Judging Credible and Unethical Statistical Data Explanations via Phrase Similarity Graph

Completed Research Paper

Kang Zhang
Graduate School of Systems Life Sciences
Kyushu University
Fukuoka, Japan
zksoda@hotmail.com

Einoshin Suzuki
Faculty of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
suzuki@inf.kyushu-u.ac.jp

Abstract

We propose a graph-based method to judge credible and unethical statistical data explanations with the exploitation of human instincts proposed by Rosling et al. Our previous work proposes three categories of statistical data explanations and three corresponding judgment methods based on phrase embedding and carefully designed comparison conditions. However, we observe that the previous method β exhibits low accuracy in the explanations of (β) category due to its counter-intuitive semantic similarities between several phrases. To address this limitation and improve the performance, our new method β^2 constructs a Phrase Similarity Graph to generate additional comparison conditions and devises a credibility score to aggregate the conditions with their importance quantified by graph entropy. The experimental results show that our β^2 achieves over 81% accuracy while the previous method β achieves about 57%. Scrutiny reveals that our β^2 mitigates the problem of the counter-intuitive semantic similarities at a satisfactory level.

Keywords: AI ethics, biased statistical data explanations, phrase similarity graph, graph entropy, text classification.

Introduction

As Artificial Intelligence (AI) systems become more prevalent and influential in our society, they are giving rise to numerous ethical concerns across various fields. The misconducts of Deepfakes pose a serious threat to truth, trust, and privacy by spreading false information and manipulating public opinions (Westerlund, 2019). Similarly, the racist, sexist, and offensive comments generated by the chatbot Tay harmed the reputation of the chatbot-creators, though Tay was designed to act in a funny and exuberant manner (Zemčík, 2021). Although the advent of ChatGPT (Thorp, 2023) has the potential to revolutionize various industries and aspects of our daily lives, such a practical language model also holds the possibility of generating and spreading seemingly convincing yet biased information (Liebrenz et al., 2023; Zhuo et al. 2023), such as fake news and inflammatory tweets. These kinds of information pose a significant challenge to the morality of our society. Among such misinformation, those that are credible and exploit human instincts are more influential than others as they are more likely to be accepted by people. Therefore, judging the credibility of unethical information is a crucial task to prevent the problem.

In this paper, we limit our scope on unethical statistical data explanations (Zhang et al., 2022). While fake news and misinformation cover broader categories of false or misleading information, an unethical statistical data explanation is a specific type of misinformation that refers to an invalid interpretation of statistical data. Following our previous work (Zhang et al., 2022), an unethical statistical data explanation is defined by considering three conditions, including 1) the statistical data seem to be valid, 2) the data can prove why the explanation is not valid, and 3) the explanation exploits at least one of the biased human instincts mentioned in a book entitled "Factfulness" (Rosling et al., 2018). The globally successful book introduces 10 human instincts and several examples of unethical statistical data explanations which exploit these instincts. The book emphasizes the importance of thinking based on facts and correct understandings derived from statistical data, instead of innate and fixed patterns in our mind, i.e., human instincts. Take as an example an explanation "Asia is the cause of the large amount of CO2 emissions"[1] with its statistical data depicting GDP per capita, total amount of CO2 emissions, and CO2 emissions per capita of four continents in Figure 1. The statistical data show that although Asia seems to be the cause in the view of total emissions, the explanation is refuted by the per person emission view with respect to the GDP per capita. However, although the statistical data clearly contradicts the explanation, some portion of people would accept the explanation as it exploits the single perspective instinct, i.e., our tendency to prefer a single cause or solution (Zhang et al., 2022). Among such unethical explanations, we believe that credible and unethical explanations deserve special attention as they pose a significant challenge to our rationality and understanding, while the non-credible explanations are less harmful as people do not believe them.

The pioneering work (Zhang et al., 2022) defines 18 types (I-XVIII) of credible and unethical explanations each with its statistical data. The 18 types of explanations can be classified into three categories based on their subjects and characteristics, including (α) habits and diseases, (β) subjects and properties, and (γ) subjects and trends. Accordingly, the work proposes three judgment methods α, β, and γ to investigate their credibilities by carefully designed comparison conditions based on the phrase embedding technique, which compares the semantic relevance between phrases specified in the explanations. For example, comparing if "women" are more relevant to "low math scores" than "men" is such a condition for judging the explanation "women have lower math scores than men". The results show that methods α and γ exhibit perfect and promising performance, respectively, due to the simpler nature of their target explanations compared with method β. This demonstrates the judgement methods are effective when the designed comparison conditions only involve a small number of phrases. However, since the phrases in (β) category are more complex, including multiple subjects and properties, several counter-intuitive semantic similarities between these subjects and properties lead to undesired results of the comparison conditions in method β. Therefore, method β achieves relatively low accuracy on the task, reflecting the difficulties and challenges for judging explanations in (β) category.

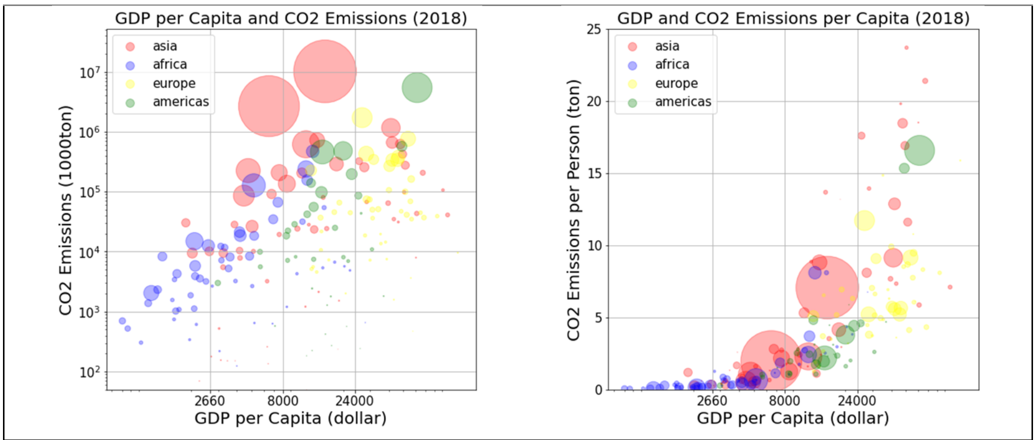


Figure 1. GDP per capita and total amount of CO2 emissions versus GDP per capita and CO2 emissions per capita.

[1] All unethical examples in this paper are either adopted from other sources or slightly modified from them and do not reflect the beliefs of the authors nor our organizations. In all cases, such examples are not believed by the authors of the sources, either.

In this paper, to achieve a higher accuracy on judging credible and unethical statistical data explanations in (β) category (Zhang et al., 2022), we propose a new judgment method β^2, which constructs a Phrase Similarity Graph to model the statistical data explanation by considering more phrases. The graph can explicitly represent these phrases and their semantic similarities, where the conditions for the judgment can be simply selected based on node combinations. Then a credibility score for judging the credibility of the explanation is proposed based on the selected conditions and graph entropy.

The main contributions of this paper are summarized as follows.

1. We propose a graph-based judgment method β^2. To improve the low accuracy of previous method β (Zhang et al., 2022), β^2 constructs a Phrase Similarity Graph to consider more phrases for generating necessary conditions and adopts graph entropy to quantify the different importance of the generated conditions for judgment.
2. When judging an explanation, our method β^2 explores the semantic relations between more phrases by considering their synonyms, which mitigates the problem of the counter-intuitive semantic similarities between limited phrases in method β.
3. We extend the dataset from Zhang et al. (2022) by adding 3 additional types of explanations to evaluate the performance of our method. The experimental results on the extended dataset demonstrate the superiority of our method β^2 compared with the baseline method β.

Related Work

Unethical and biased explanations, such as fake news and misinformation, are pervasive in various domains around the world (Scheufele et al., 2019). Misinformation can be defined as incorrect or counterfactual information, while fake news is a specific type of misinformation which is intentionally created to mislead the audience (Scheufele et al., 2019). The detection of fake news and misinformation has been extensively studied mainly based on analyzing the linguistic features (Castillo et al., 2011), the meta information (Shu et al., 2020), and fact-checking techniques (Rashkin et al., 2017). Reis et al. (2019) integrate the content of news with metadata to extract textual, source, and environment features and adopt several classic machine learning classifiers for automatic fake news detection. Similarly, through integrating meta data with texts, a hybrid Convolutional Neural Network (CNN) is devised to classify fake news based on surface-level linguistic patterns (Wang, 2017). By devising a hybrid Recurrent Neural Network (RNN) model, Ruchansky et. al (2017) incorporate texts, responses, and sources of articles for fake news classification. With the growing number of fact-checking Websites and crowdsourcing services, computer-aided fact-checking systems have been developed to judge misinformation by evaluating its semantic similarity with the truth (Nakov et al., 2021; Zeng et al., 2021). Moreover, since fake news with images or videos are becoming increasingly prevalent with the development of multimedia technology, multimodal information including visual and textual features have been explored for more accurate detection (Cao et al., 2020; Khattar et al., 2019; Wang et al., 2018).

Unethical statistical data explanations are a particular type of misinformation, which are defined by considering the validity of the data, the objectiveness of the explanation, and the exploitation of human instincts (Zhang et al., 2022). Statistical ethics refers to the ethical consideration and principles that guide the collection, analysis, interpretation, and communication of statistical information (Lesser et al., 2004). Statistical ethics covers a wide range of topics, such as the selection bias in data collection for clinical research (Tripepi et al., 2010), the misuse and abuse of statistical data for biomedical research (Thiese et al., 2015), and the survivorship bias in statistical for longitudinal mental health surveys during the COVID-19 pandemic (Czeisler et al., 2021). These works mainly focus on addressing ethical concerns in statistical data, aiming to promote the integrity and the responsible use of data in their domains. Among such works, Zhang et al. (2022) proposed that credible and unethical explanations of statistical data due to the human instinct exploitation deserve special attention, since they can lead to formation of stereotypes and prejudice for people. Such explanations may hinder people from developing correct understandings of the facts even if statistical data support them. Zhang et al. (2022) is the first work to define 18 types of unethical statistical data explanations and provide three judgment procedures to investigate their credibilities based on phrase embedding. However, as we explained in Introduction, their performance is unsatisfactory on (β) category of explanations due to the counter-intuitive semantic similarities between multiple subjects and properties, leaving room for further exploration.

As we mentioned in Introduction, we propose a graph-based method for the task. Graph structures have been widely employed in fact-checking and misinformation detection, as they can make the structure of free text explicit and are easily manageable by downstream algorithms. These works can be mainly classified into similarity-based and knowledge-based approaches. Similarity-based approaches often represent social media posts (Wu et al., 2015), sentences, or words in news articles (Balcerzak et al., 2014; Kazemi et al., 2020; Mao et al., 2022) as nodes and build edges to represent their relations in a graph. TextRank (Mihalcea et al., 2004) is adopted to identify credible statements from a graph in which the sentences and their semantic similarities represent nodes and edges (Balcerzak et al., 2014), respectively. Utilizing the same kind of graph, Biased TextRank (Kazemi et al., 2020) associates an explanation extraction with the fact-checking task by comparing the similarities between the extracted statements with the ground truth. On the other hand, knowledge-based approaches often retrieve evidence which supports or refutes the information from a large and reliable knowledge graph (Kim et al., 2020; Shu et al., 2017). Vedula et al. (2021) jointly exploit the concept-relationship structure and semantic contextual cues from the knowledge graph to detect the veracity of an input fact and generate a human-comprehensible explanation justifying the fact. For health misinformation detection, a knowledge-guided graph attention network is devised by incorporating a medical knowledge graph and an article-entity bipartite graph (Cui et al., 2020). Different from these graph-based methods for misinformation detection tasks, the Phrase Similarity Graph in our method is proposed to tackle the issue of counter-intuitive semantic similarities by considering more phrases, which improves the accuracy for judging the statistical data explanations.

Graph entropy is a measure to understand and analyze the structure and complexity of a graph, which is often utilized to quantify the degree of uncertainty for graph data. Graph entropy is usually task-specific, i.e., it depends on the characteristics of the network. These works include structure and feature entropy for node embedding dimension selection (Luo et al., 2021), parametric graph entropy for analyzing information processing (Dehmer et al., 2008), and conditional substructure entropy for graph anomaly detection (Noble et al., 2003). Among such works, Sen et al. (2018) define the sub-graph entropy by focusing on the complexity of connections between nodes in functional brain networks. The sub-graph entropy is computed by exploring the node connectivity, i.e., edge weights, to evaluate the importance of each sub-graph in a whole graph. Since our Phrase Similarity Graph considers node combinations and their connections from its sub-graphs to generate comparison conditions for judgment, we adopt sub-graph entropy to measure the importance of the comparison conditions from different sub-graphs.

Target Problem

As explained above, we focus our attention on the credible and unethical explanations of statistical data with the exploitation of human instincts. Following the definition in the previous work (Zhang et al., 2022), we assume five conditions for the credible and unethical explanations of statistical data.

1) Data seem to be valid, ideally taken from an authoritative source, e.g., WHO.
2) The explanation is significant.
3) The explanation seems to be believed by a certain number of people.
4) The data can prove why the explanation is not valid.
5) The explanation exploits at least one of the ten human instincts in Rosling et al. (2018).

The 1), 4), and 5) conditions contribute to the unethical nature of a statistical explanation, which consider its validity, objectiveness, and the exploitation of human instincts, respectively. The 2) and 3) conditions are also necessary as they consider its significance and credibility, respectively. Without 2) and 3), the explanation is not harmful as people do not pay attention to them.

As we discussed in Introduction, unethical statistical data explanations that are credible deserve more attention than those that are not because they have a greater negative impact on correct human understanding. Therefore, we tackle the same target problem as in Zhang et al. (2022), which is to classify a given statistical data explanation as either credible and unethical (class 1) or not (class 0). It is important to note that neither the previous methods in Zhang et al. (2022) nor our method takes into account the significance of the explanations in (β) category, which presents a challenge to current methods for the task. We recognize the importance of addressing this issue and consider it as a future direction for investigation and exploration.

The target problem is formulated as a binary classification task, where the goal is to predict the class labels of the explanations in (β) category. The ground-truth class labels are given by humans for the evaluation purpose only. The input of the target problem is an explanation, its statistical data, and its phrases, which will be explained in the next Section. The output is the predicted class label (0 or 1) of the explanation. To evaluate our judgment method, we utilize accuracy as the evaluation metric.

Methodology

In this section, we first introduce the explanations accompanied by their statistical data and the judgment method proposed by the most relevant work. Then we present the overall procedure of our graph-based method, including the Phrase Similarity Graph, graph entropy, and the credibility score for the task.

Preliminaries

The definition and judgment of credible and unethical explanations of statistical data are first introduced by Zhang et al. (2022). They define 18 types of explanations each of which exploits at least one human instinct. These explanations describe 7 kinds of statistical data, including (A) values of a probabilistic variable under 2 conditions, (B) a scatter plot of 2 probabilistic variables, (C) scatter plots in different categories, (D) a probability density function of a probabilistic variable and a plot of its average value, (E) a time-series chart or scatter plots in chronological order, possibly with an additional one, (F) scatter plots of 2 probabilistic variables focusing on the total values and the average values, and (G) a funnel plot. Moreover, they also clarify the reason each explanation is not valid according to its statistical data and generate its variants by providing candidate phrases.

In this paper, we concentrate on judging the credible and unethical statistical data explanations in (β) category since the previous method highlights the difficulty and challenges in judging this category. The explanation in (β) category has the form of subject X is more likely to have property Y compared with other subjects. To judge the explanation, 5 kinds of phrases X, X', Y_{base}, Y, and $\underline{Y}$ are specified. X and Y are explicitly mentioned in the explanation. X' is a subject or a set of subjects in the opposite class of X, which can be specified explicitly or generated based on knowledge on English language. $\underline{Y}$ is specified as the inverse property of Y, which is typically in the form of an adjective followed by a noun phrase Y_{base}. We show two explanations of credible and unethical explanations in (β) category accompanied by their statistical data and phrases in Figures 2 and 3. For example, in Figure 2, X and X' are "Muslims" and "Christians". Y *and* $\underline{Y}$ are "many babies" and "few babies" with respect to a base word Y_{base}, i.e., "babies".

Previous method β (Zhang et al., 2022) tackles the target problem based on carefully designed conditions between phrases for comparison. Specifically, β predicts the class label of an explanation as credible and unethical (class 1) if and only if the following two conditions hold, otherwise class 0.

$$IF\ \left(\theta_{XY} > \theta_{X\underline{Y}}\right) \wedge \forall X'(\theta_{XY} > \theta_{X'Y})\ THEN\ 1\ ELSE\ 0, \#(1)\ where\ \theta_{XY} = \frac{Sim(X,Y)}{Sim(X,Y_{base})}, \theta_{X\underline{Y}} = \frac{Sim(X,\underline{Y})}{Sim(X,Y_{base})}, \theta_{X'Y} = \frac{Sim(X',Y)}{Sim(X',Y_{base})}. \#(2)$$

Here θ_{XY}, $\theta_{X\underline{Y}}$, and $\theta_{X'Y}$ represent the semantic relevance degrees between X and Y, X and $\underline{Y}$, as well as X' and Y, respectively. The semantic similarity is a cosine-similarity of the embeddings of the phrases by Sentence-BERT (Reimers et al., 2019), which is a state-of-the-art deep model for sentence and phrase embeddings. Specifically, the semantic similarity $Sim(\cdot)$ between X and Y is given as follows.

$$Sim(X,Y) = \frac{s(X)\cdot s(Y)}{\|s(X)\|\ \|s(Y)\|}, \#(3)$$

where $s(\cdot)$ represents the embedding vector by Sentence-BERT. When judging an explanation by the two conditions, the first condition compares if X is more relevant to Y than its inverse $\underline{Y}$. The second one compares if the property Y is more relevant to the subject X than any other subject X' belonging to the opposite class. However, when judging the example in Figure 2 by the previous method β, the semantic similarity between X: Muslims and Y: many babies Sim("Muslims", "many babies") = 0.234 is counter-intuitively lower than the similarity between X: Muslims and $\underline{Y}$: few babies" $Sim($"Muslims", "few babies"$) = 0.245$, leading to a false negative.

(Type IV) **Muslims have many babies compared to Christians.**
X: Muslims. Y: many babies.
(Phrases) X: Muslims. X': Christians. Y_{base}: babies. Y: many babies. $\underline{Y}$: few babies.

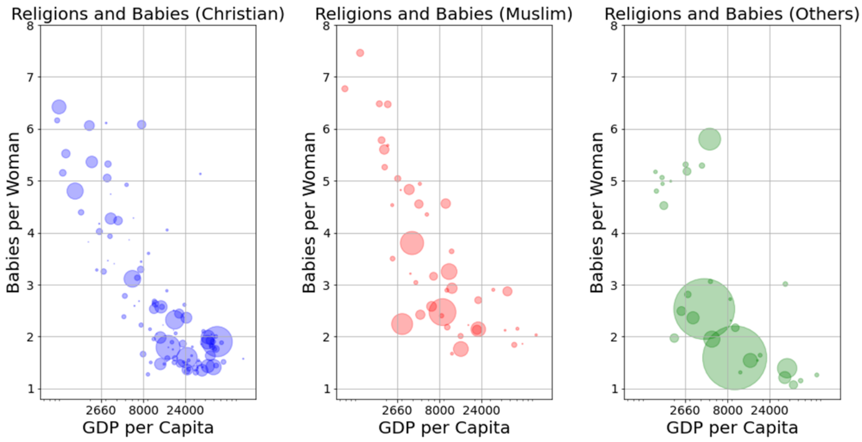


Figure 2. Religions and number of babies.

(Type II) **Cuba is the poorest of the healthiest countries.**
X: Cuba. Y: poorest.
(Phrases) X: Cuba. X': {Japan, Singapore}. Y: poorest. $\underline{Y}$: richest.

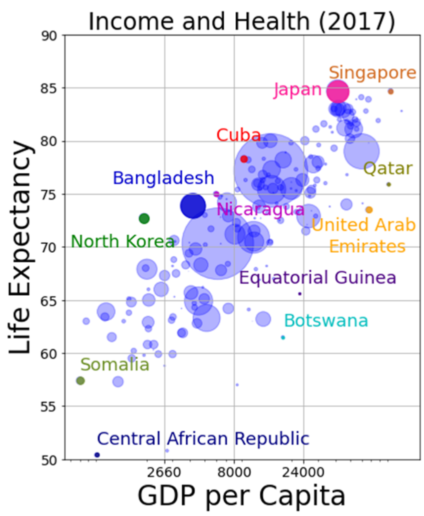


Figure 3. Life expectancy and GDP per capita in 2017.

As shown in Figure 3, type II explanations have no Y_{base} since their properties, i.e., "poorest" or "richest", are in a form of the superlative of an adjective. In such a case, the relevance between candidates is simply calculated by their semantic similarities, e.g., $\theta_{XY} = Sim(X, Y)$.

Judgment method β^2

Since the unsatisfactory performance of the previous method β is due to the counter-intuitive semantic similarities between limited phrases (Zhang et al., 2022), we introduce our graph-based method β^2 to consider more phrases and explore their relevance for judgment. Given an explanation, method β^2 first extends its phrases and constructs a Phrase Similarity Graph to model these phrases and their semantic similarities. Afterwards, the conditions for judgment are generated from subgraphs selected from the Phrase Similarity Graph. The importance of the conditions generated from the subgraphs are quantified by their sub-graph entropy. Lastly, a credibility score is devised by aggregating the conditions with their importance to judge the explanations.

We show the overall procedure of method β^2 in Algorithm 1. The phrases $X,\ X', Y_{base}, Y$, and $\underline{Y}$ are extended to phrase sets $X_{syno}, X'_{syno}, Y_{base,\ syno}, Y_{syno}$, and $\underline{Y}_{syno}$ by considering their synonyms in step 1. The Phrase Similarity Graph G is constructed based on the extended phrase sets in step 2 and the subgraphs G^k are extracted by selecting node groups from G in step 3. Then the conditions for judgment are generated via four criteria 1) − 4) based on the selected node combinations in G^k. We are going to explain the details of each step in the following sections.

Algorithm 1. **Overall procedure of method β^2.**
Input: Statistical data explanation; Phrases $X,\ X', Y_{base}, Y, \underline{Y}$; Credibility threshold $\theta_{credible}$. **Output:** Credible and unethical (class label 1) or not (class label 0) for the explanation. 1: $X_{syno}, X'_{syno}, Y_{base,\ syno}, Y_{syno}, \underline{Y}_{syno} = Extend(X,\ X', Y_{base}, Y, \underline{Y})$; 2: Phrase Similarity Graph $G = GetGraph(X_{syno}, X'_{syno}, Y_{base,\ syno}, Y_{syno}, \underline{Y}_{syno})$; 3: Subgraphs $\{G^k \mid k = 1, \dots, K\} = GetSubgraph(G)$; 4: **For** each G^k in $\mathcal{G}$: 5: Generate conditions via four criteria 1) − 4); 6: Calculate sub-score s_k via Eq. (11); 7: Calculate sub-graph entropy $H(G^k)$ via Eq. (12)-(13); 8: **End For** 8: Calculate important weight λ_k for each sub-score via Eq. (14); 9: Calculate credibility score S for the explanation via Eq. (15); 10: If $S > \theta_{credible}$, output class label 1; else output class label 0.

Phrase Similarity Graph for Statistical Data Explanations

Given a statistical data explanation with its phrases, we first generate more phrases by considering their synonyms. Then we construct a Phrase Similarity Graph to model an explanation by representing its phrases as nodes and the semantic similarities between different sets of nodes as edges.

Since the unsatisfactory performance of the previous method β is due to the counter-intuitive semantic similarities between limited phrases (Zhang et al., 2022), we propose to consider more phrases to explore their relevance for judgment. Specifically, each kind of phrase is extended to a phrase set by considering its synonyms. As shown in Figure 2, there are 5 kinds of phrases for each (β) explanation, i.e., X, X', Y_{base}, Y, and $\underline{Y}$. We adopt an emerging powerful language model ChatGPT[2] to generate top-n synonyms of each phrase, as we will show the details in Experimental Setup. The extended phrase sets are represented as $X_{syno}, X'_{syno}, Y_{base,syno}, Y_{syno}, \underline{Y}_{syno}$ according to $X, X', Y_{base}, Y, \underline{Y}$, respectively.

We propose a Phrase Similarity Graph to explicitly model the phrase sets and their semantic similarities. Following several graph-based works (Chen et al. 2020; Deng et al. 2021; Toivonen et al., 2011), our graph is an attributed graph defined as $G = (V, E, X, W)$, where $V = \{v_1, \dots, v_n\}$ represents the set of nodes. $X \in R^{n \times d}$ represents the attribute matrix, where the vector $x_i \in R^d$ in X represents the attribute of node v_i. $E = \{e_{i,j} \mid i, j = 1, \dots, N\}$ and $W = \{\omega_{v_i v_j} \mid i, j = 1, \dots, N\}$ represent the set of edges with weights between nodes v_i and v_j, respectively.

[2] https://openai.com/blog/chatgpt/

In the Phrase Similarity Graph, a node, a node attribute, and an edge with weight between two nodes represent a phrase, a phrase embedding vector, and a semantic relations computed by cosine-similarity between two phrases in the explanation, respectively. As shown in Figure 4, the graph is constructed as a tripartite graph $G = (V, E, X, W)$ with three disjoint node subsets V_{base}, $V_{subject}$, and $V_{property}$, where nodes in V_{base} represent the phrases of base words in $Y_{base,syno}$, nodes in $V_{subject}$ represent the phrases of subjects in $X_{syno} \cup X'_{syno}$, and nodes in $V_{property}$ represent the phrases of properties in $Y_{syno} \cup \underline{Y}_{syno}$, respectively.

Following Zhang et al. (2022), we consider the semantic similarities between different kinds of phrases for judgment, i.e., the similarities between subjects and base words and the similarities between subjects and properties. Therefore, the edges E are built between nodes in V_{base} *and* $V_{subject}$, as well as between nodes in $V_{subject}$ *and* $V_{property}$, respectively. The node attributes X are embedding vectors of phrases, which are generated by Sentence-BERT (Reimers et al., 2019). The edge weight ω_{v_i,v_j} in W represents the semantic similarity between two nodes v_i and v_j, which is calculated by the cosine-similarity $\mathrm{Sim}(\cdot)$ between their node attributes x_i and x_j. Formally, the edge weight ω_{v_i,v_j} between nodes v_i and v_j is given as follows.

$$\omega_{v_i,v_j} = Sim(v_i, v_j) = \frac{x_i \cdot x_j}{\|x_i\| \, \|x_j\|}. \#(4)$$

We present an example of a Phrase Similarity Graph for a statistical data explanation in Figure 5. Given an explanation, each kind of phrase are first extended as a phrase set by considering its synonyms. Then the Phrase Similarity Graph is constructed to represent all the phrases in the phrase sets as nodes and the semantic similarities between nodes from different subsets of nodes as edges. For simplicity, heavy edge weights are shown by thick width of edges in the graph in Figure 5.

Additional Conditions by Subgraphs in Phrase Similarity Graph

In addition to the two conditions in the previous method β, we propose that further conditions should be considered for judgment. Take the explanation in Figure 2 as an example. The two conditions in method β are to compare the relevance between subject X and different properties Y and $\underline{Y}$, as well as property Y with different subjects X and X', represented as $\theta_{XY} > \theta_{X\underline{Y}}$ and $\theta_{XY} > \theta_{X'Y}$. However, the relevance between the opposite subjects X' with different properties Y and $\underline{Y}$, as well as the opposite property $\underline{Y}$ with different subjects X and X', represented as $\theta_{X'\underline{Y}} > \theta_{X'Y}$ and $\theta_{X'\underline{Y}} > \theta_{X\underline{Y}}$, has not been considered, while it may potentially contribute to the judgment. Nevertheless, as shown in Figure 5, as the number of the phrases increases in the extended phrase sets, designing necessary comparison conditions for judgment becomes more difficult and complex. To simplify the design of conditions, we propose to generate necessary conditions from the subgraphs extracted from the Phrase Similarity Graph.

Specifically, each subgraph is extracted by selecting one node from each phrase set, e.g., X, X', Y_{base}, Y, and $\underline{Y}$ from $X_{syno}, X'_{syno}, Y_{base,syno}, Y_{syno}$, and $\underline{Y}_{syno}$ with the weighted edges, represented as $G^k = (V^k, E^k, X^k, W^k), k = 1, \ldots, K$, where K is the number of all subgraphs extracted from Phrase Similarity Graph G. As the conditions are to compare the relevance between subjects and properties for judgment, the

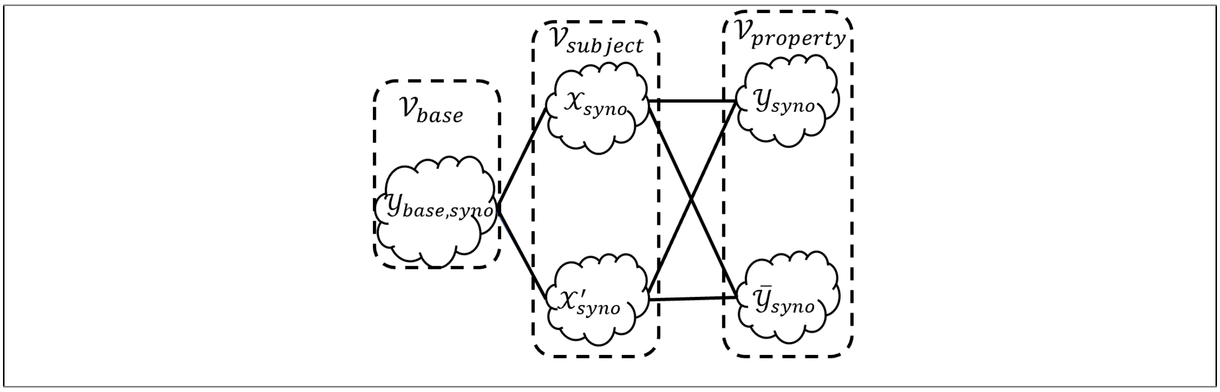


Figure 4. Phrase Similarity Graph to model phrase sets and their semantic similarities.

(Type XII) **Asia is the cause of the large amount of CO2 emissions.**
X: Asia. Y: large amount of CO2 emissions.
(Phrases)
X: Asia. $X' \in$ {Africa, Europe}. Y_{base}: CO2 emissions.
Y: large amount of CO2 emissions. $\underline{Y}$: small amount of CO2 emissions.
(Phrase sets)
X_{syno}: {Asia, Asian countries, Asian nations}.
X'_{syno}: {Europe, European countries, European nations, Africa, African countries, African nations}.
$Y_{base,syno}$: {CO2 emissions, greenhouse gas emissions, carbon dioxide emissions}.
Y_{syno}: {large amount of CO2 emissions, large amount of greenhouse gas emissions, large amount of carbon dioxide emissions}.
$\underline{Y}_{syno}$: {small amount of CO2 emissions, small amount of greenhouse gas emissions, small amount of carbon dioxide emissions}

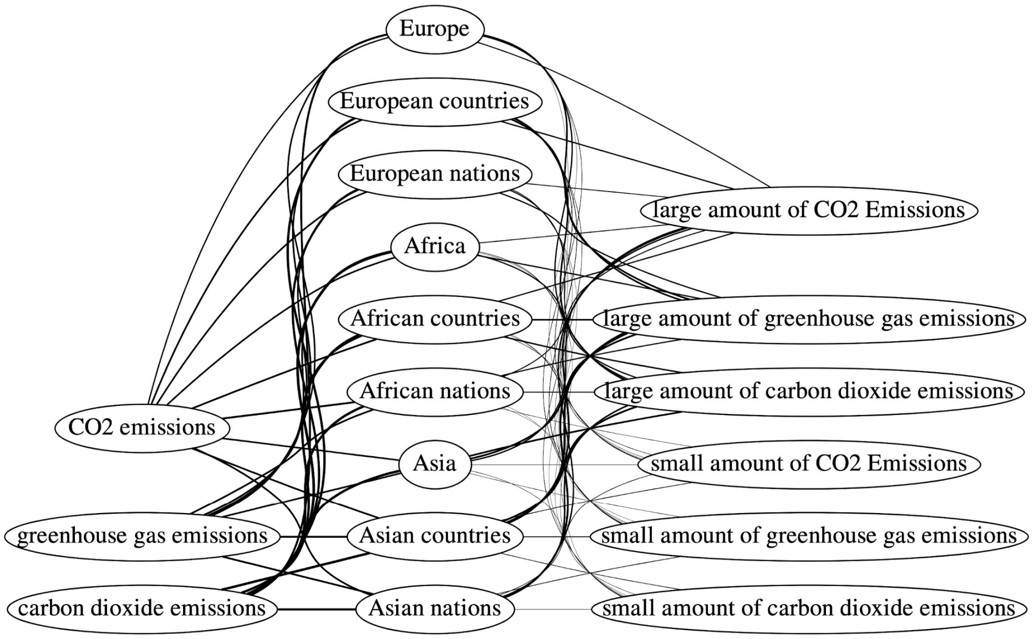


Figure 5. Example of constructing the Phrase Similarity Graph from a statistical data explanation.

subjects in the opposite classes are selected in pairs, i.e., nodes in X_{syno} and X'_{syno}. Similarly, two opposite properties are also selected in pairs with a same base word, i.e., nodes in Y_{syno} and $\underline{Y}_{syno}$. Take the phrase sets in Figure 5 as an example. The subject "Asian countries" is selected together with the other subjects "African countries" and "European countries". Similarly, the properties "large amount of CO2 emissions" and "small among of CO2 emissions" are selected together with the base word "CO2 emissions". Therefore, when extracting a subgraph from the Phrase Similarity Graph, the nodes from X_{syno} and X'_{syno}, as well as the nodes from $Y_{base,syno}$, Y_{syno}, and $\underline{Y}_{syno}$ are selected in pairs to generate conditions for judgment.

As each subgraph G^k represents a group of subjects and properties with their semantic similarities, the aforementioned conditions from a subgraph can be simply generated by considering the node combinations constructed by each node and its neighboring nodes in $V^k_{subject} \cup V^k_{property}$. Given the node combinations, we design the conditions by comparing the relevance between the nodes of subjects and the nodes of properties. The conditions for judgment are designed following four criteria.

1) Nodes in X_{syno} are more relevant to nodes in Y_{syno} than nodes in $\underline{Y}_{syno}$;
2) Nodes in X'_{syno} are more relevant to nodes in $\underline{Y}_{syno}$ than nodes in Y_{syno};
3) Nodes in Y_{syno} are more relevant to nodes in X_{syno} than nodes in X'_{syno};
4) Nodes in $\underline{Y}_{syno}$ are more relevant to nodes in X'_{syno} than nodes in X_{syno}.

Following the previous work (Zhang et al. 2022), the relevance degree θ_{XY} between two nodes X and Y is defined as follows, which can be calculated by the edge weights in our graph.

$$\theta_{XY} = \frac{Sim(X,Y)}{Sim(X,Y_{base})} = \frac{\omega_{X,Y}}{\omega_{X,Y_{base}}}. \#(5)$$

Figure 6 shows an example of an extracted subgraph from the Phrase Similarity Graph in Figure 5 by selecting a group of nodes $X, X_1', X_2', Y_{base}, Y, \underline{Y}$. By selecting each node and its neighboring nodes except Y_{base} in the subgraph, conditions are generated by following the four criteria as follows.

$$X \cup N(X) = \{X, Y, \underline{Y}\} \rightarrow IF\ \theta_{XY} > \theta_{X\underline{Y}}, \#(6)\ X_1' \cup N(X_1') = \{X_1', Y, \underline{Y}\} \rightarrow IF\ \theta_{X_1'\underline{Y}} > \theta_{X_1'Y}, \#(7)\ X_2' \cup N(X_2') = \{X_2', Y, \underline{Y}\} \rightarrow IF\ \theta_{X_2'\underline{Y}} > \theta_{X_2'Y}, \#(8)\ Y \cup N(Y) = \{Y, X, X_1', X_2'\} \rightarrow \{IF\ \theta_{XY} > \theta_{X_1'Y} \#\ IF\ \theta_{XY} > \theta_{X_2'Y}, \#(9)\ \&\underline{Y} \cup N(\underline{Y}) = \{\underline{Y}, X, X_1', X_2'\} \rightarrow \{IF\ \theta_{X_1'\underline{Y}} > \theta_{X\underline{Y}}\ IF\ \theta_{X_2'\underline{Y}} > \theta_{X\underline{Y}}, \#(10)$$

where $N(X)$ represents the neighboring nodes of node X. Among the group of conditions from the subgraph, each satisfied condition increases the credibility of the explanation. We define a sub-score s_k to represent the proportion of satisfied conditions over all conditions from each subgraph G^k as follows.

$$s_k = \frac{the\ number\ of\ satisfied\ conditions\ in\ G^k}{the\ number\ of\ all\ conditions\ in\ G^k}. \#(11)$$

(Type XII) **Asia is the cause of the large amount of CO2 emissions.**
(Selected nodes from Phrase Similarity Graph)
X: Asian countries. X_1': African countries. X_2': European countries.
Y_{base}: CO2 emissions. Y: large CO2 emissions. $\underline{Y}$: small CO2 emissions.

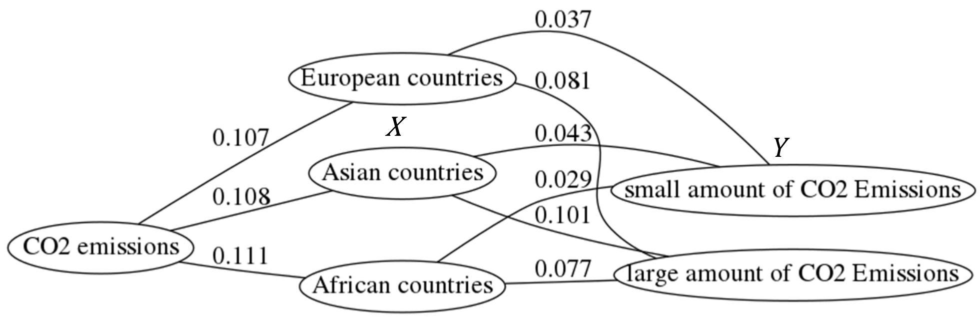


Figure 6. Example of a subgraph and its generated conditions for judgment.

Graph Entropy for Importance of Conditions

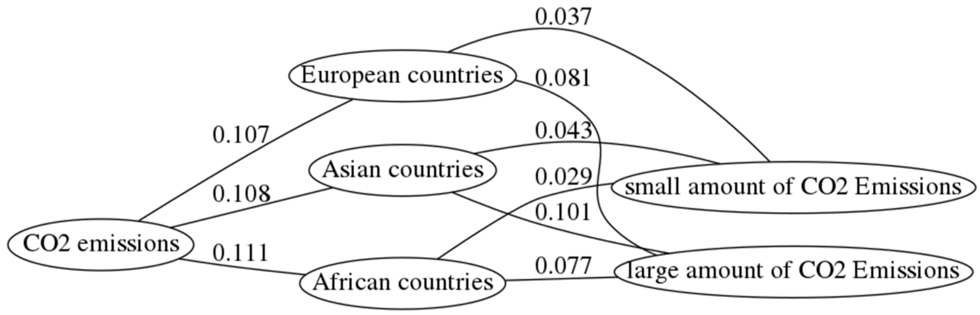
As we mentioned above, a group of conditions for judging an explanation is generated from each subgraph. Since the conditions for judgment are based on the diverse node attributes and edge weights from different subgraphs, they should be assigned different importance to judge an explanation. For example, Figure 7 shows two subgraphs G^1 and G^2 extracted from the Phrase Similarity Graph in Figure 5. As the nodes and the edge weights representing their semantic similarities are different in the two subgraphs, the semantic similarity-based conditions generated from G^1 and G^2 should have different importance for judgment.

Sen et al. (2018) utilize the sub-graph entropy based on edge weights to calculate the importance of subgraphs in functional brain networks. Following this work, we adopt the sub-graph entropy to quantify the importance of the generated conditions from each subgraph. In our approach, the edge weights refer to the semantic similarities between nodes, so the graph entropy measures the uncertainty of semantic similarities between nodes in a subgraph. A subgraph with high graph entropy indicates a greater uncertainty in the semantic similarities between its nodes, which suggests that the conditions generated from this subgraph should be assigned less importance. The graph entropy $H(G^k)$ of a subgraph G^k is negatively related to the importance of its generated conditions. To keep the weight value within the range of 0 to 1, we utilize the normalized exponential function of negative graph entropy $e^{-H(G^k)}$ as the weight λ_k to represent the importance of the conditions from the subgraph G^k.

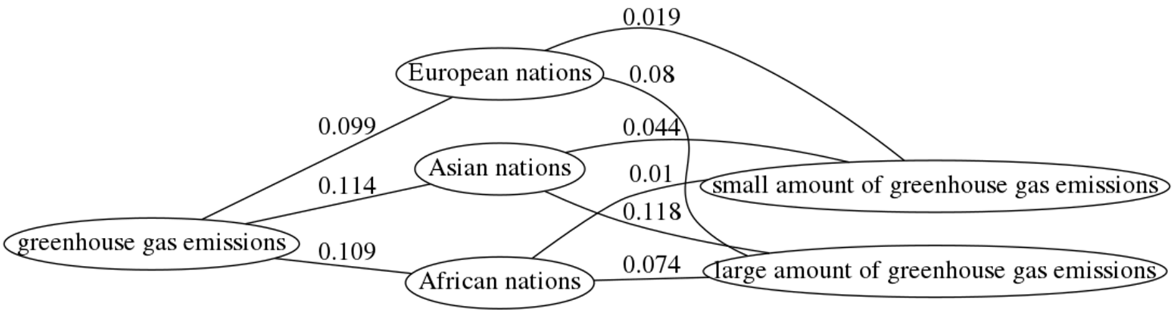
Following Sen et al. (2018), we adopt sub-graph entropy to measure the uncertainty of a subgraph within a whole graph. Sub-graph entropy is calculated by the normalized edge weights, which allows a fair comparison between subgraphs with different ranges of edge weights. Formally, given a subgraph $G^k = (V^k, E^k, W^k)$, its sub-graph entropy $H(G^k)$ is calculated as follows.

$$H(G^k) = -\sum_{i,j} p^k_{v_i,v_j} log p^k_{v_i,v_j}, \#(12)$$

$$where\ \ p^k_{v_i,v_j} = \frac{\omega^k_{v_i,v_j}}{\sum_{i,j} \omega^k_{v_i,v_j}}. \#(13)$$



(a) Subgraph G^1



(b) Subgraph G^2

Figure 7. Two subgraphs extracted from the Phrase Similarity Graph in Figure 5.

Take the two subgraphs G^1 and G^2 in Figure 7 as an example. Based on the normalized edge weights between nodes, the sub-graph entropy for G^1 and G^2 are calculated as $H(G^1) = 3.04$ and $H(G^2) = 2.93$, which indicates that G^2 has less uncertainty in the semantic similarities between its nodes, and thus conditions generated from G^2 should be assigned more importance. The weight λ_k representing the importance of the conditions generated from subgraph G^k is calculated by the normalized exponential function of negative sub-graph entropy $e^{-H(G^k)}$ as follows.

$$\lambda_k = \frac{e^{-H(G^k)}}{\sum_K^{k=1} \quad e^{-H(G^k)}} . \#(14)$$

Credibility Score of Explanation for Judgment

The credibility score of an explanation is defined by summing up all sub-scores and their corresponding weights, which are determined by the conditions generated from all subgraphs and their importance evaluated by sub-graph entropy. Given the sub-scores s_k and the weight λ_k of all subgraphs $\{G^k | \, k = 1, \ldots, K\}$, the credibility score S is calculated as follows.

$$S = \sum_K^{k=1} \quad \lambda_k s_k . \#(15)$$

The credibility score S ranges from 0 to 1 and a higher score indicates stronger credibility of the explanation. We define a use-supplied threshold $\theta_{credible}$ for our judgment method. The explanation is judged as credible and unethical if $S > \theta_{credible}$, else not.

Complexity Analysis

We analyze the time complexity of the proposed method β^2 when judging a statistical data explanation. Given a statistical data explanation, let m be the number of its phrases and we consider n synonyms for each phrase. The number of nodes in the Phrase Similarity Graph is mn. By considering the semantic similarities between nodes in different subsets to build edges, the time complexity of constructing a Phrase Similarity Graph is $O(mn^2)$. We propose to extract the subgraphs in the Phrase Similarity Graph by selecting nodes in groups from subjects and properties, respectively, so the time complexity for the extraction is $O(n^2)$. For each subgraph, the time complexities for generating conditions and calculating its graph entropy is $O(m^2)$ and $O(m)$, respectively. Therefore, the time complexity for judging an explanation based on the graph is $O(m^2n^2)$. To sum up, the overall time complexity for method β^2 is $O(m^2n^2)$. In our experiments, the values of m and n are less than ten and there are hundreds of explanations, which demonstrate that our method is fast and efficient for the target problem.

Experiments

In this section, we conduct experiments to evaluate the performance of the proposed method β^2. The experimental results are illustrated including a comparison of performance and detailed analysis.

Datasets

Our method is evaluated on statistical data explanations in (β) category. To conduct a more comprehensive evaluation, we have extended the dataset proposed by Zhang et al. (2022) by adding about 32% instances. The extended dataset contains 14 types of statistical data explanations within (β) category, where types II, IV-VII, XII, and XIV-XVIII are from Zhang et al. (2022) and we construct additional 3 types, i.e., XIX-XXI.

The 14 types of explanations describe 6 kinds of statistical data, including (B)-(G) in Preliminaries, which cover a wide range of topics. Specifically, type II and XVII involve the topics of health and economy, which explain the data of countries, life expectancy, and GDP per capita, as well as countries, continents, and GDP per capita, respectively. Type V and XIX involves the topics of education and collaboration, which explain the data of sex and math scores as well as countries and members of the United Nations, respectively. Type IV, VI, and VII involve the topic of children, which explain the data of babies and religions, babies and countries, as well as infant mortality rates and countries, respectively. Type XII, XIV, XX, and XXI involve the topic of energy, which explain the data of continents and CO2 emissions, countries and CO2 emissions, countries and mismanaged plastic waste, as well as countries and fossil fuel consumption, respectively.

Type XV, XVI, and XVIII involve the topic of health, which explain the data of hospitals and mortality rates, infected people and deaths from COVID-19 variants, as well as survival rates and diseases, respectively. The examples of the explanations accompanied with their corresponding statistical data and phrases have been introduced in Figures 2 and 3. The details, including the subject and the property, of each explanation is shown in Table 2.

The total number of the explanations is 122, consisting of 59 credible and unethical explanations and 63 not credible and unethical explanations. We settle on an approximate $50-50$ class balance in our experiments as it is the most difficult setting for a classification task. The ratio of the anomalies in the real world can vary. We avoid the problem of an arbitrary ratio of anomalies by this equal distribution setting. The ground-truth class labels of these explanations were manually assigned through a careful and consistent discussion among the authors (Zhang et al., 2022).

Experimental Setup

We utilize a large language model, ChatGPT with the released version named "ChatGPT Jan 9 Version" in 2023, to search for the top-n synonyms of each kind of phrase. Specifically, the top-n synonyms are obtained by utilizing the template "what are similar words to <phrase>" and selecting the top-n answers, where <phrase> is replaced by each phrase when searching for its synonyms. In our experiments, by investigating the qualities of the generated synonyms, n is set to 3. We notice that some of the synonyms of the proper nouns generated by ChatGPT are far from their original meanings, e.g., "East Asia" is generated as the synonym for "China". Therefore, we exclude the synonyms for phrases which are proper nouns, including countries and disease names in types II, XIV, XVIII, XX, and XXI. When generating phrase embeddings, we choose a Sentence-BERT model named "all-mpnet-base-v2"[3] trained on a large amount of data (more than 1 billion training pairs), which can map each phrase to a 768 dimensional dense vector. The credibility threshold $\theta_{credible}$ is set to 0.5.

Experimental Results and Analysis

The experimental results were obtained by measuring the agreement between the predicted class labels and the ground-truth class labels. Table 1 shows the confusion matrices of our method β^2 and the method β (Zhang et al., 2022) on the 14 types of statistical data explanations. Compared with method β, our method β^2 shows a significant improvement, which achieves an accuracy of 0.811. Due to the large number of false negatives in the results, the previous method β exhibits a relatively low accuracy, which is 0.574. In summary, our method β^2 significantly outperforms the baseline method β with about 0.237 improvement in accuracies. The results demonstrate the effectiveness of the proposed method β^2 for the target problem.

Table 1. Results by method β^2 compared with previous method β.

β	Predicted Positive	Predicted Negative	β^2	Predicted Positive	Predicted Negative
Actual Positive	20	39	Actual Positive	48	11
Actual Negative	13	50	Actual Negative	12	51

We show detailed results by method β^2 on 14 types of explanations in Table 2. In the Table, each explanation is represented by subject X with property Y in each row, where property Y in parentheses represents that the explanation belongs to class 0. For example, in type IV, X: Muslims and Y: many babies represent an explanation "Muslims have many babies compared to Christians" with class label 1. By replacing the property Y, X: Muslims and Y: (few babies) represent its variant "Muslims have few babies compared to Christians" with class label 0. FP and FN with bold fonts represent that the explanation is

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

judged as a false positive and a false negative, respectively. A blank in Result column either represents a true positive or a true negative.

Based on the results in Table 2, our method β^2 achieves almost perfect performance on 10 types of explanations, including types II (16-0), IV (6-0), V (7-1), VI (8-0), XII (6-2), XIV (7-1), XVIII (9-1), XIX (6-2), XX (16-0), and XXI (6-0)[4], where the numbers in parentheses represent correct and wrong predictions in this order. The previous method β obtains 31 false predictions on these 10 types, where 26 false predictions are caused by the counter-intuitive semantic similarities between subjects and properties. In contrast, our method β^2 yields 7 false predictions, with only 5 false predictions caused by this issue. This demonstrates that β^2 is capable of providing more accurate answers compared with the previous method β. Take an explanation of class 1 in type VI from Table 2 as an example, i.e., “Iranians have many children compared to Americans in the 21st century”. The previous method β obtains a false negative due to the counter-intuitive semantic similarities between “Iranians” and “many children” (0.285) and between “Iranians” and “few children” (0.294). While our method β^2 gives a correct answer of this explanation by considering the synonyms in the phrase sets, e.g., “Iranian nationals”, “many babies”, and “few babies”, which do not have counter-intuitive similarities. The 2 false predictions in type V and XIV are attributed to the fact that both two explanations describing one subject with two opposite properties are assigned class 0. The class labels of these two explanations reflect the subjectivity of persons, which is difficult to be estimated with no mistake.

On the other hand, our method β^2 achieves relatively low accuracies on 4 types, including VII (4-4), XV (0-4), XVI (6-4), XVII (4-4), by obtaining 16 false predictions, while the previous method β obtains 21 false predictions. There exist several explanations where our method β^2 fails while the previous method β succeeds. Take an explanation of class 0 in type XVII from Table 2 as an example, i.e., “Europe has lower GDP per capita than other regions”. The previous method β gives a correct prediction for it as the similarities between “Europe” and “low GDP” (0.300) is intuitively lower than the similarities between “Europe” and “high GDP” (0.369). On the other hand, our method β^2 yields a false positive because several synonyms in the phrase sets, e.g., “Europe”, “weak economy”, and “strong economy”, have counter-intuitive similarities. However, it is worth noting that our method β^2 achieves equal or higher accuracies on 12 types (106 explanations) while lower accuracies on only 2 types (16 explanations) compared with the previous method β. In addition, type XV poses a challenge because small hospitals are in general less well-equipped but receives fewer serious patients than large hospitals. Their degrees of safety are controversial, which might have influenced the phrase embeddings. Type XVI shows the difficulty in handling a serious issue related to the recent pandemic, which hasn’t been clarified scientifically and is a subject of a fierce debate. Omitting these two controversial types, our method β^2 can achieve an accuracy of 0.861 compared to the accuracy of 0.639 achieved by the previous method β. We believe these results show the performance of the two methods more appropriately.

We investigate the issue of the counter-intuitive semantic similarities between phrases in the results of the previous method β and our method β^2 under scrutiny. Take the explanation in Figure 2 as an example. The previous method β fails in judging it because the semantic similarities between “Muslims” and “many babies” (0.234) is counter-intuitively lower than the similarities between “Muslims” and “few babies” (0.245). In contrast, our method β^2 succeeds because the majority of the synonyms of “Muslims” exhibits higher semantic similarities to the synonyms of “many babies” compared to the synonyms of “few babies” in the extended phrase sets. For instance, Sim(“Muslims”, “many infants”) = 0.230, Sim(“Islam followers”, “many babies”) = 0.173, and Sim(“Islam followers”, “many kids”) = 0.195 are higher than Sim(“Muslims”, “few infants”) = 0.228, Sim(“Islam followers”, “few babies”) = 0.165, and Sim(“Islam followers”, “few kids”) = 0.177, respectively. The investigation suggests that the intuitive semantic similarities among the majority of the synonyms mitigate the problem of the counter-intuitive similarities between specific phrases, and thus helps our credibility score for accurate judgment.

[4] Our method 2 obtains 1 false positive and 1 false negative for the three new types XIX, XX and XXI. On the other hand, the previous method obtains 7 false negatives.

Table 2. Results and credibility score by method β^2, where the abbreviations MR, ER, CO2E, UNs, MPW, and PE represent mortality rates, enrollment rates, CO2 emissions, United Nations, mismanaged plastic waste, and plastic emissions, respectively.

Type	X	Y	Score	Result
II-1 8-0	Cuba	poorest	0.750	
		(richest)	0.250	
	Nicaragua	poorest	1.000	
		(richest)	0.000	
	Bangladesh	poorest	0.644	
		(richest)	0.356	
	North Korea	poorest	0.571	
		(richest)	0.429	
II-2 8-0	United Arab Emirates	richest	0.892	
		(poorest)	0.108	
	Qatar	richest	0.679	
		(poorest)	0.321	
	Equatorial Guinea	richest	0.785	
		(poorest)	0.215	
	Botswana	richest	0.536	
		(poorest)	0.464	
IV 6-0	Muslims	many babies	0.515	
		(few babies)	0.485	
	Judaisms	many babies	0.531	
		(few babies)	0.469	
	Christians	few babies	0.515	
		(many babies)	0.485	
V 7-1	women	low math score	0.527	
		(high math score)	0.473	
	men	high math score	0.527	
		(low math score)	0.473	
	women	(low English score)	0.395	
		high English score	0.605	
	men	(high English score)	0.395	
		(low English score)	0.605	**FP**
VI 8-0	Iranians	many children	0.583	
		(few children)	0.417	
	Afghans	many children	0.708	
		(few children)	0.292	
	French	few children	0.434	
		(many children)	0.566	
	Americans	few children	0.391	
		(many children)	0.609	
VII 4-4	developing countries	high infant MR	0.468	**FN**
		(low infant MR)	0.532	**FP**
	advanced countries	low infant MR	0.468	**FN**
		(high infant MR)	0.532	**FP**
	developing countries	low ER	0.718	
		(high ER)	0.282	
	advanced countries	high ER	0.718	
		(low ER)	0.282	
XII 4-2	Asia	large amount of CO2E	0.470	**FN**
		(small amount of CO2E)	0.530	**FP**
	Africa	small amount of CO2E	0.571	
		(large amount of CO2E)	0.429	
	Europe	small amount of CO2E	0.613	
		(large amount of CO2E)	0.387	
XIV 7-1	China	large amount of CO2E	0.750	
		(small amount of CO2E)	0.250	
	India	large amount of CO2E	0.686	
		(small amount of CO2E)	0.314	
	United States	large amount of CO2E	0.573	
		(small amount of CO2E)	0.427	
	United Kingdom	(large amount of CO2E)	0.251	
		(small amount of CO2E)	0.749	**FP**
XV	small hospitals	dangerous hospitals	0.354	**FN**

Type	X	Y	Score	Result
0-4	large hospitals	(safe hospitals)	0.646	**FP**
		safe hospitals	0.354	**FN**
		(dangerous hospitals)	0.646	**FP**
XVI 6-4	Omicron strain	less dangerous	0.512	
		(more dangerous)	0.488	
	Alpha strain	less dangerous	0.497	**FN**
		(more dangerous)	0.503	**FP**
	Beta strain	less dangerous	0.531	
		(more dangerous)	0.469	
	Gamma strain	less dangerous	0.483	**FN**
		(more dangerous)	0.517	**FP**
	Delta strain	more dangerous	0.512	
		(less dangerous)	0.488	
XVII 4-4	Africa	low GDP	0.795	
		(high GDP)	0.205	
	Asia	high GDP	0.600	
		(low GDP)	0.400	
	Americas	high GDP	0.481	**FN**
		(low GDP)	0.519	**FP**
	Europe	high GDP	0.424	**FN**
		(low GDP)	0.575	**FP**
XVIII 9-1	cancer	(long life expectancy)	0.371	
		short life expectancy	0.629	
	Alzheimer's disease	(long life expectancy)	0.500	
		short life expectancy	0.500	**FN**
	heart disease	(long life expectancy)	0.436	
		short life expectancy	0.564	
	pneumonia	(long life expectancy)	0.309	
		short life expectancy	0.691	
	periodontal disease	(short life expectancy)	0.326	
		long life expectancy	0.674	
XIX 6-2	Americas	many members of the UNs	0.513	
		(few members of the UNs)	0.487	
	Europe	many members of the UNs	0.515	
		(few members of the UNs)	0.485	
	Asia	many members of the UNs	0.438	**FN**
		(few members of the UNs)	0.562	**FP**
	Africa	few members of the UNs	0.530	
		(many members of the UNs)	0.470	
XX 16-0	India	large amount of MPW	0.576	
		(small amount of MPW)	0.424	
	China	large amount of MPW	0.781	
		(small amount of MPW)	0.219	
	United Kingdom	small amount of MPW	0.856	
		(large amount of MPW)	0.144	
	United States	small amount of MPW	0.536	
		(large amount of MPW)	0.464	
	India	large amount of PE	0.573	
		(small amount of PE)	0.427	
	China	large amount of PE	0.644	
		(small amount of PE)	0.356	
	United Kingdom	small amount of PE	0.713	
		(large amount of PE)	0.287	
	United States	small amount of PE	0.664	
		(large amount of PE)	0.336	
XXI 6-0	Australia	low fossil fuel consumption	0.750	
		(high fossil fuel consumption)	0.250	
	United Kingdom	low fossil fuel consumption	0.810	
		(high fossil fuel consumption)	0.190	
	United States	low fossil fuel consumption	0.750	
		(high fossil fuel consumption)	0.250	

Conclusion

In this paper we proposed a graph-based method to judge credible and unethical statistical data explanations. The Phrase Similarity Graph is constructed to explicitly model the phrases in phrase sets and their semantic similarities, where the sets are generated by considering synonyms of phrases specified from the explanation. Then the credibility score is devised by combining the conditions generated from the

Phrase Similarity Graph with their corresponding importance measured by sub-graph entropy. Experiments on 14 types of statistical data explanations demonstrate the effectiveness and superiority of the proposed method on the target problem compared with the baseline method.

We expect that this paper opens a new opportunity to bridge the gap between graph models and explanations of statistical data, enabling more effective judgement of the credibility of such explanations. As we mentioned in Target Problem, the significance of the statistical data explanations in (β) category has not been judged either in the previous method or in our method. Developing an objective measure of the significance is a challenging task due to the diverse individual perspectives and subjectivity. We plan to address this challenge in our future work. Another potential direction is to estimate more fine-grained credibility degrees of statistical data explanations using our proposed credibility score. However, for a fair evaluation, it is necessary to determine the ground truth of the credibility degrees by conducting cognitive experiments with a carefully designed approach.

Acknowledgements

A part of this work was supported by JSPS KAKENHI Grant Number JP21K19795. The first author is supported by China Scholarship Council (Grant No. 201906330075).

References

Balcerzak, B., Jaworski, W., & Wierzbicki, A. (2014). Application of TextRank algorithm for credibility assessment. *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, 1, 451-454.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web*, 675-684.

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., & Li, J. (2020). Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, 141-161.

Chen, W., Chen, L., Xie, Y., Cao, W., Gao, Y., & Feng, X. (2020). Multi-range attentive bicomponent graph convolutional network for traffic forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4), 3529-3536.

Cui, L., Seo, H., Tabar, M., Ma, F., Wang, S., & Lee, D. (2020). Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 492-502.

Czeisler, M. É., Wiley, J. F., Czeisler, C. A., Rajaratnam, S. M., & Howard, M. E. (2021). Uncovering survivorship bias in longitudinal mental health surveys during the COVID-19 pandemic. *Epidemiology and Psychiatric Sciences*, 30, e45.

Dehmer, M. (2008). Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation*, 201(1-2), 82-94.

Deng, J., Deng, Y., & Cheong, K. H. (2021). Combining conflicting evidence based on Pearson correlation coefficient and weighted graph. *International Journal of Intelligent Systems*, 36(12), 7443-7460.

Kazemi, A., Pérez-Rosas, V., & Mihalcea, R. (2020). Biased TextRank: Unsupervised graph-based content extraction. *Proceedings of the 28th International Conference on Computational Linguistics*, 1642-1652.

Kim, J., & Choi, K. S. (2020). Unsupervised fact checking by counter-weighted positive and negative evidential paths in a knowledge graph. *Proceedings of the 28th International Conference on Computational Linguistics*, 1677-1686.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). Mvae: Multimodal variational autoencoder for fake news detection. *Proceedings of 2019 World Wide Web Conference*, 2915-2921.

Lesser, L. M., & Nordenhaug, E. (2004). Ethical Statistics and Statistical Ethics: Making an Interdisciplinary Module, *Journal of Statistics Education*, 12(3).

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *The Lancet Digital Health*, 5(3), E105-E106.

Luo, G., Li, J., Su, J., Peng, H., Yang, C., Sun, L., Yu, P., & He, L. (2021). Graph entropy guided node embedding dimension selection for graph neural networks. *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2767-2774.

Mao, Q., Wang, Y., Yang, C., Du, L., Peng, H., Wu, J., Li, J., & Wang, Z. (2022). HiGIL: Hierarchical graph inference learning for fact checking. *Proceeding of the 2022 IEEE International Conference on Data Mining*, 329-337.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404-411.

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., & Martino, G. D. S. (2021). Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv*:2103.07769.

Noble, C. C., & Cook, D. J. (2003). Graph-based anomaly detection. *Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data mining*, 631-636.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931-2937.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3980-3990.

Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2), 76-81.

Rosling, H., Rosling, O., & Roennlund, A. R. (2018). Factfulness: Ten reasons we're wrong about the world - and why things are better than you think. Sceptre.

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797-806.

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16), 7662-7669.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3), 171-188.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36.

Thiese, M. S., Arnold, Z. C., & Walker, S. D. (2015). The misuse and abuse of statistics in biomedical research. *Biochemia Medica*, 25(1), 5-11.

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), 313-313.

Toivonen, H., Zhou, F., Hartikainen, A., & Hinkka, A. (2011). Compression of weighted graphs. *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data mining*, 965-973.

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2010). Selection bias and information bias in clinical research. *Nephron Clinical Practice*, 115(2), c94-c99.

Vedula, N., & Parthasarathy, S. (2021). Face-keg: Fact checking explained using knowledge graphs. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 526-534.

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 422-426.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., ... & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 849-857.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39-52.

Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. *31st IEEE International Conference on Data Engineering*, 651-662.

Zemčík, T. (2021). Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI & Society*, 36, 361-367.

Zeng, X., Abumansour, A. S., & Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10), e12438.

Zhang, K., Shinden, H., Mutsuro, T., & Suzuki, E. (2022). Judging instinct exploitation in statistical data explanations based on word embedding. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 867-879.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring AI ethics of ChatGPT: A diagnostic analysis. *arXiv preprint arXiv*:2301.12867.