
46. Borkan D, Dixon L, Sorensen J, Thain N, Vasserman L. Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion proceedings of the 2019 world wide web conference; 2019. p. 491–500.
47. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers); 2016. p. 1715–25.
48. Mao R, Li X. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. In: Proceedings of the AAAI conference on artificial intelligence; 2021. vol. 35, p. 13534–42.
49. Mao R, Li X, Ge M, Cambria E. MetaPro: a computational metaphor processing model for text pre-processing. Inf Fusion. 2022;86–87:30–43.
50. Ge M, Mao R, Cambria E. Explainable metaphor identification inspired by conceptual metaphor theory. In: Proceedings of the AAAI conference on artificial intelligence; 2022. vol. 36, p. 10681–9.
51. Mao R, He K, Ong CB, Liu Q, Cambria E. MetaPro 2.0: computational metaphor processing on the effectiveness of anomalous language modeling. In: Findings of the association for computational linguistics: ACL; 2024. p. 9891–908.
52. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on management of data; 1993. p. 207–16.
53. Adams CJ, Borkan D, Sorensen J, Dixon L, Vasserman L, Thain N. Jigsaw unintended bias in toxicity classification. Kaggle; 2019. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification
54. Brassard-Gourdeau E, Khoury R. Subversive toxicity detection using sentiment information. In: Proceedings of the third workshop on abusive language online; 2019. p. 1–10.
55. Nejadgholi I, Balkir E, Fraser KC, Kiritchenko S. Towards procedural fairness: uncovering biases in how a toxic language classifier uses sentiment information. In: Proceedings of the fifth BlackboxNLP workshop on analyzing and interpreting neural networks for NLP; 2022. p. 225–37.
56. Zhang Y, Ding L, Zhang L, Tao D. Intention analysis prompting makes large language models a good jailbreak defender; 2024. arXiv:2401.06561
57. Roth PL, Purvis KL, Bobko P. A meta-analysis of gender group differences for measures of job performance in field studies. J Manag. 2012;38(2):719–39.
58. Diamond LM, Pardo ST, Butterworth MR. Transgender experience and identity. Handbook of identity theory and research; 2011. 629–47.
59. Simon W, Gagnon JH. Homosexuality: the formulation of a sociological perspective. J Health Soc Behav. 1967;177–85.