applied sciences

MDPI

Article

# Automatically Distinguishing People's Explicit and Implicit Attitude Bias by Bridging Psychological Measurements with Sentiment Analysis on Large Corpora

**Bo Wang [1,2,*], Runxin Zhang [1], Baixiang Xue [2], Yachao Zhao [2], Li Yang [1] and Hongxiang Liang [2]**

[1] Institute of Applied Psychology, Tianjin University, Tianjin 300350, China; 2021212061@tju.edu.cn (R.Z.); yangli@tju.edu.cn (L.Y.)
[2] College of Intelligence and Computing, Tianjin University, Tianjin 300072, China; baixiangxue@tju.edu.cn (B.X.); 2022244009@tju.edu.cn (Y.Z.); superlianghx@gmail.com (H.L.)
* Correspondence: bo_wang@tju.edu.cn

**Abstract:** Social psychological studies show that people's explicit attitude bias in public expression can differ from their implicit attitude bias in mind. However, the current automatic attitude analysis does not distinguish between explicit and implicit attitude bias. Simulating the psychological measurements of explicit and implicit attitude bias, i.e., self-report assessment (SRA) and implicit association test (IAT), we propose an automatic language-based analysis to distinguish explicit and implicit attitude bias in a large population. By connecting the criteria of SRA and IAT with the statements containing patterns of special words, we derive explicit and implicit attitude bias with the sentiment scores of the statements, which are obtained by pre-trained machine-learning methods. Extensive experiments on four English and Chinese corpora and four pairs of concepts show that the attitude biases obtained by our method on a large population are consistent with those of traditional psychological experiments in the costly small-scale experiments. The maximum gap between the sentiment scores of explicit and implicit biases reaches 0.9329. Furthermore, we achieve new findings on the difference between the evolution of explicit and implicit attitude bias. The maximum variance gap of sentiment scores in the dynamic changes between explicit and implicit biases reaches 0.249.

**Keywords:** attitude bias; self-report assessment; implicit association test; sentiment analysis

check for updates

**Citation:** Wang, B.; Zhang, R.; Xue, B.; Zhao, Y.; Yang, L.; Liang, H. Automatically Distinguishing People's Explicit and Implicit Attitude Bias by Bridging Psychological Measurements with Sentiment Analysis on Large Corpora. *Appl. Sci.* **2024**, *14*, 4191. https://doi.org/10.3390/app14104191

Academic Editor: Raúl Quevedo-Blasco

Received: 13 March 2024
Revised: 29 April 2024
Accepted: 6 May 2024
Published: 15 May 2024

## 1. Introduction

Attitude bias (abbreviated as "bias") is people's tendentious beliefs about concepts. As an inherent motivation of people's interests and consequent behaviors, bias is essential to human-centered research, such as social psychology, computational opinion analysis, and recommendation. In current natural-language processing research, we can understand biases by analyzing the sentiment and semantics of people's public expressions. However, social psychology research reveals that people's explicitly expressed bias in public does not always agree with their implicit bias in mind [1]. People's explicit biases are conscious, controllable, and easy to report, while implicit biases are uncontrollable and cannot be consciously acquired [2]. Explicit and implicit biases play different roles in social life. Explicit bias spreads in public and forms the mainstream value, whereas implicit bias can determine the behavior without conscious awareness, e.g., voting and purchasing [3]. A well-known example is American racial bias, where self-reports reveal a near-absence of preference difference between White and Black people [4]. In contrast, implicit bias reveals a widespread preference for White people relative to Black people [5]. Implicit bias is also found to be a better predictor of certain kinds of behavior (e.g., risky flight [6], and anxiety [7]) than explicit bias.

In psychology, the Self-Report Assessment (SRA) [8] and the Implicit Association Test (IAT) [5] are capable of measuring explicit and implicit attitude bias, respectively.

However, both SRA and IAT need cooperation from subjects, and it is very costly for a large population. To understand the bias automatically, researchers apply natural-language processing techniques to understand bias with the semantic relations between people's words [9]. For example, Recasents [10] introduced that biased language in Wikipedia uncovers framing and epistemological bias and can be identified by familiar linguistic cues. Recent works show that word and sentence embedding can capture common bias in training corpora [11]. Word Embedding Association Test (WEAT) proposes to [12] measure the bias by regarding the semantic distance between pre-trained vectors of certain words as the implicit association strength in IAT. Sentence Encoder Association Test (SEAT) [13] extends the WEAT by constructing sentences containing specific words so that the word encoders (word2vec and GloVe) in WEAT can be replaced by the sentence encoders (ELMo and BERT). Recent work [14,15] explores the subjects of word embedding biases and finds that these biases are centered around males. Moreover, Dobrzeniecka et al. propose hierarchical Bayesian modeling (HBM) [16] as an alternative to WEAT. HBM scrutinizes biases in word embeddings across various levels of granularity and yields robust results in bias assessments concerning gender, race, and religion. However, WEAT, SEAT, and earlier language-based approaches do not distinguish explicit and implicit biases. Without this distinguishment, if explicit and implicit biases contained in language are different, the bias identified by current methods will be a confusion of explicit and implicit biases.

Against the background mentioned above, the primary research motivation of this paper is to draw inspiration from the psychological measurement of biases and employ computational techniques to address the problem of separately identifying explicit and implicit bias in a large population automatically.

Based on this motivation, our research objective is to propose a novel strategy to connect the key factors in SRA and IAT with special words and statements and automatically simulate SRA and IAT with sentiment analysis of the statements. The framework is shown in Figure 1. First, we define the categories of special words named concept words, exemplar words and attribute words, corresponding to the concept, exemplar and attribute in SRA and IAT. Second, we simulate explicit response in SRA and implicit response in IAT with statements containing different combinations of special words, respectively. Finally, by measuring the sentiment of the statements, we derive explicit and implicit bias separately.

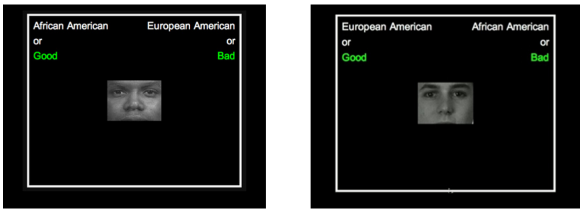


**Figure 1.** A typical interface of computer-aided IAT on the implicit bias of African-American vs. European-American.

In experiments on corpora of different languages (English and Chinese) and different types (Wikipedia and Social Media), our method successfully reproduces the well-known observations of SRA and IAT: (1) people's explicit and implicit biases can be different, especially on well-known cases reported in psychological studies, e.g., racial bias; (2) the difference degree is related to the sociality of the concepts, the scenario of language use, and the cultural background. Furthermore, we also find that implicit biases are more stable than explicit biases over time, and the stability is also related to social factors.

We summarize our contributions as follows:

(1) Simulate psychological tests, i.e., SRA and IAT, with automatic language-based analysis, and do not acquire extra modification on the state-of-the-art natural-language processing techniques.

(2) Automatically reproduce the small-scale psychological observations on a large population.

(3) Achieve new findings, indicating further advantages of the proposed automatic method compared with psychological tests.

## 2. Related Works

In this section, we will introduce the main related works in the fields of psychology and computer technology that are pertinent to the present study. These primarily include methods in psychology for measuring explicit and implicit biases, as well as word vector techniques and sentiment polarity measurement techniques in computer technology that can be used to assess biases.

### *2.1. Psychological Tests of Bias*

We commence by introducing the SRA and IAT methods, which are utilized in psychology to measure explicit and implicit biases.

Self-report assessment (SRA) is a typical method of measuring explicit bias. Reading questions and choosing responses in the assessment is the most direct method of asking about one's feelings, prejudices, beliefs, and other information [17]. SRA can be a survey, questionnaire, or poll in which subjects express their attitude explicitly. The main strength of SRA is allowing participants to describe their own experiences rather than inferring this from observing them.

SRA is more reliable for explicit bias than implicit bias [18]. However, SRA also has specific disadvantages because of the way subjects behave [19]. There are two main reasons for the untruthful response in SRA. The first reason is "social desirability bias", in which the subjects wish to present themselves in a socially acceptable manner. The second reason is the "lack of awareness" that the subjects are not aware of their genuine attitude in mind, which is also known as "unconscious attitude" [20].

There are also other reasons leading to untruthful responses in SRA: (1) The order of questions: response can be influenced by the different orders of questions. (2) Response options: The response of a given option can be influenced by the compared options. (3) The wording of questions: response also can be influenced by the description of the questions.

These disadvantages can be leading because they may unwittingly force the respondent to give a particular reply [21]. Therefore, we will not be collecting valid data based on the explicit attitude expressed by SRA. Psychologists have developed indirect measurements, such as the Implicit Association Test (IAT), to measure implicit bias.

Implicit Association Test(IAT) [5] is the most widely used psychological test of implicit bias. In contrast to SRA, in IAT, to avoid conscious explicit bias of the concept, the test is based on the exemplars of the concept (e.g., rose) instead of the concept itself (e.g., Flower). A set of exemplar stimuli (e.g., pictures or names of exemplars) and attribute stimuli are used to prime the concept and attribute in subjects' minds unconsciously. The subjects are encouraged to classify "exemplar stimuli + attribute stimuli" combinations. During the classification, their response (e.g., button click) latency measures the implicit association between the concepts w.r.t the attributes. Finally, an effect size of response latency is used to calculate the implicit bias [22].

For example, as shown in Figure 1, in computer-aided IAT https://implicit.harvard.edu/implicit/takeatest.html (accessed on 12 March 2024), to measure the implicit bias toward "African-American vs. European-American", given two keys on the left and right hand, respectively, an "exemplar stimuli + attribute stimuli" combination is shown to the subject every time. The left screenshot pairs 'African-American' with 'Good' and 'European-American' with 'Bad,' while the right screenshot does the opposite. In the center of each screenshot is an exemplar or attribute stimuli. If the subject's response under the instruction: "click left when you see an African-American or a pleasant stimulus; click right when you see a European-American or an unpleasant stimulus" is faster than his response under the instruction: "click left when you see a European-American or a pleasant stimulus; click right when you see an African-American or an unpleasant stimulus", IAT will say that the subject has an implicit preference for “African-American” compared with “European-American”.

Figure 2 illustrates a schematic diagram of a real scene where a subject is taking an IAT test in front of a computer. On the left, ‘Black’ is associated with ‘good’ and ‘White’ with ‘bad,’. However, on the right, the associations are inverted. The test measures the subject’s reaction times to these pairings to evaluate implicit racial attitudes.

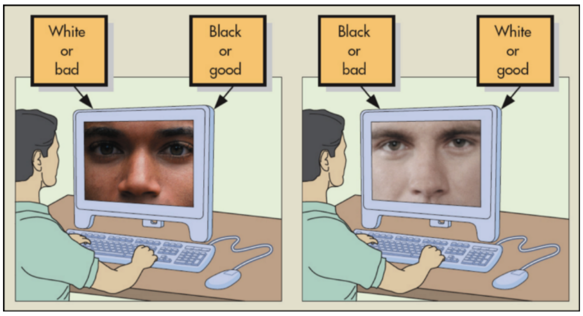


**Figure 2.** A schematic diagram of a real scene where a subject is taking an IAT test in front of a computer.

The following brief introduction of the Implicit Association Test (IAT) is selected from the home page of the IAT project: https://implicit.harvard.edu/implicit/index.jsp (accessed on 12 March 2024). On this website, users can try an IAT test online to learn more information.

The IAT measures the strength of associations between concepts (e.g., African-American, European-American) and attributes (e.g., good, bad) or stereotypes (e.g., athletic, clumsy). The main idea is that making a response is more accessible when closely related items share the same response key.

When doing an IAT, you are asked to quickly sort words into categories that are on the left and right-hand side of the computer screen by pressing the “e” key if the word belongs to the category on the left and the “i” key if the word belongs to the category on the right. The IAT has five main parts [23].

In the first part of the IAT, you sort words relating to the concepts (e.g., African-American, European-American) into categories. If the category “African-American” was on the left, and a picture of a Black person was on the screen, you would press the “e” key.

In the second part of the IAT, you sort words relating to the attributes (e.g., good, bad). So, if the category “good” was on the left, and a pleasant word appeared on the screen, you would press the “e” key.

In the third part of the IAT, the categories are combined, and you are asked to sort both concept and evaluation words. So, the categories on the left-hand side would be African-American/Good, and the categories on the right-hand side would be European-American/Bad. It is important to note that the order in which the blocks are presented varies across participants. Some people will do the African-American/Good, European-American/Bad part first, and others will do the African-American/Bad, European-American/Good part first.

The placement of the concepts switches in the fourth part of the IAT. If the category “African-American” were previously on the left, it would now be on the right. Notably, the number of trials in this part of the IAT is increased to minimize the effects of the practice.

In the final part of the IAT, the categories are combined in a way that is opposite to what they were before. If the category on the left were previously African-American/Good, it would now be African-American/Bad.

The IAT score is based on how long it takes a person, on average, to sort the words in the third part of the IAT versus the fifth part of the IAT. We would say that one has an implicit preference for thin people relative to fat people if they are faster to categorize words when Thin People and Good share a response key and Fat People and Bad share a response key relative to the reverse.

Although the Implicit Association Test (IAT) is currently the mainstream method for measuring implicit biases in psychological research, it has certain limitations when studying implicit biases in large-scale populations. First, the IAT requires participants to actively cooperate in the test in a laboratory or designated website, which raises the threshold for testing. Second, as participants are aware that they are being tested, the IAT falls into the category of contract testing, potentially affecting the test results. If we hope to conduct more accurate testing on a large-scale population, it is necessary to explore an automated and non-contact testing method.

### *2.2. Embedding Association Test of Bias*

Next, we introduce a commonly used computer technology that can measure biases, embedding association test (WEAT) [12]. Embedding is a semantic representation of words or sentences according to their context in the corpus [24]. Words that are closer in embedding space are supposed to be semantically similar, assuming that speakers' mental associations between concepts and attributes can be represented by the semantic similarity between the words naming the concepts and attributes. Word Embedding Association Test (WEAT) [12] is proposed to measure the bias with the distance between the embeddings of the words naming the exemplars and attributes in IAT.

The details of the WEAT are as follows. It calculates the implicit association between two concept words and two sets of attribute words using Equations (1) and (2). Given two exemplar words set $Ew_i$, $Ew_j$ corresponding to two concepts $C_i$, $C_j$, respectively (e.g., $\{orchid, rose, \ldots\}$ for the concept "*Flower*" and $\{ant, spider, \ldots\}$ for the concept "*Insect*"), and two attribute words set $Aw_p$, $Aw_q$ corresponding to two attributes $A_p$, $A_q$, respectively (e.g., $\{happy, great, pretty, \ldots\}$ for the attribute "*Pleasant*" and $\{awful, ugly, sad, \ldots\}$ for the attribute "*Unpleasant*"), the null hypothesis is that there is no difference between $Ew_i$, $Ew_j$ in terms of their relative similarity to $Aw_p$, $Aw_q$. Let $\vec{x}$ be the embedding vector of word $x$, $cos(\vec{x}, \vec{y})$ denotes the cosine similarity between $\vec{x}$ and $\vec{y}$. Specifically, WEAT [12] calculated the bias on $C_i$, $C_j$ with respect to $A_p$, $A_q$ as:

$$Bias(Ew_i, Ew_j, Aw_p, Aw_q) = \sum_{e\in Ew_i} s(e, Aw_p, Aw_q) - \sum_{e\in Ew_j} s(e, Aw_p, Aw_q) \tag{1}$$

where

$$s(e, Aw_p, Aw_q) = \underset{a\in Aw_p}{\text{mean}}\, cos(\vec{e}, \vec{a}) - \underset{a\in Aw_q}{\text{mean}}\, cos(\vec{e}, \vec{a}) \tag{2}$$

Then, WEAT [12] measures the strength of association through the effect size. The effect size ('ES' in Equation (3)) of the bias is calculated as:

$$ES = \frac{\underset{e\in Ew_j}{\text{mean}}\, s(e,Aw_p,Aw_q) - \underset{e\in Ew_i}{\text{mean}}\, s(e,Aw_p,Aw_q)}{\underset{e\in Ew_i\cup Ew_j}{\text{std\_dev}}\ s(e,Aw_p,Aw_q)} \tag{3}$$

In this direction, SEAT [13] proposes a generalization of WEAT to sentences. SEAT generates sentences generated by inserting words from WEAT's tests into simple templates such as "This is a[n] $word$." SEAT finds evidence of human-like bias in sentence encoders like WEAT does in word encoders. Specifically, a study conducted by [25] quantifies the degree to which gender bias differs with the corpora used for the pre-trained model and fine-tuning with additional data. Schroder et al. [26] proposes various metrics of embedding biases and compares their strengths and weaknesses. Garrido et al. [27] explore biases in embeddings through the lens of geometric spaces, providing a fresh perspective on the subject. Some studies [28,29] utilize embedding testing to investigate the evolution of biases across extensive temporal scales, revealing persistent statistical features throughout history. Furthermore, a recent study [30] reveals extensive potential for embedding testing.

Beyond uncovering biases and stereotypes, embeddings also provide precise assessments of the environments in which biases manifest within human culture and thought.

Although embedding-based approaches are related to IAT in connecting semantic distance and mental association, measured biases are a joint effect of explicit and implicit bias, where neither explicit nor implicit bias is correctly understood. For example, if people dislike flower in implicit bias but claim to like it in explicit bias, in their statements explicitly expressing their attitude toward flowers, concept "*Flower*" will be semantically closer to attribute "*Pleasant*" than "*Unpleasant*". However, in his/her rest sentences, "*Flower*" can be closer to "*Unpleasant*" than "*Pleasant*", which reveals their real implicit bias. Significantly, with the recent emergence of large language models (LLMs) extensively utilized in real-world applications [31,32], it is essential to differentiate between explicit and implicit biases for a thorough assessment of biases present in LLMs.

### *2.3. Automatic Sentiment Analysis*

Lastly, we introduce the automatic sentiment analysis utilized for analyzing sentiment polarity. Sentiment analysis, otherwise known as attitude and opinion mining, refers to the use of natural-language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information [9]. It aims to determine the attitude of a speaker, writer, or other subjects with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. Sentiment analysis involves classifying attitudes in text into categories like "positive", "negative", or "neutral" [33]. Generally, current sentiment analyzers often represent an attitude with the following key factors [34]:

Object: an entity which can be a person, event, product, organization, topic, etc.;

Attribute: a feature of objects with respect to which evaluation is made;

Attitude orientation or polarity: the orientation of an attitude on an object representing whether the attitude is positive, negative, or neutral;

Attitude holder: the person or entity that expresses the attitude.

The definitions of "object" and "attitude" in sentiment analyzers of attitude can match those of "concept" and "attitude" in psychological measurements. Although "attribute" has different definitions in automatic methods and psychological measurements, we do not need to match them in this work because we are not working on aspect-based attitude analysis [35]. It is noted that state-of-the-art sentiment-based attitude measurements also do not clearly distinguish between explicit and implicit attitudes bias [36–38]. In this study, we employ pre-trained machine-learning models [39] to analyze the sentiment conveyed in statements. Pre-trained machine-learning methods refer to the process of initially training a model on a large-scale dataset to acquire knowledge of the underlying language structure, followed by fine-tuning for specific tasks. Specifically, two sentiment analysis models, namely Stanford CoreNLP [40] and Baidu Sentiment Analyzer https://ai.baidu.com/tech/nlp_apply/sentiment_classify (accessed on 12 March 2024), were employed. These two models combine probabilistic models, machine learning, and linguistic knowledge to analyze sentiments of statements.

Although automated sentiment analysis has been widely studied and applied in fields such as public opinion analysis and human-machine dialogue, there has been no research yet on applying sentiment analysis techniques to identify people's biases. This not only requires us to theoretically attempt to correlate the key elements of sentiment analysis with the elements of bias measurement in the field of psychology but also necessitates the design of reasonable automated methods to apply this correlation to real-world data.

### *2.4. Highlight the Issues of Related Works*

Here, we enumerate the key issues that need to be addressed in current related research and are also the main focus of this paper:

(1) The SRA and IAT tests in psychology, which target explicit and implicit biases, require active cooperation from the subjects and are contact-based tests, making it difficult to accurately measure them in large-scale populations.
(2) While the WEAT method based on word vector technology can automatically measure biases in large-scale populations, it is unable to distinguish between explicit and implicit biases.
(3) Although automated sentiment analysis is widely used, it has not been applied to the automated measurement of explicit and implicit biases, and there is also a lack of analysis on the correlation between the two.

## 3. Our Approach

In this section, we will first introduce the high-level idea of the method proposed in this paper, followed by a detailed discussion of the key techniques involved.

### *3.1. Overview of the High-Level Idea of Connecting SRA/IAT and Linguistic Statements*

Automatic measurement of attitude bias plays an essential role in many social computing studies, e.g., customized recommendation, public opinion analysis, influence analysis, and user profiling. Unlike the traditional psychological measurements in the laboratory, these tasks often require attitude measurement on a massive population instead of a limited sample set. Furthermore, different types of attitudes are expected for different applications. For example, in public opinion and influence analysis, people's explicit attitude bias is focused on, which is transferred in the group and influences each other through language. However, in customized recommendation and user profiling, people's implicit attitude bias can be more important, as it determines people's behavior inherently.

In psychology studies, subjects' explicit bias of the concepts can be measured when the subjects are consciously aware of the measured concepts, e.g., being asked questions containing the concepts in SRA. Subjects' implicit bias of the concepts can be measured when the concepts are implicitly primed in subjects' minds, but they are not consciously aware of it, e.g., giving responses to the exemplars instead of concepts in IAT [1].

For this task, both classical psychological measurements and current computational sentiment or semantic measures have certain defects.

First, in classical psychological measurements, although SRA and IAT can measure the explicit and implicit attitude separately with good theoretical validation, they have two main disadvantages for automatic massive measurements.

1. Sample size: Due to SRA and IAT requiring active cooperation from the subjects, the size and distribution of the sample size are highly limited by how many and what kind of subjects can be invited.
2. Independence: In both SRA and IAT, the subjects are aware that they are attending a measurement and may behave differently from their usual behaviors, i.e., the subjects cannot be exactly independent of the measurement.

Second, in current language-based computational measurements, although sentiment and semantic analysis can be applied to massive subjects, the main disadvantage is ignoring the difference between explicit and implicit bias. This will be confused in the case that the explicit and implicit biases are different. For example, if a group of people expresses positive explicit bias toward Black people, but they have a negative implicit bias in their mind, the sentiment or semantic scores on their language about Black people can be neutral in general. This result can mislead both the public influence analysis (explicit bias) and the user preference profiling (implicit bias).

To address the above problems, in this task, we propose to combine the advantages of psychological measurements and language-based computational measurements. Figure 3 illustrates the overview of the high-level idea of connecting SRA/IAT and linguistic statements. We simulate SRA and IAT with the sentiment of the statements containing combinations of concept/exemplar words (e.g., Flower/Rose) and attribute words (e.g., beautiful). In a set of statements expressing the attitude of the exemplars (e.g., Rose is beautiful.), speakers' implicit bias can be understood statistically because there is a chance for every statement that the concepts (e.g., Flower) implicitly influence speakers' attitudes toward the exemplars (e.g., Rose), but the speakers are not consciously aware of it. Furthermore, to connect the criteria of SRA/IAT to the linguistic features, we make the following hypothesis. In the statements expressing the attitude of the concepts, speakers' explicit bias can be understood because the speakers must be consciously aware of the concepts they are discussing. For example, in Figure 3, the statement "Flower is not beautiful at all" must express the explicit bias on the concept "Flower".

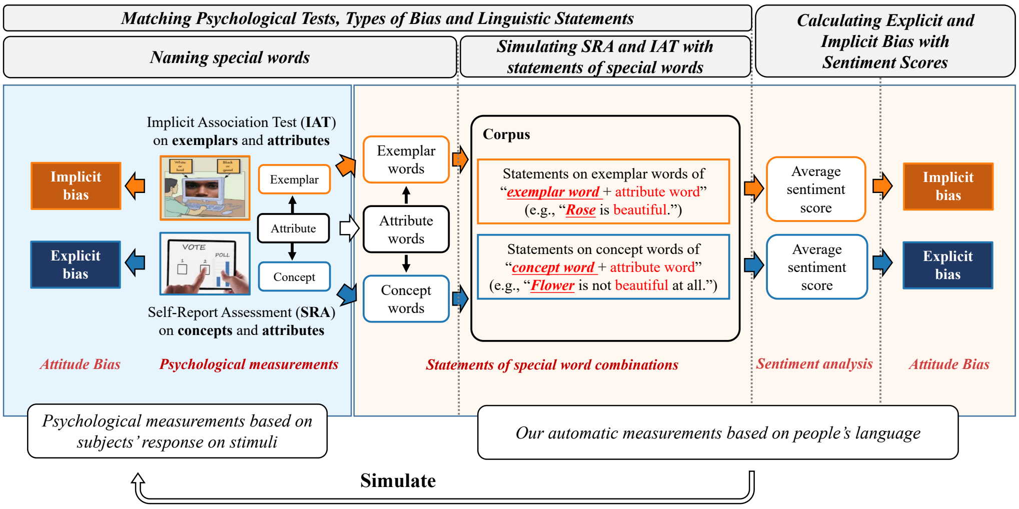

Matching Psychological Tests, Types of Bias and Linguistic Statements
Calculating Explicit and Implicit Bias with Sentiment Scores
Naming special words
Simulating SRA and IAT with statements of special words
Implicit Association Test (IAT) on exemplars and attributes
Implicit bias
Explicit bias
Exemplar
Attribute
Concept
VOTE
POLL
Self-Report Assessment (SRA) on concepts and attributes
Exemplar words
Attribute words
Concept words
Corpus
Statements on exemplar words of "exemplar word + attribute word" (e.g., "Rose is beautiful.")
Statements on concept words of "concept word + attribute word" (e.g., "Flower is not beautiful at all.")
Average sentiment score
Average sentiment score
Implicit bias
Explicit bias
Attitude Bias
Psychological measurements
Statements of special word combinations
Sentiment analysis
Attitude Bias
Psychological measurements based on subjects' response on stimuli
Our automatic measurements based on people's language
Simulate

**Figure 3.** Overview of the high-level idea of connecting SRA/IAT and linguistic statements.

However, in the statements expressing the attitude of the exemplars, speakers' implicit bias may be understood because there is a chance that the concepts implicitly influence speakers' attitudes toward the exemplars, but the speakers are not consciously aware of it. For example, in Figure 3, the statement "Rose is beautiful at all" may express the implicit bias on the concept "Flower". Although this implicit influence is not guaranteed in each statement of exemplars, it is still likely to understand this concept's implicit bias if we have many statements containing the exemplars of a concept.

Therefore, our high-level idea is to simulate the criteria of SRA or IAT by measuring the sentiment of people's statements on concepts or exemplars w.r.t attributes, thus distinguishing two kinds of statements involving explicit and implicit bias, respectively. In this way, we can understand people's explicit and implicit biases by calculating the sentiment scores of each type of statement. This simple idea has three advantages:

(1) There is no need to modify state-of-the-art sentiment analyzers.
(2) The massive population statements are easy to obtain from online corpora. Furthermore, by narrowing the properties of the subjects or the corpus, it is possible to control the properties of the samples. For example, it is possible to find the bias of the female subjects or find people's bias in formal scenarios (from serious text like Wikipedia or news) or informal scenarios (from social media)
(3) The proposed test is strictly independent of the data. The measurement of the bias in language will not impact the generation of the bias in language, which is a good feature for the measurement of psychological characteristics (neither SRA nor IAT exactly satisfies this independence).

### *3.2. Matching Psychological Tests, Types of Bias and Linguistic Statements*

As shown in the left half of Figure 3 (the part covered by the indicator bar labeled "Matching Psychological Tests, Types of Bias and Linguistic Statements"), we propose two stages to simulate SRA/IAT linguistically: 'naming special words' and 'simulating SRA/IAT with statements of special words'.

#### 3.2.1. Naming Special Words

As shown in the left half of Figure 3 (the part covered by the indicator bar labeled "Naming special words"), we define the special words of our approach. In SRA and IAT, there are three key factors: Concept, Exemplar and Attribute. We define three kinds of words naming these three key factors:

(1) Concept words: concept words are the word naming concept, e.g., the concept words of concept "Flower" can be flower, flowers, . . . }.
(2) Exemplar words: exemplar words are the words naming the exemplars of a concept, e.g., the exemplars words of concept "Flower" can be {aster, clover, hyacinth, . . . }.
(3) Attribute words: attribute words are the words naming the connotation of an attribute, e.g., the attribute words of attribute "Pleasant" can be {happy, good, beautiful, . . . }.

#### 3.2.2. Simulating SRA and IAT with Statements of Special Words

Then, as shown in the left half of Figure 3 (the part covered by the indicator bar labeled "Simulating SRA and IAT with statements of special words"), we simulate SRA and IAT with statements containing the combinations of special words.

(1) Explicit bias and statements of "concept word + attribute word": in SRA, subjects' biased expression on a concept with respect to an attribute is regarded as explicit bias on the concept. In this work, corresponding linguistic criteria is: if a statement mentions a concept word with respect to an attribute word, the sentiment of this statement is regarded as a case for explicit bias on the concept, e.g., the sentiment of "Flower is not beautiful" is a case of negative explicit bias on "Flower".
(2) Implicit bias and statements of "exemplar word + attribute word": in IAT, the subject's biased response on an exemplar with respect to an attribute is regarded as implicit bias on the concept involving the exemplar. In this work, the corresponding linguistic criterion is: if a statement mentions an exemplar word with respect to an attribute word, the sentiment of this statement is regarded as a case of implicit bias on the concept involving the exemplar. For example, the sentiment of "Rose is beautiful" is a case of positive implicit bias on "Flower".

It is noted that we cannot only use the sentiment polarity of attribute words to determine the sentiment of statements. For example, in a statement, "Rose is not beautiful at all.", though the attribute word "beautiful" is positive in sentiment, the statement is negative in the implicit bias on "Flower".

In practice, the statements containing the "*concept/exemplar word + attribute word*" are identified according to grammatical dependency between the concept/exemplar word and attribute word. In particular, a statement is selected if and only if it is a "*subject-verb-object*" or "*subject-link verb-predicative*" where a concept/exemplar word and an attribute word are both involved.

### *3.3. Calculating Explicit and Implicit Bias with Sentiment Scores*

As shown in the right half of Figure 3 (the part covered by the indicator bar labeled "Calculating Explicit and Implicit Bias with Sentiment Scores"), matching the two categories of statements with SRA and IAT, we separately calculate the sentiment scores of the collection of each category of statements to measure the explicit and implicit bias.

Given a pair of concepts $C_i$ vs. $C_j$ , $Cw_i$ and $Cw_j$ are the concept word sets of $C_i$ and $C_j$, respectively. $Ew_i$ and $Ew_j$ are the exemplar words sets of $C_i$ and $C_j$, respectively. $Aw_p$ and $Aw_q$ are the attribute words sets of a pair of attribute $A_p$ vs. $A_q$, respectively. Given statements set $S$, the following steps calculate the explicit and implicit bias between $C_i$ and $C_j$ of the people who generate $S$.

$S_{Cw\_i}, S_{Cw\_j}, S_{Ew\_i}, S_{Ew\_j} \subseteq S$ are four statements collections in which each statement simultaneously contains one word from $Cw_i, Cw_j, Ew_i, Ew_j$ (concept or exemplar words), and one word from $Aw_p \cup Aw_q$ (attribute words), respectively. $S_{Cw\_i}$, $S_{Cw\_j}$, $S_{Ew\_i}$ and $S_{Ew\_j}$ are used to identify the explicit bias on $C_i$, the explicit bias on $C_j$, the implicit bias on $C_i$ and the implicit bias on $C_j$, respectively. Equations (4) and (5) calculate the explicit and implicit bias on $C_i$ vs. $C_j$ with respect to $A_p$ vs. $A_q$, respectively, where $p(s)$ is the sentiment score of statement $s$.

$$ExplicitBias(C_i, C_j, A_p, A_q) = \underset{s \in S_{Cw\_j}}{\text{mean}}\; p(s) - \underset{s \in S_{Cw\_i}}{\text{mean}}\; p(s) \tag{4}$$

$$ImplicitBias(C_i, C_j, A_p, A_q) = \underset{s \in S_{Ew\_j}}{\text{mean}}\; p(s) - \underset{s \in S_{Ew\_i}}{\text{mean}}\; p(s) \tag{5}$$

## 4. Experiments

In this section, we will present the experiments and analysis conducted in this paper, including the grouping and objectives of the experiments, the experimental design, the results obtained, and a discussion of the findings.

### *4.1. Aims of Experiments*

In the experiments, we investigated two questions:

The first question is: applying our method on corpora, can we reproduce well-known psychological observations (ground truth) of explicit and implicit bias?

The second question is: with our method, can we have new findings besides the well-known observations?

Then, we had three groups of experiments:

The first experiment is: on various online corpora, comparing the experimental observations of our proposed method with those reported by SRA and IAT (Section 4.3).

The second experiment is: examining the effects of the selection of special words and sentiment analyzers (Section 4.4). The concept of sentiment analyzers is introduced in Section 2.3.

The third experiment is: comparing the evolution of explicit and implicit bias over time (Section 4.5).

### *4.2. Global Settings*

We introduce the basic settings used in our experiments, including the selection of different corpora, the target words of the study, the sentiment analysis tools utilized, and the evaluation methods. These settings allow us to systematically investigate and compare explicit and implicit biases across different languages, media, and cultural contexts.

#### 4.2.1. Corpora

Four corpora of different languages and media were explored. For serious corpora, English Wiki (99 M sentences) and Chinese Wiki (7 M sentences); for social media corpora, Twitter (100 M English sentences) and Weibo (77 M Chinese sentences). For each concept and each corpus, we selected two subsets of two kinds of statements defined in Section 3.2.2, i.e., (1) subsets of the statements of "*concept word* + *attribute word*"; (2) subsets of the statements of "*exemplar word* + *attribute word*". Stanford Dependency Parser [41] was used to identify the "*subject-verb-object*" or "*subject-link verb-predicative*" pattern in statements.

#### 4.2.2. Concept Pairs and Special Words

In experiments, people's bias on four concept pairs was investigated. To compare our results to well-known psychological observations, four concept pairs reported in IAT were selected including "$\underline{\textit{Insect}}$ vs. $\underline{\textit{Flower}}$", "$\underline{\textit{Weapon}}$ vs. $\underline{\textit{Instrument}}$", "$\underline{\textit{Afri-American}}$ vs. $\underline{\textit{Euro-American}}$" and "$\underline{\textit{China}}$ vs. $\underline{\textit{America}}$". The selected concepts involve general social bias, controversial social bias, and cultural differences between China and America to make the experiments more representative.

In English corpora, concept words were selected to name the concept directly. The exemplar words and attribute (*pleasant* and *unpleasant*) words were selected following IAT [5], WEAT [12], and Chinese version of Harvard's online IAT https://implicit.harvard.

edu/implicit/china/ (accessed on 12 March 2024). We manually translated the special words used in English experiments into Chinese in Chinese experiments.

The detailed list of selected special words for each pair of concepts is as follows:

**(1) Concept pair: Insect vs. Flower**

**Concept words of 'Insect':** insect, insects

**Concept words of 'Flower':** flower, flowers

**Exemplar words of 'Insect':** ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil

**Exemplar words of 'Flower':** aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia

**Attribute words of 'Pleasant':** caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

**Attribute words of 'Unpleasant':** abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

**(2) Concept pairs: Weapon vs. Instrument**

**Concept words of 'Weapon':** weapon, weapons

**Concept words of 'Instrument':** instrument, instruments

**Exemplar words of 'Weapon':** arrow, club, gun, missile, spear, axe, dagger, harpoon, pistol, sword, blade, dynamite, hatchet, rifle, tank, bomb, firearm, knife, shotgun, teargas, cannon, grenade, mace, slingshot, whip

**Exemplar words of 'Instrument':** bagpipe, cello, guitar, lute, trombone, banjo, clarinet, harmonica, mandolin, trumpet, bassoon, drum, harp, oboe, tuba, bell, fiddle, harpsichord, piano, viola, bongo, flute, horn, saxophone, violin

**Attribute words of 'Pleasant':** caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

**Attribute words of 'Unpleasant':** abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

(3) Concept pair: African-American vs. European-American

**Concept words of 'African-American':** african_americans, african_american, black_people, African_American, Black_people, African_Americans

**Concept words of 'European-American':** european_americans, european_american, white_people, European_American, White_people, European_Americans

**Exemplar words of 'African-American':** Alonzo, Jamel, Theo, Alphonse, Jerome, Leroy, Torrance, Darnell, Lamar, Lionel, Tyree, Deion, Lamont, Malik, Terrence, Tyrone, Lavon, Marcellus, Wardell, Nichelle, Shereen, Ebony, Latisha, Shaniqua, Jasmine, Tanisha, Tia, Lakisha, Latoya, Yolanda, Malika, Yvette

**Exemplar words of 'European-American':** Adam, Harry, Josh, Roger, Alan, Frank, Justin, Ryan, Andrew, Jack, Matthew, Stephen, Brad, Greg, Paul, Jonathan, Peter, Amanda, Courtney, Heather, Melanie, Katie, Betsy, Kristin, Nancy, Stephanie, Ellen, Lauren, Colleen, Emily, Megan, Rachel

**Attribute words of 'Pleasant':** caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

**Attribute words of 'Unpleasant':** abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

**(4) Concept pair: China vs. America**

**Concept words of 'China':** China, PRC, china

**Concept words of 'America':** America, USA, US

**Exemplar words of 'China':** Chinese, CNY, MaoZedong, socialist, Beijing

**Exemplar words of 'America':** American, USD, Washington, capitalism, New York

**Attribute words of 'Pleasant':** caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

**Attribute words of 'Unpleasant':** abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

Correspondingly, in Chinese experiments, we manually translate the special words used in English experiments into Chinese. Finally, according to concept pairs and special words, the counts of collected statements are in Table 1, wherein each NumA/NumB, NumA counts explicit statements containing "concept word + attribute word," NumB counts implicit statements containing "exemplar word + attribute word."

**Table 1.** Count of explicit and implicit statements in experiments.

| | Concept Pairs | | | |
|---|---|---|---|---|
| **Corpora** | **Insect vs. Flower** | **Weapon vs. Instrument** | **Afri-A vs. Euro-A** | **China vs. America** |
| Eng-Wiki | 152k/395k | 273k/136k | 97k/1868k | 838k/202k |
| Twitter | 79k/171k | 60k/391k | 36k/113k | 613k/468k |
| Chn-Wiki | 10k/17k | 21k/162k | 20k/505k | 380k/91k |
| Weibo | 243k/240k | 30k/562k | 76k/664k | 1674k/694k |

### 4.2.3. Sentiment Analyzers

In experiments, sentiment scores for English and Chinese statements were obtained using two sentiment analysis tools, referred to as Stanford CoreNLP and Baidu Sentiment Analyzer, as mentioned in Section 2.3. The reason for choosing these two sentiment analyzers is that they are among the best-performing sentiment analyzers in English and Chinese, respectively. To verify the performance and compatibility of these two analyzers, we constructed 100 Chinese-English test statement pairs (listed in Appendix A.2), where each pair consists of semantically identical English and Chinese statements containing both concept/exemplar words and attribute words. We analyzed the sentiment polarity of English and Chinese statements in each pair using CoreNLP (https://stanfordnlp.github.io/CoreNLP/, accessed on 12 March 2024) and Baidu Analyzer (https://intl.cloud.baidu.com/product/body.html, accessed on 12 March 2024), respectively. The results showed that CoreNLP achieved a 94% accuracy rate on 100 English statements, while Baidu Analyzer achieved a 97% accuracy rate on 100 Chinese statements, with a 96% matching degree between the two analyzers. This indicates good consistency between their sentiment analysis results.

To examine the effects of different sentiment analyzers, in the testing of effects of procedural variables, we chose two comparative analyzers: NLTK sentiment package for English and Tencent Sentiment Analyzer https://ai.qq.com/doc/nlpemo.shtml (accessed on 12 March 2024) for Chinese.

### 4.2.4. Evaluation Method

To the best of our knowledge, our work is the first to automatically distinguish explicit and implicit bias in corpora. We build a high-level and intuitive connection between SRA/IAT and linguistic statements. Although the validity of such a connection cannot be theoretically guaranteed, we propose to use well-known psychological observations [5] as ground truth and investigate whether the proposed method can reproduce these observations qualitatively. The observations will be compared in Section 4.3. In Section 4.3, we employed sentiment scores as evaluation metrics to compare explicit and implicit biases across different datasets and concept pairs, revealing disparities between explicit and implicit biases under various conditions. Furthermore, in Section 4.5, we utilized the temporal changes in sentiment scores to compare the trends of explicit and implicit biases across different datasets and concept pairs, highlighting the differences in their dynamic evolutionary patterns, which have not been explored in traditional psychological research.

### *4.3. Comparing Explicit and Implicit Bias*

**Results**: In the first group of experiments, we investigated the performance of our method of distinguishing explicit and implicit bias. We calculated the explicit and implicit bias separately with our method on each corpus for each concept pair. The significance of the difference between explicit and implicit bias pairs was tested with a permutation test using no difference as the null hypothesis. The results are shown in Figure 4, where the explicit and implicit bias of each concept pair on each corpus is compared using sentiment scores calculated with Equations (4) and (5). Positive and negative values indicate the bias preferring the right and left concept of this column, respectively (e.g., in the first column, Insect is the left concept, Flower is the right concept). The red star indicates that the difference between this pair of explicit/implicit biases is significant ($p$ = 0.0001). "Afri-A" = "Afri-American", "Euro-A" = "Euro-American". For example, in English Wiki, people prefer to Afri-American in explicit bias and prefer to Euro-American in implicit bias.

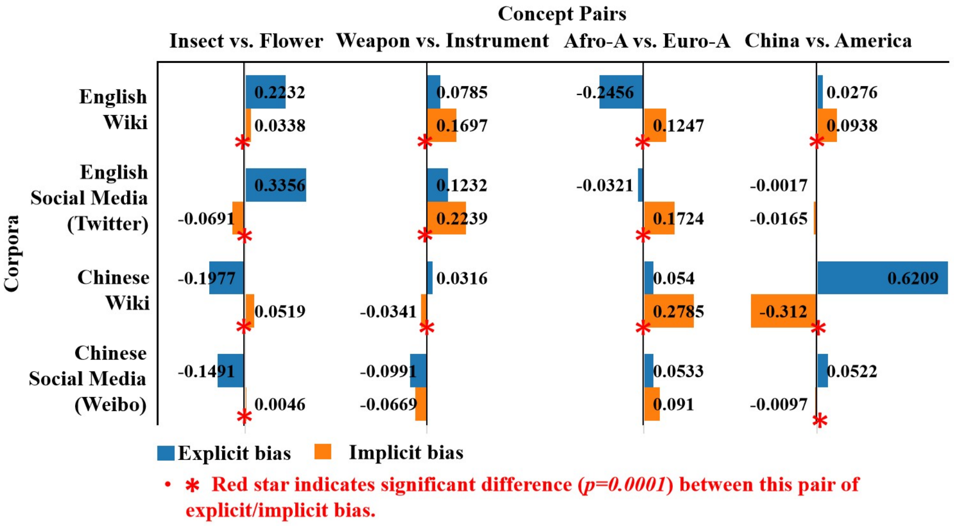


**Figure 4.** Compare explicit and implicit bias of each concept pair on each corpus.

**Discussion**: In Figure 4, we have four main observations:

(1) **In general:** implicit bias can be significantly different from explicit bias, which agrees with the IAT observation. The difference is sensitive to the sociality of concepts, scenarios, and cultures, which agrees with the psychological assumptions of implicit cognition: the difference between explicit and implicit bias can be explained by the social desire [5].

(2) **On concepts:** in most cases, the difference between implicit and explicit bias is more significant on social and controversial concept pairs (i.e., "Afri-American vs. Euro-American" and "China vs. America") than the difference on unsocial concept pairs (i.e., "Insect vs. Flower" and "Weapon vs. Instrument"). This observation agrees with the psychological assumption that for social and controversial concepts, people are more motivated to adjust their explicit bias to adapt to social desire, which leads to greater differences between implicit and explicit bias.

(3) **On scenario:** for controversial concept pairs, implicit and explicit biases are closer to each other in social media than in Wikipedia. From social desire, users are more strongly required to conform to mainstream social values in Wikipedia than in social media.

(4) **On language (culture):** though on both two English corpora, explicit and implicit bias on "*Afri-American* vs. *Euro-American*" have opposite polarity (consistent with IAT reports), but this does not happen in Chinese corpora. Similarly, explicit and implicit bias on "*China* vs. *America*" are opposite on Chinese corpora instead of on English corpora. From social desire, the importance of consistency with mainstream social values can vary across countries or cultures for a concept pair.

From the view of the relationship with previous work, our experimental results line with psychological observations in general. Our innovation lies in the automation of explicit and implicit bias analysis by correlating concepts from psychological measurements with elements of automated sentiment analysis. Furthermore, we addressed the shortcomings of the previous word vector-based automated analysis method, which was unable to distinguish between explicit and implicit biases. The limitation of our approach is that further explanation of the difference in specific concept pairs still needs help from psychology scientists in future work. This explanation might achieve new findings considering the absence of observations on a massive population in the current psychological study on implicit bias.

### *4.4. Effects of Procedural Variables*

Besides $p$-values of the difference between explicit and implicit bias, we also tested the significance of the qualitative observations in Section 4.3 by testing the effects of procedural variables. The proposed method has two procedural variables: special words and sentiment analyzers' selection, and this section investigates whether these two variables affect the experimental observations.

Table 2 illustrates the percentage that experimental observations are maintained across procedural variables, where 'Words' is special words, 'Analyzer' is sentiment analyzers, 'E-' is English, and 'C-' is Chinese.

For the effect of exemplar word selection, we randomly selected a subset from the original exemplar word set and repeated the experiments. We repeated the experiment in Section 4.3 1000 times and calculated the percentage of the qualitatively consistent observations in Section 4.3.

For the effect of sentiment analyzer choice, we randomly selected a subset from the original exemplar words set, repeated the experiments on the subset with comparative analyzers indicated in Section 4.2.3, and compared the original analyzers' results. We repeated the test 1000 times and calculated the percentage that the results of comparative and original analyzers were consistent with each other.

In Table 2, the observations drawn in Section 4.3 are maintained at high probability across different exemplar words and sentiment analyzers. The concept of sentiment analyzers is introduced in Section 2.3.

**Table 2.** The percentage that experimental observations are maintained across procedural variables.

| Variable | Twitter | E-Wiki | Weibo | C-Wiki |
|---|---|---|---|---|
| Words | 97.8% | 99.8% | 96.4% | 98.3% |
| Analyzers | 96.7% | 90.4% | 92.8% | 91.2% |

### *4.5. Evolution of Explicit and Implicit Bias*

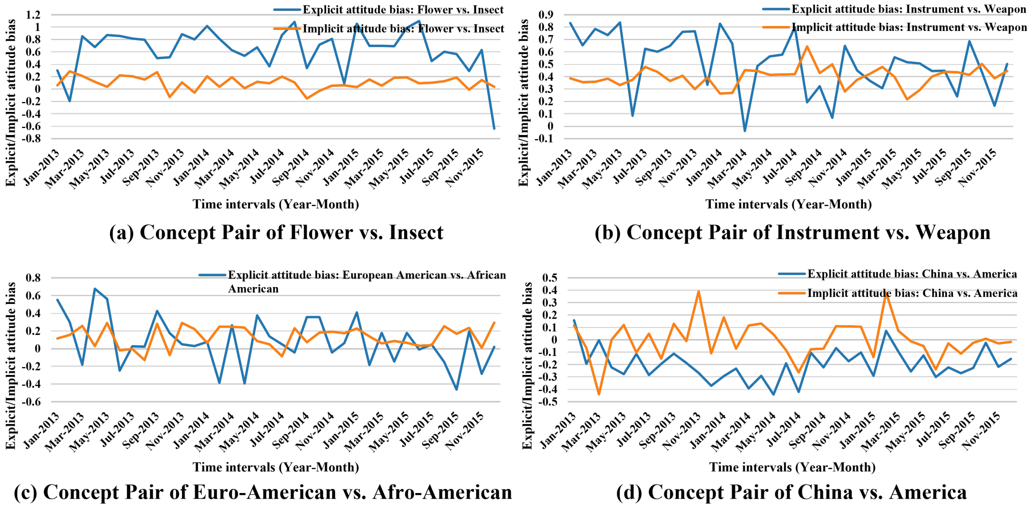
**Results**: In this group of experiments, we compare the evolution of explicit and implicit bias over time. We use English and Chinese social media corpora (i.e., Twitter and Weibo), with a timestamp for each statement. To track the evolution of the biases, we broke each corpus into subsets of months, spreading over 36 months. Then, we measured explicit and implicit bias month by month using the proposed sentiment-based bias measurement. In Figures 5 and 6, we illustrate the detailed evolution tracks of each concept pair on each corpus. In Figure 7, we compare the stability of explicit and implicit bias in evolution. The stability is calculated with the standard deviation of the bias over time.

**Discussion**: In Figures 5–7, we have two main observations:

(1) **Explicit and implicit bias is different in stability:** beyond existing psychological studies, a new observation of evolution experiments is that implicit biases are more stable than explicit biases in most cases. In Figures 5–7, with the exception of the concept pair "China vs. America", implicit biases in other concept pairs are more stable than explicit biases. This may indicate that although people's explicit public bias may vary due to the changing social context (e.g., the scenario or the talking partner), their internal implicit bias tends to be more consistent over time. This possibly explains the finding that implicit bias is a better predictor of certain kinds of behavior than explicit bias [6,7].

(2) **Factors influencing the stability:** in this experiment, the sociality of concepts also influences the evolution of biases. In Figures 5 and 6, implicit biases of unsocial concept pairs (i.e., "<u>*Insect*</u> vs. <u>*Flower*</u>" and "<u>*Weapon*</u> vs. <u>*Instrument*</u>") are more stable than that of social concept pairs (i.e., "<u>*Afri-American*</u> vs. <u>*Euro-American*</u>" and "<u>*China*</u> vs. <u>*America*</u>").

In terms of its relationship with previous work, the dynamic analysis presented in this paper builds upon the existing concepts of psychological bias measurement. However, its novelty lies in the fact that existing psychological measurements primarily focus on static analysis, whereas our study delves into the dynamic evolution and comparison of explicit and implicit biases. While our work has successfully revealed the differences in the evolution of these two types of biases, its limitation lies in the fact that we have not yet empirically and deeply analyzed the impact of these differences on behavior. This aspect remains an area for further exploration in future research.



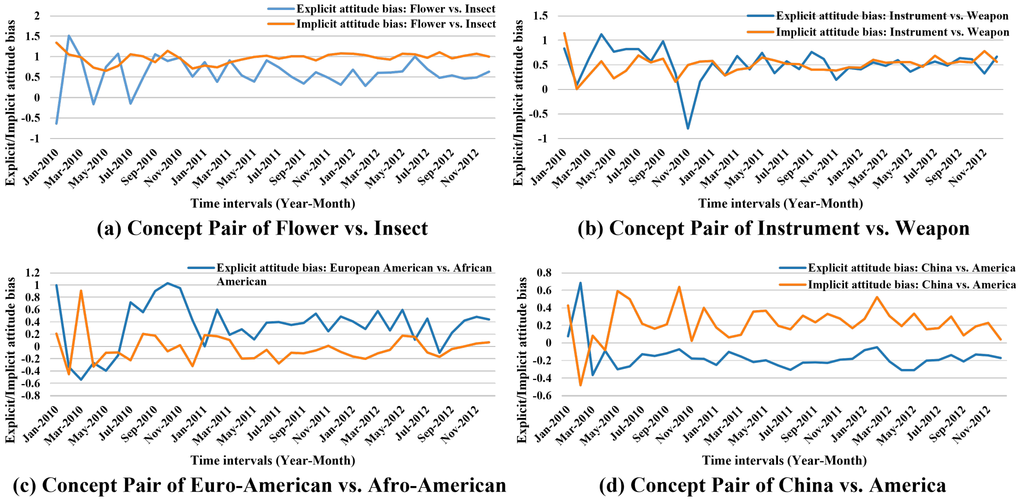
(a) Concept Pair of Flower vs. Insect

(b) Concept Pair of Instrument vs. Weapon

(c) Concept Pair of Euro-American vs. Afro-American

(d) Concept Pair of China vs. America

**Figure 5.** The evolution of explicit and implicit bias on Twitter.



(a) Concept Pair of Flower vs. Insect

(b) Concept Pair of Instrument vs. Weapon

(c) Concept Pair of Euro-American vs. Afro-American

(d) Concept Pair of China vs. America

**Figure 6.** The evolution of explicit and implicit bias on Weibo.

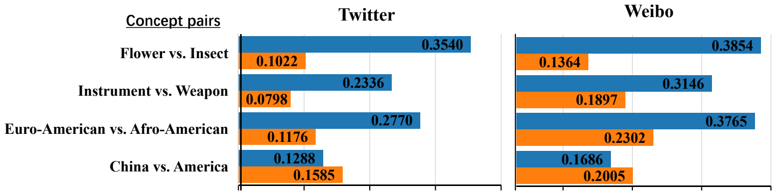


**Figure 7.** The stability (standard deviation over time) of explicit and implicit bias in social media. The blue and orange bars illustrate the standard deviation of explicit and implicit bias, respectively.

### 4.6. Comparing WEAT and SEAT with Proposed Sentiment-Based Method

As introduced in Section 2.2, WEAT [12] and SEAT [13] are a joint effect of explicit and implicit bias. The factors of our method are more naturally connected with those of SRA and IAT, and it can distinguish explicit and implicit bias. Experimental results support the effectiveness of this connection. In the case that explicit and implicit bias might be quite different, and the difference is seriously concerned with the task, our method can be a better choice for bias identification. Furthermore, due to the requirement of massive data in high-quality training of word or sentence embedding, WEAT and SEAT do not work well on small data (e.g., the language of an individual). However, our method derives bias cases from every single statement and is not as sensitive to the data size as WEAT does.

### 4.7. The Application of This Work in Eliminating Bias in Generative AI

Bias is also a key issue in current research on generative AI. There is research indicating that the content generated by generative AI may contain bias [42–44]. This bias is considered inappropriate and may affect the user's cognition. Although there have been some attempts to identify and eliminate biases in generative AI, an important drawback of current efforts is the lack of clear differentiation between explicit and implicit biases. This drawback makes it difficult to accurately evaluate the cognitive characteristics of bias in generative AI, therefore reducing the effectiveness of bias elimination, especially implicit bias elimination. Our latest research shows that biases in generative AI can also be distinguished as explicit and implicit [45]. Therefore, the work of this article will help automate the identification and differentiation of explicit and implicit bias in generative AI, thus more effectively guiding research on bias elimination in generative AI.

## 5. Conclusions and Future Work

### 5.1. Summary

Using automatically obtained sentiment scores of statements containing the combination of special words on various corpora, we automatically reproduce psychological measurements of explicit and implicit attitude bias in a large population. We examine the special words of four pairs of concepts and the statements from four large corpora, including English and Chinese Wikipedia and social media. The experimental results reveal that our automatic method qualitatively agrees with the costly small-scale observations from psychological studies. Specifically, explicit and implicit attitude bias can be significantly different. The significance of the difference is shown to be influenced by the sociality of concepts, the scenario of language use, and the culture behind the language. We also find that implicit attitude biases are more stable than explicit attitude biases over time, which is absent in existing psychological studies. The theoretical significance of our work is to provide a new perspective based on sentiment analysis for the concept of implicit association in traditional psychological research. The technical significance is in proposing an automatic measurement method that can replace high-cost traditional psychological measurements. The practical significance is in providing new research tools for explicit and implicit cognition targeting larger-scale populations.

### *5.2. Future Work*

In future work, there are two main directions to further improve this work. The first direction is to examine the conclusions of this work with more datasets. Although the corpora used in this work cover two main languages (English and Chinese) and the two most common forms of Internet text (serious text from Wikipedia and free text from social media), the experimental results of data from more different languages and sources are expected to provide more evidence for the conclusions of this work and help to understand the explicit and implicit bias of people from different cultural backgrounds. Another important future work is to design algorithms to automatically discover the concepts on which people have bias instead of selecting the concepts manually. Our work in this paper manually adopts the specified concepts to facilitate comparison with existing research in psychology to verify our method. Once the effectiveness of our method is verified, in practical applications, we need to be able to automatically discover more biased concepts that have not been addressed in existing research by improving our method.

**Author Contributions:** Conceptualization, B.W.; Methodology, B.W., R.Z. and B.X.; Software, B.X.; Validation, B.X.; Writing–original draft preparation, Y.Z.; Writing–review and editing, L.Y. and H.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Key R&D Program of China(2022YFC3301900; 2022YFC3301901), National Natural Science Foundation of China (62376188, 62272340, 62276187, 62376192).

**Data Availability Statement:** The original data presented in this study are openly available for Chinese Wikipedia at https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2 (accessed on 12 March 2024) and for English Wikipedia at https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2 (accessed on 12 March 2024). Due to the large scale of the datasets collected from Weibo and Twitter, interested readers may request access to them by contacting the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A

### *Appendix A.1. Definitions*

Attitude Bias: attitude is one's evaluation of a concept [46]. Attitude bias (abbreviated as "bias"), in psychology, indicates one's preference in his attitudes toward two comparative concepts [47].

Explicit bias: explicit bias is the bias that one deliberately expresses [48]. For example, if one claims to prefer African Americans to European Americans in SRA, he/she has an explicit bias against African Americans compared to European Americans [8].

Implicit bias: implicit bias is one's internal bias without conscious awareness or honestly reporting [47]. An individual's consciousness does not control implicit bias. The primary purpose of measuring implicit bias is to predict behaviors that explicitly held biases cannot predict [49]. For example, although one deliberately claims to prefer African Americans to European Americans, he/she may associate African Americans with negativity in implicit cognition [5].

Concept: concepts name the objects on which people have bias [47]. As psychological tests do, we pair concepts to understand bias, e.g., Afri-American vs. Euro-American.

Exemplar: exemplars are stimuli of concepts. Exemplars should be objects which can be clearly classified into one concept in a concepts pair, e.g., typical African-American names or photos can be the exemplars of concept "Afri-American" w.r.t the concepts pair "Afri-American vs. Euro-American" [47].

Attribute: attribute is the connotation of bias. For example, in IAT, attributes of positive and negative bias can be "Pleasant" and "Unpleasant", respectively [12,47].

Self-Report Assessment (SRA): Self-report assessment is a typical explicit bias method. Reading questions and choosing responses in the assessment is the most direct method of asking about one's feelings, prejudices, beliefs, and other information [17]. SRA can be a survey, questionnaire, or poll in which subjects express their attitude explicitly. The main strength of SRA is allowing participants to describe their own experiences rather than inferring this from observing them.

Implicit Association Test (IAT): IAT [5] is the most widely used psychological test of implicit bias. IAT is designed to measure the strength of subjects' implicit associations between a pair of concepts (e.g., Flower vs. Insect) and a pair of attributes (e.g., Pleasant vs. Unpleasant).

### *Appendix A.2. Sentence Pairs for Comparing the Two Sentiment Analyzers*

In this section, we enumerate all sentences employed to compare the consistency of sentiment analyzers between Stanford CoreNLP and Baidu Sentiment Analyzer. The sentences comprise two groups, positive and negative, with 50 pairs in each group. Each pair includes an English sentence and the corresponding semantically equivalent Chinese sentence.

The positive sentiment sentences are enumerated as follows:

The gentle melody of the violin brings peace to my soul; 小提琴的温柔旋律让我的灵魂感到平静.

Roses add beauty and love to every garden they grace; 玫瑰花为每一个花园增添了美丽与爱.

The lilac's fragrance is a miracle of nature, evoking pure happiness; 丁香花的香气是大自然的奇迹, 唤起纯粹的幸福.

A family picnic in the park, surrounded by blooming daisies, feels like heaven; 在公园里举行家庭野餐, 周围布满开放的雏菊, 感觉如同天堂.

The laughter at the concert was a testament to the harmonica's cheerful sound; 音乐会上的笑声证明了口琴的愉快声音.

A rainbow appeared over the field just as we spotted a beautiful bluebell; 当我们看到美丽的风铃草时, 天空中出现了一道彩虹.

The honest lyricism of the guitar struck a chord with everyone in the room; 吉他的诚实歌词与在场的每个人产生了共鸣.

The breeze through the magnolia flowers gently caressed me, which was really peaceful; 穿过木兰花的微风轻抚, 真是太平静了.

Playing the piano at sunrise is a magical experience, filled with pleasure; 在日出时弹奏钢琴是一种神奇的体验, 充满了愉悦.

Learning to play the flute has been a journey filled with joy and friendship; 学习吹长笛的过程充满了欢乐和友谊.

The sight of a butterfly on a zinnia is a small, everyday miracle; 蝴蝶停留在紫罗兰上, 这是一个微小的日常奇迹.

The love song played on the saxophone was pure romance; 萨克斯管演奏的恋歌纯粹是浪漫.

Spending a vacation surrounded by wildflowers is truly a slice of heaven on earth; 在野花丛中度假是人间天堂.

The health benefits of walking among orchids are as pleasant as their beauty; 在兰花中散步的健康益处与它们的美丽一样令人愉悦.

The tulip festival was a rainbow of colors and a feast for the senses; 郁金香节是色彩的彩虹, 是感官的盛宴.

The early-blooming crocuses, fortunately, discovered the joy that heralds the arrival of spring; 早开的番红花幸运地发现, 预示着春天的欢乐.

The harp's melody at the wedding was a gift of heavenly sound; 婚礼上的竖琴旋律是天籁之音.

A loyal bee diligently pollinates the flowers, ensuring their beauty multiplies; 忠诚的蜜蜂勤奋地给花朵授粉，确保它们的美丽繁衍.

The family gathers annually to enjoy the cherry blossoms, a tradition of happiness; 家人每年聚集一起欣赏樱花, 这是幸福的传统.

A gentle dragonfly rests on a lily, a peaceful scene in nature's playground; 一只温柔的蜻蜓停在百合花上, 这是大自然游乐场中的一个和平场景.

The peony garden was a place of peace and gentle beauty; 牡丹花园是和平与温柔美丽的地方.

The trumpet's blast at the parade spread cheer among the crowd; 游行中小号的声音在人群中传播欢乐.

The pleasure of watching hummingbirds dart among the azaleas is unmatched; 观看蜂鸟在杜鹃花间穿梭的乐趣无与伦比.

The violinist's performance was a diamond sparkling in the evening; 小提琴手的表演如同晚上的一颗闪耀的钻石.

The laughter of children chasing butterflies is pure happiness; 孩子们追逐蝴蝶的笑声是纯粹的幸福.

A rose in full bloom is a symbol of love and natural beauty; 盛开的玫瑰是自然美和爱的象征.

The clarinet's soft notes added a layer of peace to the quiet afternoon; 单簧管的柔和音符为安静的午后增添了一层平和.

The paradise of a spring garden is enhanced by the scent of hyacinths; 春日花园的天堂由风信子的香味增强.

The loyal old dog peacefully lies among the buttercups in the sunshine; 忠诚的年老的狗在阳光中平和地躺在金凤花中.

The enchanting melodies of the mandolin near the campfire create a stronger sense of camaraderie among friends; 营火旁的曼陀林声音使朋友们更亲近.

The miracle of nature is evident in every unfolding daffodil; 每一朵绽放的水仙花都显现自然的奇迹.

The marigold's bright colors bring happiness to any sunny spot; 金盏花的鲜艳色彩为任何阳光充足的地方带来快乐.

The lucky discovery of a rare orchid made the hike unforgettable; 偶然发现一朵罕见的兰花使徒步之旅难忘.

The pleasure of a trombone's slide brings jazz to life in the city; 滑音大号的声音让城市的爵士乐生动起来.

A sunny vacation day is best spent surrounded by blooming violets; 阳光明媚的假日最适合被盛开的紫罗兰环绕.

The laughter of friends around the drum circle creates a cheerful atmosphere; 鼓圈周围朋友的笑声营造出愉快的氛围.

The diploma hung proudly above the piano, a symbol of achievement and music; 毕业证书骄傲地挂在钢琴上方, 是成就与音乐的象征.

The gentle breeze brushing against the iris bed feels soothing; 微风轻抚鸢尾花床, 感觉很舒缓.

A peace lily brightens the office with its gentle presence; 和平百合在办公室以其温柔的存在点亮了环境.

Playing the harpsichord is like touching a piece of history with a pleasant tune; 弹奏拨弦键琴就像触摸历史的一部分, 带来愉快的旋律.

The healthy sight of the garden, with bees and butterflies fluttering around, is a delightful picture; 花园的健康景象, 有蜜蜂和蝴蝶飞舞, 是一幅令人欣喜的画面.

The cheer of the crowd as the band struck up was infectious; 乐队开始演奏时, 人群的欢呼声具有感染力.

A rainbow over a field of carnations is a sight to treasure; 在康乃馨田野上空出现的彩虹是一幕值得珍藏的景象.

The honest simplicity of a sunflower brings a smile to anyone's face; 向日葵的质朴诚实让每个人都会微笑.

The gift of a first guitar can inspire a lifetime of music and happiness; 收到第一把吉他的礼物可以激发一生的音乐和幸福.

The lucky charm of finding a four-leaf clover among the clovers is delightful; 在三叶草中幸运地找到四叶草是令人愉快的幸运符.

The experience of stumbling upon paradise in a vibrant poppy field is nearly surreal; 在盛开的罂粟花田中找到天堂几乎是超现实的.

A family outing to a concert is a vacation filled with joyful sounds; 全家去参加音乐会就像是在享受充满欢乐声音的假期.

The miracle of a bee pollinating a gladiola is nature at its finest; 蜜蜂授粉给剑兰的景象是大自然最精彩的展示.

The gentle strumming of a banjo under a starlit sky brings peace to the heart; 在星光下轻弹班卓琴，给心灵带来平和.

**The negative sentiment sentences are enumerated as follows:**

The mosquito's persistent buzzing is a source of agony for many; 蚊子持续的嗡嗡声是许多人的痛苦来源.

The roach's presence in the kitchen was a sign of filth and neglect; 厨房里蟑螂的出现是脏乱和疏忽的标志.

The sight of a spider often invokes irrational fear and disgust; 蜘蛛的出现常常引起人们的恐惧和厌恶.

The deafening blast of gunfire ripped apart the silence, unleashing a maelstrom of horror and despair that ravaged the community; 枪声打破了平静，给社区带来了悲剧.

The stink of rotting flowers in the vase was overwhelmingly unpleasant; 花瓶里腐烂的花朵散发出极其令人不快的气味.

The appearance of cockroaches in the kitchen signifies a grimy, chaotic, and negligent environment; 蝗虫群造成了该地区的饥荒和贫困.

A wasp sting brings about distressing swelling and sharp pain; 黄蜂蜇伤导致肿胀和显著的疼痛.

The termite damage to the house led to a costly divorce; 白蚁对房屋的损害导致了一场昂贵的离婚.

Polluted air choked the vibrant tulips, turning them brown and ugly; 污染的空气使生机勃勃的郁金香变成了棕色和丑陋.

The knife was a grim reminder of the murder that had taken place; 刀子是一次发生的谋杀案的冷酷提醒.

The tank rumbled through the streets, a symbol of war and disaster; 坦克在街道上不断开炮，代表战争和灾难.

The prison was filled with the grief of those separated from their families; 监狱里充满了与家人分别的悲伤.

The sight of maggots in the trash can was utterly disgusting; 垃圾桶里的蛆虫令人极其厌恶.

The grenade explosion left a scene filled with chaos and death; 手雷爆炸留下了混乱和死亡的现场.

The rifle was a constant reminder of the violence that plagued the area; 该地区的暴力犯罪让步枪成为一个持续的提醒.

The repulsive fly, with its filth and disorder, marred the pure beauty of the delicate daisy; 脏乱的苍蝇落在美丽的雏菊上, 破坏了它的纯洁外观.

The poison ivy spread quickly through the garden, killing the nearby plants; 毒藤在花园中迅速蔓延, 杀死了附近的植物.

The crash of the cymbals was too loud, causing headaches among the audience; 钹声过大, 导致观众头痛.

The desecrated water has irreversibly marred the once enchanting landscape of the windbell grass marsh; 污染的水破坏了曾经美丽的风铃草沼泽.

The cockroach infestation in the apartment complex led to widespread complaints; 公寓大楼的蟑螂问题引发了广泛的投诉.

The harpoon was a tool of death in the whaling industry; 鱼叉是捕鲸工业中的死亡工具.

The divorce papers lay on the piano, a symbol of a love turned sour; 钢琴上的离婚文件, 是一段变酸的爱情的象征.

The dragonfly was trapped in an oil slick, struggling and doomed; 龙蝇被困在油污中, 挣扎且注定失败.

The bomb scare at the concert led to panic and chaos; 音乐会上的炸弹恐慌导致恐慌和混乱.

The rifle misfired, causing a tragic accident during the hunting trip; 狩猎旅行中步枪走火, 造成了悲剧性的事故.

The ugly graffiti defaced the once stunning mural of a magnolia; 丑陋的涂鸦破坏了曾经壮观的木兰壁画.

The centipede's appearance in the shower caused screams of fright; 洗澡时出现的蜈蚣引起了尖叫声.

The sickness spread by mosquitoes devastated the local community; 蚊子传播的疾病摧毁了当地社区.

Uncovering a dagger alongside the nearby corpse is a foreboding indication; 旁边的尸体发现了匕首, 这是一个不祥的迹象.

The cancer diagnosis shattered the peace and happiness of the family; 癌症诊断打破了家庭的平静和幸福.

The repulsive pond fails to attract the once joyful frogs; 肮脏的池塘不再吸引快乐的青蛙.

The assault in the quiet park marred its reputation as a safe place; 安静公园的袭击破坏了它作为安全场所的名声.

The bedbug bites caused discomfort and sleepless nights for the hotel guests; 酒店客人因床虱咬伤而感到不适和失眠.

The stink of the dead wasp was a repulsive discovery in the attic; 阁楼中死去的黄蜂散发出令人厌恶的气味.

The prison sentence for the innocent man was a grave injustice; 无辜者的监狱判决是严重的不公.

The rotten smell from the garbage attracted unwanted flies; 垃圾的腐烂气味吸引了不受欢迎的苍蝇.

The poverty in the area was evident in the neglected state of the public gardens; 该地区的贫困在被忽视的公共花园状态中显而易见.

The agony of the bee sting ruined the day's picnic; 蜜蜂蜇伤毁了一天的野餐.

The gnat swarm at the barbecue was a constant annoyance; 烧烤时的蚊子群是持续的烦恼.

The plight of the shipwreck weighs heavily on the seaside village; 沉船的悲剧困扰着海边的村庄.

The jail was a bleak and depressing place; 监狱是一个阴暗和压抑的地方.

The vomit on the sidewalk outside the bar was a disgusting sight; 酒吧外人行道上的呕吐物是一道令人厌恶的景象.

The cancer treatment center was a place of both hope and profound sadness; 癌症治疗中心既是希望之地也是深刻悲伤之地.

The death of the cherished rose garden was caused by a harsh winter; 严冬导致珍爱的玫瑰园死亡.

The ugly dispute over the land poisoned relationships in the family; 土地争议的丑陋争吵毒化了家庭关系.

The abrupt shutting of the piano lid frightened each person in the tranquil room; 钢琴盖突然合上, 惊吓了安静房间里的每个人.

The pollution in the river killed thousands of fish, creating a tragic environmental disaster; 河流的污染导致成千上万的鱼死亡, 造成了悲剧性的环境灾难.

The hatred visible in the community divided it deeply; 社区中可见的仇恨深刻地分裂了它.

The accident at the factory left several workers with serious injuries; 工厂的事故导致几名工人严重受伤.

The grief of losing a pet was evident in every corner of their home; 家中每一个角落都流露出失去宠物的悲伤.

## References

1. Greenwald, A.G.; Banaji, M.R. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychol. Rev.* **1995**, *102*, 4. [CrossRef]
2. Wilson, T.D.; Lindsey, S.; Schooler, T.Y. A model of dual attitudes. *Psychol. Rev.* **2000**, *107*, 101–126. [CrossRef] [PubMed]
3. Chang, J.H.; Zhu, Y.Q.; Wang, S.H.; Li, Y.J. Would you change your mind? An empirical study of social impact theory on Facebook. *Telemat. Inform.* **2018**, *35*, 282–292. [CrossRef]
4. Dasgupta, N.; Mcghee, D.E.; Greenwald, A.G.; Banaji, M.R. Automatic Preference for White Americans: Eliminating the Familiarity Explanation. *J. Exp. Soc. Psychol.* **2000**, *36*, 316–328. [CrossRef]
5. Greenwald, A.G.; McGhee, D.E.; Schwartz, J.L. Measuring individual differences in implicit cognition: The implicit association test. *J. Personal. Soc. Psychol.* **1998**, *74*, 1464. [CrossRef] [PubMed]
6. Molesworth, B.R.; Chang, B. Predicting pilots' risk-taking behavior through an implicit association test. *Hum. Factors* **2009**, *51*, 845–857. [CrossRef]
7. Teachman, B.A.; Woody, S.R. Staying tuned to research in implicit cognition: Relevance for clinical practice with anxiety disorders. *Cogn. Behav. Pract.* **2004**, *11*, 149–159. [CrossRef]
8. Northrup, D.A. *The Problem of the Self-Report in Survey Research*; York University: Toronto, ON, Canada, 1997.
9. Liu, B. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing*; Routledge: Oxford, UK, 2010; Volume 2, pp. 627–666.
10. Recasens, M.; Danescu-Niculescu-Mizil, C.; Jurafsky, D. Linguistic models for analyzing and detecting biased language. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL, Sofia, Bulgaria, 4–9 August 2013; pp. 1650–1659.
11. Bolukbasi, T.; Chang, K.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Proceedings of the 30th Advances in Neural Information Processing Systems, NIPS, Barcelona, Spain, 5–10 December 2016; pp. 4349–4357.
12. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186. [CrossRef]
13. May, C.; Wang, A.; Bordia, S.; Bowman, S.R.; Rudinger, R. On measuring social biases in sentence encoders. In Proceedings of the North American Chapter of the Association for Computational Linguistics, NAACL, Minneapolis, MN, USA, 3–5 June 2019; pp. 622–628.
14. Petreski, D.; Hashim, I.C. Word embeddings are biased. But whose bias are they reflecting? *AI Soc.* **2023**, *38*, 975–982. [CrossRef]
15. Lima, R.M.d.; Pisker, B.; Corrêa, V.S. Gender bias in artificial intelligence. *J. Telecommun. Digit. Econ.* **2023**, *11*, 8–30. [CrossRef]
16. Dobrzeniecka, A.; Urbaniak, R. A Bayesian approach to uncertainty in word embedding bias estimation. *Comput. Linguist.* **2024**, 1–56. [CrossRef]
17. Garg, N.; Schiebinger, L.; Jurafsky, D.; Zou, J. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proc. Nat. Acad. Sci. USA* **2017**, *115*, E3635–E3644. [CrossRef] [PubMed]
18. Paulhus, D.L.; Vazire, S. The self-report method. In *Handbook of Research Methods in Personality Psychology*; Guilford Press: New York, NY, USA, 2007; Volume 1, pp. 224–239.
19. Stone, A.A. *The Science of Self-Report. Implications for Research and Practice*; Routledge: Oxford, UK, 2000.
20. Westen, D. The scientific status of unconscious processes: Is Freud really dead? *J. Am. Psychoanal. Assoc.* **1999**, *47*, 1061–1106. [CrossRef]
21. Litt, E.; Hargittai, E. The Imagined Audience on Social Network Sites. *Soc. Media + Soc.* **2016**, *2*. [CrossRef]

22. Greenwald, A.G.; Nosek, B.A.; Banaji, M.R. Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *J. Personal. Soc. Psychol.* **2003**, *85*, 197–216. [CrossRef] [PubMed]
23. Greenwald, A.G.; Hummert, M.L.; Garstka, T.A.; O'Brien, L.T.; Mellott, D.S. Using the Implicit Association Test to Measure Age Differences in Implicit Social Cognitions. *Psychol. Aging* **2002**, *17*, 482–495.
24. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
25. Babaeianjelodar, M.; Lorenz, S.; Gordon, J.; Matthews, J.; Freitag, E. Quantifying Gender Bias in Different Corpora. In Proceedings of the Web Conference 2020, WWW, Taipei, Taiwan, 20–24 April 2020; pp. 752–759.
26. Schröder, S.; Schulz, A.; Kenneweg, P.; Feldhans, R.; Hinder, F.; Hammer, B. Evaluating metrics for bias in word embeddings. *arXiv* **2021**, arXiv:2111.07864.
27. Garrido-Muñoz, I.; Montejo-Ráez, A.; Martínez-Santiago, F.; Ureña-López, L.A. A survey on bias in deep NLP. *Appl. Sci.* **2021**, *11*, 3184. [CrossRef]
28. Charlesworth, T.E.; Caliskan, A.; Banaji, M.R. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2121798119. [CrossRef]
29. Morehouse, K.; Rouduri, V.; Cunningham, W.; Charlesworth, T. Traces of Human Attitudes in Contemporary and Historical Word Embeddings (1800–2000). *Res. Sq.* **2023**, *preprint*.
30. Durrheim, K.; Schuld, M.; Mafunda, M.; Mazibuko, S. Using word embeddings to investigate cultural biases. *Br. J. Soc. Psychol.* **2023**, *62*, 617–629. [CrossRef] [PubMed]
31. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
32. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
33. Igor, M.; Miha, G.; Jasmina, S.; Matjaz, P. Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PLoS ONE* **2016**, *11*, e0155036.
34. Pozzi, F.A.; Fersini, E.; Messina, E.; Liu, B. *Sentiment Analysis in Social Networks*; Elsevier: Amsterdam, The Netherlands, 2016.
35. Poria, S.; Gelbukh, A.; Cambria, E. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowl. Based Syst.* **2016**, *108*, 42–49. [CrossRef]
36. Xin, W.; Liu, Y.; Sun, C.; Wang, B.; Wang, X. Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26–31 July 2015.
37. Zhu, X.; Huang, M.; Qian, Q. Encoding Syntactic Knowledge in Neural Networks for Sentiment Classification. *Acm Trans. Inf. Syst.* **2017**, *35*, 26.
38. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1253. [CrossRef]
39. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. Pre-trained models: Past, present and future. *AI Open* **2021**, *2*, 225–250. [CrossRef]
40. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.Y.; Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing EMNLP, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
41. Chen, D.; Manning, C.D. A fast and accurate dependency parser using neural networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Doha, Qatar, 25–29 October 2014; pp. 740–750.
42. Cao, Y.T.; Sotnikova, A.; Daumé, H., III; Rudinger, R.; Zou, L. Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V., Eds.; ACL: Seattle, WA, USA, 2022; pp. 1276–1295. [CrossRef]
43. Meister, C.; Stokowiec, W.; Pimentel, T.; Yu, L.; Rimell, L.; Kuncoro, A. A Natural Bias for Language Generation Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; ACL: Toronto, ON, Canada, 2023; pp. 243–255. [CrossRef]
44. Cheng, M.; Durmus, E.; Jurafsky, D. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; ACL: Toronto, ON, Canada, 2023; pp. 1504–1532. [CrossRef]
45. Zhao, Y.; Wang, B.; Zhao, D.; Huang, K.; Wang, Y.; He, R.; Hou, Y. Mind vs. Mouth: On Measuring Re-judge Inconsistency of Social Bias in Large Language Models. *arXiv* **2023**, arXiv:2308.12578.
46. Perloff, R.M. *The Dynamics of Persuasion: Communication and Attitudes in the Twenty-First Century*; Routledge: Oxford, UK, 2020.
47. Greenwald, A.G.; Banaji, M.R. The implicit revolution: Reconceiving the relation between conscious and unconscious. *Am. Psychol.* **2017**, *72*, 861. [CrossRef]

48. Sap, M.; Prasettio, M.C.; Holtzman, A.; Rashkin, H.; Choi, Y. Connotation Frames of Agency and Power in Modern Films. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017.
49. Carpenter, J.; Preotiuc-Pietro, D.; Flekova, L.; Giorgi, S.; Hagan, C.; Kern, M.L.; Buffone, A.; Ungar, L.; Seligman, M. Real Men don't say 'cute': Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes. *Soc. Psychol. Personal. Sci.* **2017**, *8*, 310–322. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.