Judging Instinct Exploitation in Statistical Data Explanations Based on Word Embedding

Anonymous Author(s)

ABSTRACT

This paper proposes 18 types of statistical data explanations and three kinds of procedures to investigate credibility in unethical and biased explanations due to exploitation of the 10 instincts proposed by Rosling et al. The explanation "men are better at math scores than women" accompanied with the averages and the distributions of their scores is an example of such an explanation, as it exploits the gap instinct, i.e., our tendency to divide all kinds of things into two distinct and often conflicting groups. It becomes much less credible if we replace the word "math" with "English", even if we keep the data as they are, as the exploitation seems less credible. Our judging procedures are based on phrase embedding and carefully designed comparisons to capture such an exploitation. The results of our experiments comparing the 18 types with their variants show promising results and clues for further developments.

KEYWORDS

instinct, word embedding, AI and ethics

ACM Reference Format:
Anonymous Author(s). 2022. Judging Instinct Exploitation in Statistical Data Explanations Based on Word Embedding. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As the impact and the presence of AI systems on our societies increase, their unethical misconducts are prone to severe reproach. The hijacking event of the chatbot Tay clearly shows that pure benevolence could turn into an opposite outcome [33]. Inflammatory tweets by a chatbot are often unethical, harms the reputation of its producer, and challenges the moral of our society. For the last issue, several tweets are more influential than others, as they are likely to be believed due to several reasons.

In this article, among such reasons, we tackle exploitation of human instincts in statistical data explanation. Rosling et al.'s book "Factfulness" has known a global success and emphasizes the importance of thinking based on facts and correct understandings [25]. The book includes examples of unethical and biased explanations each of which is denied by the accompanied statistical data. We, however, argue that such a thinking attitude is not always adopted and even accepted. Take as an example an explanation "men are better at math than women"[1] with chronological scores at SAT tests in the US and the score distributions of men and women in the 2016 test [25] in V in Figure 1. The latter statistical data clearly show the absurdness of discussing men and women in mass, as individually women or men who are good at math exist as those who are not. However, due to the gap instinct [25], i.e., our tendency to divide all kinds of things into two distinct and often conflicting groups, some portion of the public would believe the explanation, though the statistical data clearly refutes it. Such a situation deserves special attention as it highlights challenges to our rationality. In this paper, we are going to define 18 types of such credible unethical explanations each with its statistical data.

We also provide countermeasures to such explanations. Word embedding [11, 13, 19–22], which projects a word in a high dimensional space, keeping its semantics relative to other words, has known notable successes [2, 10, 29]. We leverage its extensions, phrase embedding [7, 23, 30], to three methods that judge whether an explanation is credible and unethical.

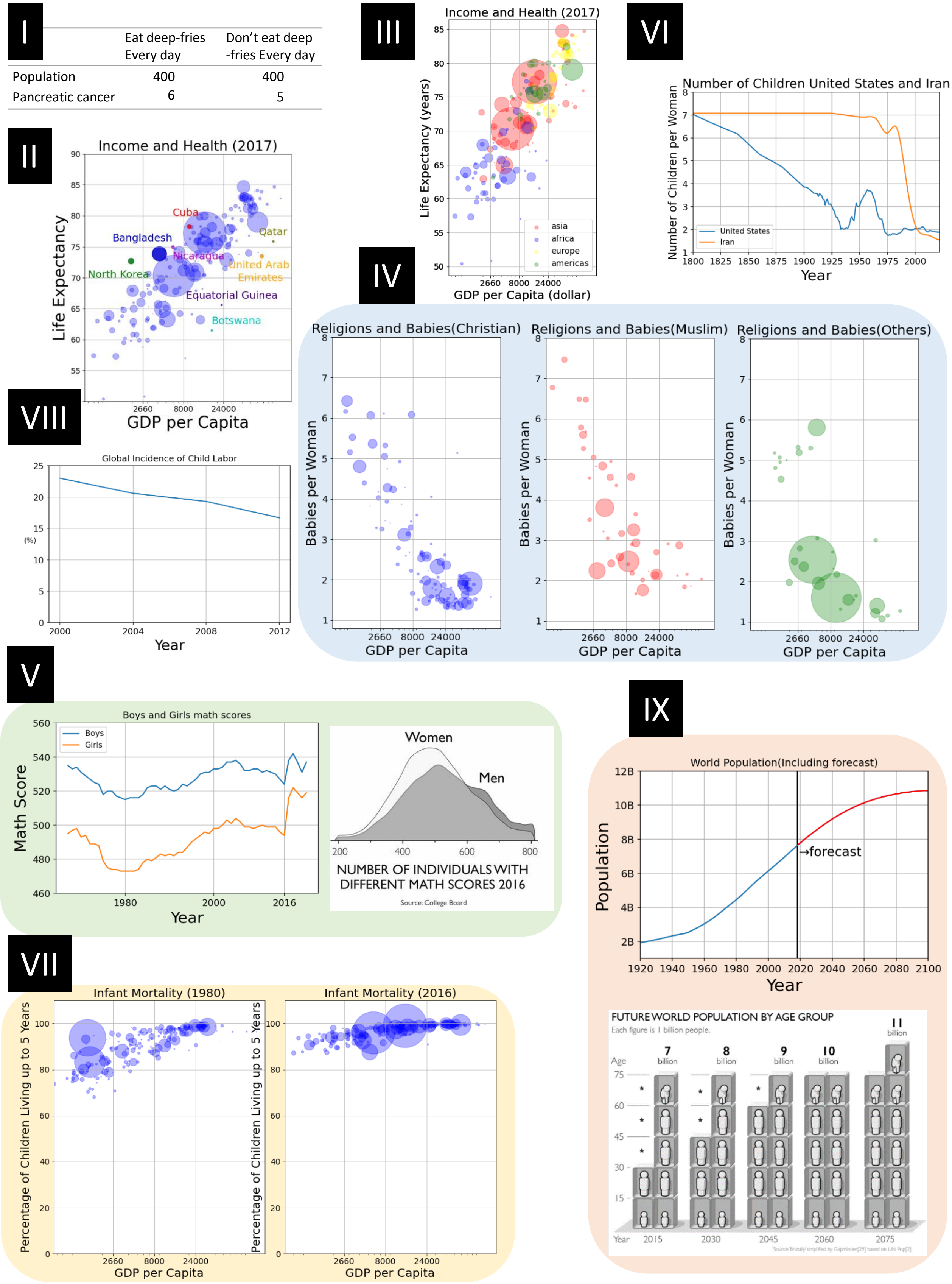
The organization of this paper is as follows. Section 2 reviews relevant works. We define the target problem in Section 3 and propose our method in Section 4. Section 5 tests the method by experiments and Section 6 concludes.

2 RELATED WORK

Unethical and biased explanations are widely generated in diverse fields around the world, such as fake news and hoaxes[1]. Detecting fake news is a challenging natural language processing (NLP) task involving two problems: characterization and detection [27]. Considering feature selection and extraction, Reis et al. [24] designed informative features, which consider semantic and syntactic properties, political biases, credibility, and environments of news, for automatic detection of fake news. Vlachos et al. [28] introduced fact checking tasks and discussed baseline approaches to assess truthfulness of explanations by measuring their semantic similarities. Detecting fake news is usually formulated as a classification task in a supervised manner [18, 32]. Through integrating meta data with texts, a hybrid Convolutional Neural Network (CNN) is devised to classify fake news based on surface-level linguistic

Conference'17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[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.



	Eat deep-fries Every day	Don't eat deep -fries Every day
Population	400	400
Pancreatic cancer	6	5

Figure 1: Statistical data in explanations I - IX. Data are adopted or modified from [25] or Gapminder [26].

patterns [31]. In this paper, we limit our attention to explanations of statistical data and focus on their unethical nature and credibility due to instinct exploitation.

As we explained above, our judging procedure is based on phrase embedding and carefully chosen comparisons to capture such exploitation. Measuring semantic similarity between various text components such as words, sentences, or documents has been explored in a wide range of downstream NLP tasks, such as machine translation [34], information retrieval [14] and question answering [17]. Li et al. [16] measured semantic similarity between words using multiple information sources, including the attributes path lengths, depths, and local densities in a hierarchical semantic knowledge base. To reduce the ambiguity in words, a robust semantic similarity measure [3] was devised by utilizing the information including page counts and lexico-syntactic patterns from text snippets of a Web search engine. Similar to [3], Normalized Google Distance (NGD) [8] was proposed to measure the similarity between two terms based on the results of Google search engine.

Semantic similarity methods have exploited the recent developments in neural networks and word embedding to enhance their performance [6]. In contrast to adopting traditional static word embedding [15, 20] for semantic similarity measurement between words [4], contextualized word embedding generated from modern neural language models, such as ELMo [21], GPT-2 [22], and BERT [11], has been widely employed for semantic similarity tasks [12]. The latter possesses over the former an advantage of capturing rich syntactic and semantic properties of words under diverse linguistic contexts. Moreover, for semantic similarity tasks between two sets of multiple words, such as phrases and sentences, InferSent [9] employes a bi-directional Long-Short Term Memory (LSTM) with a max-pooling operator as a sentence encoder to generate sentence embedding. Trained on a number of natural language prediction tasks, Universal Sentence Encoder [5] modeled the meaning of word sequences to encode sentences into high dimensional vectors. Sentence-BERT [23] adopted Siamese and triplet architectures based on the pre-trained BERT network to generate semantically meaningful embedding for sentences. Furthermore, the semantic similarities of sentences can be directly compared with cosine-similarity between their embedding.

To the best of our knowledge, no previous work tackles the problem of judging credible unethical explanations on statistical data by AI methods. This paper is the first work to define and investigate such explanations through AI techniques.

3 TARGET PROBLEM

3.1 Rosling's Ten Instincts

We focus our attention on Rosling et al.'s ten instincts [25], which are listed below. The ten instincts could be considered as innate, typically fixed patterns of human thinking.

(1) The gap instinct: our tendency to divide all kinds of things into two distinct and often conflicting groups, with an imagined, huge gap in between.
(2) The negativity instinct: our tendency to notice the bad more than the good.
(3) The straight line instinct: our tendency to believe that the increase is a straight line.
(4) The fear instinct: our tendency to focus our attention to what we are afraid of.
(5) The size instinct: our tendency to misjudge the size of things or the importance of a single number/instance.
(6) The generalization instinct: our tendency to categorize and generalize things all the time.
(7) The destiny instinct: our tendency to consider that several things never change due to their innate characteristics.
(8) The single perspective instinct: our tendency to prefer a single cause or solution.
(9) The blame instinct: our tendency to find a clear, simple reason for why something bad has happened.
(10) The urgency instinct: our tendency to want to take an immediate action in the face of a perceived imminent danger.

3.2 Credible Unethical Explanation of Statistical Data that Exploits the Instincts

We assume the following five conditions for our credible unethical explanation of statistical data.

(1) Data seems 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 instincts in Section 3.1.

Our target problem is to judge whether a given explanation is credible and unethical (class 1) or not (class 0). The types of the explanation are defined in the next section.

4 PROPOSED METHOD

4.1 Eighteen Types

The 18 types of explanations (I)-(XVIII) explain 7 kinds of statistical data. The data are (A) values of a probabilistic variable under 2 conditions, (B) a scatter plot of 2 probabilistic variables, (C) scatter plots or bubble charts 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) plots of a probabilistic variable and its average value (or frequency), and (G) a funnel plot. Examples of the statistical data are shown in Figures 1 and 2.

For each type of explanation, we code the exploited instincts and its statistical data. For example, in explanation (I), A-2 represents that the explanation exploits instinct 2 to explain the statistical data (A). In addition, we clarify why these explanations are not valid according to statistical

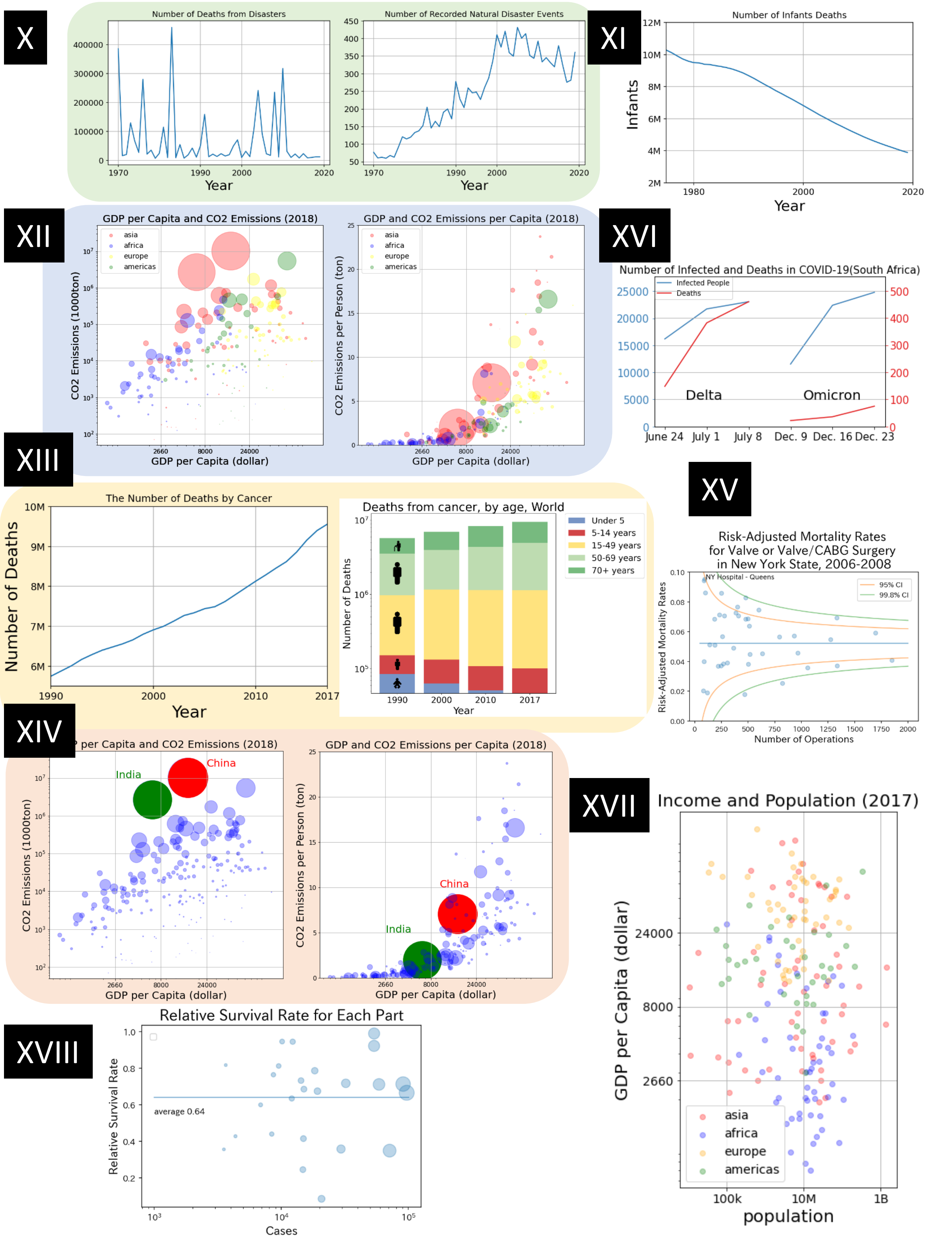


Figure 2: Statistical data in explanations X - XVIII. Data are adopted or modified from [25] or Gapminder [26].

data. Lastly, we provide candidate phrases for replacing the corresponded phrases.

(I) A-2, A-4: Deep-fried food boosts pancreatic cancer risk. $PhraseX$: deep-fried food. $PhraseY$: pancreatic cancer.

(Clarification) The relative risk of pancreatic cancer is only increased by 0.25%. Statistically testing the difference between the two groups will fail.

(Candidates for variants) $PhraseX$: alcohol abuse, heavy drinking, long-distance running. $PhraseY$: Alzheimer's disease, periodontal disease ,flu, alopecia areata, bone fracture, nosebleeds.

We have two variations for explanations II.
(II-1) B-2, B-8: Cuba is the poorest of the healthiest countries.
$PhraseX$: Cuba. $PhraseY$: the poorest of the healthiest countries.
(II-2) B-2, B-8: United Arab Emirates is the richest of the unhealthiest countries.
$PhraseX$: United Arab Emirates. $PhraseY$: the richest of the unhealthiest countries.

(Clarification) (II-1) Cuba is also the healthiest of the poorest countries. It is inappropriate to consider only one side.

(Candidates for variants) (II-1) $PhraseX$: Bangladesh, North Korea, Nicaragua. $PhraseY1$: richest. $PhraseY2$: unhealthiest.
(II-2) $PhraseX$: Qatar, Equatorial Guinea, Botswana. $PhraseY1$: poorest. $PhraseY2$: healthiest.

(III) B-3: Life expectancy continues to grow in proportion to GDP per capita.
$PhraseX$: life expectancy. $PhraseY$: proportional to GDP.

(Clarification) Note that the horizontal axis in the figure is a logarithmic scale, which is non-linear. Average life expectancy cannot grow without limit.

(Candidates for variants) $PhraseX$: healthy life expectancy. $PhraseY$: inversely proportional to GDP, not correlated to GDP.

(IV) C-1, C-6, C-7: Muslims have many babies compared to other religions.
$PhraseX$: Muslims. $PhraseY$: many babies.

(Clarification) All the 3 plots show that the number of babies decreases as the income increases, and there is no significant difference in the distribution. In fact, the average number of children per woman is 3.1 among Christians and 2.7 among Muslims.

(Candidates for variants) $PhraseX$: Judaism, Christian. $PhraseY$: few babies.

(V) D-1, D-2, D-6: Girls have lower math scores than boys. $PhraseX$: girl. $PhraseY$: low math score.

(Clarification) The left plot shows that girls have lower average scores than boys. The right plot shows that there exists an almost complete overlap between the two groups.

(Candidates for variants) $PhraseX$: boys.
$PhraseY$: high math score, low English score, high English score.

(VI) E-7, E-8: Iranians have many children compared to Americans in the 21st century.
$PhraseX$: Iranians. $PhraseY$: many children.

(Clarification) In the past centuries, Iranians had more children than Americans. In this century the two groups are similar in the number of children.

(Candidates for variants) $PhraseX$: Afghans, Americans, French. $PhraseY$: few children.

(VII) E-1, E-6, E-7, E-8: Infant mortality rates in developing countries are still significantly higher than in advanced countries.
$PhraseX$: developing country. $PhraseY$: high infant mortality rates.

(Clarification) The percentage of children living up to 5 years is now over 85% in most countries, and there is no significant difference between advanced and developing countries.

(Candidates for variants) $PhraseX$: advanced country.
$PhraseY$: low infant mortality rates, low enrollment rate, high enrollment rate.

(VIII) E-2, E-5: Child labor is about 15% and is not decreasing.
$PhraseX$: child labor. $PhraseY$: not decreasing.

(Clarification) The percentage of child labor is decreasing.

(Candidates for variants) $PhraseX$: child hunger, child mortality. $PhraseY$: increasing, decreasing, not increasing, constant.

(IX) E-3: The world's population will just increase.
$PhraseX$: world population. $PhraseY$: will just increase.

(Clarification) The right plot shows that the populations of younger generations are stable and those of older ones slowly increase. As the results, the population growth will be controlled.

(Candidates for variants) $PhraseY$: will rapidly increase, will just decrease, will rapidly decrease, will keep constant.

(X) E-4: Since year 2000, compared to 1980, there is an increasing in natural disasters and an increasing in deaths from natural disasters.
$PhraseX$: increasing in natural disasters. $PhraseY$: increasing in deaths from natural disasters.

(Clarification) The number of natural disasters is increasing, whereas the number of deaths from disasters is fluctuating and tends to decrease.

(Candidates for variants) From this type, we use {} as there are many candidates. $PhraseX$: {increasing in, decreasing in, constant} {natural disasters, epidemic damages, industrial accidents}. *PhraseY*: {increasing in, decreasing in, constant} deaths from {natural disasters, epidemic damages, industrial accidents}.

(XI) E-5, E-10: The death of many babies (4 million) is not decreasing.
PhraseX: death of many babies. *PhraseY*: not decreasing.

(Clarification) Nearly 10 million babies died 40 years ago, but recently the number has fallen to 4 million and the situation is improving.

(Candidates for variants) *PhraseX*: death of many {children, adults, old people}. *PhraseY*: increasing, decreasing, not increasing, constant.

(XII) F-1, F-6, F-7, F-8, F-9: Asia is the cause of the large amount of CO2 emissions.
PhraseX: Asia. *PhraseY*: large amount of CO2 emissions.

(Clarification) Asian countries seem to be the cause in the view of total emissions, which is denied by the per person emission view with respect to the GDP per capita.

(Candidates for variants) *PhraseX*: Africa, Americas, Europe.
PhraseY: CO2 emissions: {large, small} amount of {CO2, methane, freon gas} emissions.

(XIII) F-2, F-8: The risk of death from cancer is increasing worldwide.
PhraseX: risk of death from cancer. *PhraseY*: increasing.

(Clarification) The number of deaths from cancer is increasing, which is the result of the increase of the elderly in number.

(Candidates for variants) *PhraseX*: risk of death from {Alzheimer', periodonta} disease. *PhraseY*: decreasing, constant.

(XIV) F-8, F-9, F-10: China is the cause of the large amount of CO2 emissions. *PhraseX*: China. *PhraseY*: large amount of CO2 emissions.

(Clarification) A large population inevitably leads to an increase in CO2 emissions. In terms of CO2 emissions per person, the explanation is denied.

(Candidates for variants) *PhraseX*: United Kingdom, India, United States. *PhraseY*: small amount of CO2 emissions.

(XV) G-1, G-9: Small hospitals are dangerous hospitals. *PhraseX*: small hospital. *PhraseY*: dangerous hospital.

(Clarification) Funnel plot shows that most of the data points are within the confidence interval. Thus there is no such tendency.

(Candidates for variants) *PhraseX*: large hospital. *PhraseY*: safe hospital.

(XVI) A-1, A-6, A-7, A-8: Omicron strain of COVID-19 is less dangerous. *PhraseX*: Omicron strain. *PhraseY*: less dangerous.

(Clarification) Judging the dangerous degree of Omicron strain only by the number of deaths is inadequate. Omicron strain is more dangerous than Delta strain in the view of infections.

(Candidates for variants) *PhraseX*: Alpha strain, Beta strain, Delta strain, Gamma strain. *PhraseY*: more dangerous.

(XVII) B-1, B-6, B-7: Africa has lower GDP per capita than other regions.
PhraseX: Africa. *PhraseY*: low GDP.

(Clarification) Not all African countries have lower GDP per capita than other regions.

(Candidates for variants) *PhraseX*: Asia, Americas, Europe. *PhraseY*: high GDP.

(XVIII) G-6, G-7, G-8: The average 5-year survival rate for cancer is 64% so long life expectancy is expected.
PhraseX: cancer. *PhraseY*: long life expectancy.

(Clarification) The explanation is an overgeneralization because the survival rate for some dangerous cancers is less than 10%.

(Candidates for variants) *PhraseX*: Alzheimer's disease, periodontal disease, heart disease, pneumonia. *PhraseY*: short life expectancy.

4.2 Three Judgement Methods

We devise three detection methods (α), (β) and (γ) to assess the credibility of the 18 types of explanations. Detection method (α) is devised to judge explanation (I) on a habit *PhraseX* and a disease *PhraseY*. Detection method (β) is devised to judge explanations (II), (IV)-(VII), (XII), (XIV)-(XVIII), which describes the subject *PhraseX* has a property *PhraseY*. Detection method (γ) is devised to judge explanations (III), (VIII)-(XI), (XIII), which describes the subject *PhraseX* has a trend *PhraseY*.

The three methods all employ relevance degrees as the basis of their judgments. Each relevance degree is either a ratio of semantic similarities or a semantic similarity between a pair of phrases. The semantic similarity is a cosine-similarity of the embedded vectors of the phrases with SBERT [23]. Specifically the semantic similarity Sim($\cdot$) between *PhraseX* and *PhraseY* is computed by

$$\mathrm{Sim}(PhraseX,\ PhraseY) = \frac{s(PhraseX)\cdot s(PhraseY)}{\|s(PhraseX)\|\|s(PhraseY)\|}, \quad (1)$$

where $s(\cdot)$ represents the output embedding by SBERT.

4.2.1 *Detection method* (α).

This method judges an explanation as credible unethical if and only if the habit *PhraseX* is bad, the disease *PhraseY* is dangerous, and the two are highly relevant.

$$\text{IF } (\theta_{\text{relevance}} > \theta_1) \wedge (\theta_{\text{fear}} > \theta_2) \wedge (\theta_{\text{bad habit}} > \theta_3) \text{ THEN } 1 \text{ ELSE } 0, \quad (2)$$

where $\theta_1, \theta_2, \theta_3$ are user-supplied thresholds and $\theta_{\text{relevance}}$, θ_{fear}, $\theta_{\text{bad habit}}$ are the relevance ratio, the fear ratio, and the

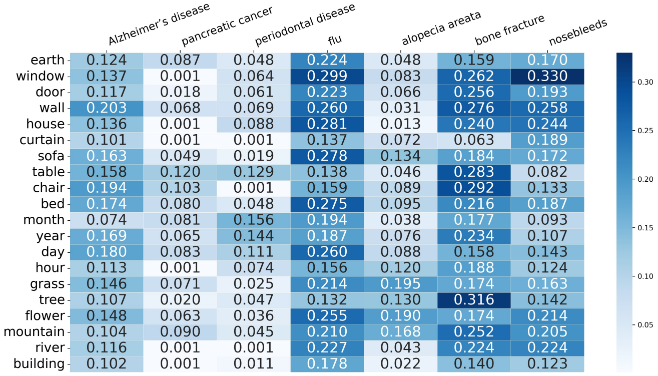


Figure 3: Relevance degrees between diseases and base words.

bad habit ratio, respectively.

$$\theta_{\text{relevance}} = \frac{\text{Sim}(PhraseX,\ PhraseY)}{\text{Sim}(PhraseY,\ \text{base word})}, \tag{3}$$

$$\theta_{\text{fear}} = \frac{\text{Sim}(PhraseY,\ \text{“}major\ illness''\text{)}}{\text{Sim}(PhraseY,\ \text{“}minor\ illness''\text{)}}, \tag{4}$$

$$\theta_{\text{bad habit}} = \frac{\text{Sim}(PhraseX,\ \text{“}bad\ habit''\text{)}}{\text{Sim}(PhraseX,\ \text{“}good\ habit''\text{)}} \tag{5}$$

For instance, explanation (I) is judged credible and unethical because eating deep-fried food is a bad habit, pancreatic cancer is a dangerous disease, and the two seem to be relevant each other.

In Eq. (3)-(5), the base word should be neutral with any habit and disease. We conducted a preliminary experiments on the relevance degrees between a disease and such a wordl The results are shown in Figure 3. We see that “day” and “earth” have steady degrees to all the tested diseases. We select the former as the base word.

Another series of preliminary experiments proved that Eq. (4) shows quite counter-intuitive results, probably due to our highly-variable subjectivity in assessing the risk of diseases. Thus we use a summary (GBD Cause nd Risk Summaries in https://www.thelancet.com/gbd/summaries) of DALYs (Disability Adjusted Life Years) in Burden of Disease (https://ourworldindat.org/burden-of-disease) as θ_{fear}. DALYs measures the loss in health quantitatively, allowing us to compare different diseases and other kinds of damages.

4.2.2 Detection method (β). This method judges an explanation as credible and unethical if and only if the relevance degree θ_{XY} between the subject $PhraseX$ and the property $PhraseY$ is larger than the relevance degree $\theta_{X\overline{Y}}$ between $PhraseX$ and the inverse property $Phrase\overline{Y}$, and the relevance degrees $\theta_{X^{'}Y}$ between similar subjects $PhraseX^{'}$ and $PhraseY$.

$$\begin{aligned}&\text{IF } (\theta_{XY} > \theta_{X\overline{Y}}) \wedge \forall X^{'}(\theta_{XY} > \theta_{X^{'}Y}) \text{ THEN } 1\\&\text{ELSE } 0\end{aligned} \tag{6}$$

Explanations II-1 and II-2 each has two properties $PhraseY1$ and $PhraseY2$ and thus the above procedure is applied to each of them, by first replacing Y with $Y1$ and then Y with $Y2$. The relevance degrees are defined as follows.

$$\begin{aligned}\theta_{XY} &= \frac{\text{Sim}(PhraseX, PhraseY)}{\text{Sim}(PhraseX, PhraseY_{\text{base}})}\\ \theta_{X\overline{Y}} &= \frac{\text{Sim}(PhraseX, Phrase\overline{Y})}{\text{Sim}(PhraseX, PhraseY_{\text{base}})}\\ \theta_{X^{'}Y} &= \frac{\text{Sim}(PhraseX^{'}, PhraseY)}{\text{Sim}(PhraseX', PhraseY_{\text{base}})}\end{aligned} \tag{7}$$

For instance, explanation (XVII) is judged credible and unethical because the above conditions are satisfied for $PhraseX$: Africa, $PhraseY$: low GDP, $Phrase\overline{Y}$: high GDP, $PhraseY_{\text{base}}$: GDP, and $PhraseX^{'} \in$ {Asia, Americas, Europe}.

Note that $PhraseY_{\text{base}}$ serves as a base in comparison. For (II-1), we set $PhraseY_{1base}$: poor and $PhraseY_{2base}$: healthy. $PhraseY^{'}_{1base}$: rich and $PhraseY^{'}_{2base}$: unhealthy. For other kinds of explanations, $PhraseY_{\text{base}}$ is equivalent to $PhraseY^{'}_{\text{base}}$ due to the single property. They are “babies” for (IV), “math score” for (V), “children” for (VI), “infant mortality rate” for (VII), “CO2 emissions” for (XII) and (XIV), “hospital” for (XV), “dangerous” for (XVI), “GDB” for (XVII), and “life expectancy” for (XVIII).

4.2.3 Detection method (γ). This method judges an explanation as credible and unethical if and only if the relevance degree between the subject $PhraseX$ and the trend $PhraseY$ is larger than the relevance degree between similar subjects $PhraseX^{'}$ and $PhraseY$.

$$\begin{aligned}&\text{IF } \forall X^{'}(\theta_{XY} > \theta_{X^{'}Y}) \text{ THEN } 1\\&\text{ELSE } 0\end{aligned} \tag{8}$$

The relevance degrees are defined as follows.

$$\begin{aligned}\theta_{XY} &= \text{Sim}(PhraseX, PhraseY)\\ \theta_{XY^{'}} &= \text{Sim}(PhraseX, PhraseY^{'})\end{aligned} \tag{9}$$

For instance, explanation (XIII) is judged credible and unethical because the above conditions are satisfied for $PhraseX$: risk of death from cancer, $PhraseY$: increasing, and $PhraseY^{'} \in$ {decreasing, constant}.

5 EXPERIMENTS

We choose an SBERT model named “all-mpnet-base-v2” trained on a large amount of data (more than 1 billion training pairs) which can map each phrase to a 768 dimensional dense vector. In detection method α, the thresholds $\theta_1, \theta_2, \theta_3$ are all set to 1.

The accuracies of our detection methods (α), (β) and (γ) on the 18 types of explanations are 0.893, 0.519, and 0.688, respectively. We see that (α) and (γ) exhibit relatively high accuracies, probably due to the simpler forms of their target explanations. Their confusions matrices are shown in Table 1. The Tables show clues for further improvements such as investing the three and five false positive of (α) and (γ), respectively. Though (β) needs a substantial refinement, we

Table 1: Results of (α), (β) and (γ)

(α)	Predicted Positive	Predicted Negative	(β)	Predicted Positive	Predicted Negative	(γ)	Predicted Positive	Predicted Negative
Actual Positive	6	3	Actual Positive	10	17	Actual Positive	11	5
Actual Negative	0	19	Actual Negative	9	18	Actual Negative	0	0

believe that the results are quite promising as the first step toward judging credible unethical explanations of statistical data.

Tables 2 and 3 show the detailed results for α and γ, respectively. They deserve detailed analyses unlike β. A pair of parenthesis represents that the corresponding explanation is not a credible unethical one.

In summary, the relevance degree defined by semantic similarity exhibits encouraging performance on judging the credibility of the explanations on statistical data. The underlying semantic relatedness between phrases is worth exploring in the next step.

6 CONCLUSIONS

In this paper we have investigated exploitation of ten instincts in statistical data explanations as a first yet important step toward ethical AI. Our goal is not in abusing our investigation but to prevent such unethical conducts through deep understanding. Our 18 prototypes together with their variants, our three kinds of judgement method, and our experiments serve as a milestone toward the goal.

This paper opens promising avenues for further research. Beyond the judgement methods, neutralizing credible unethical explanations through transformation represents a challenging and yet important problem. A fully automatic generation of unethical explanations will be the next step, though our earlier investigations with generative deep neural networks knew limited success. Large-scale cognitive experiments on the credibility of the variants of each explanation is definitively highly rewarding. Targeting at the right population is the key to success, though the authors feel that our communities lack diversity in this respect. Web services such as Amazon Mechanical Turk provide a powerful solution, though a careful design is mandatory.

Table 2: Results by method (α).

Type	PhraseX	PhraseY	$\theta_{relevance}$	θ_{fear}	θ_{bad_habit}	result	Type	PhraseX	PhraseY	$\theta_{relevance}$	θ_{fear}	θ_{bad_habit}	result
(I)	deep-fried food	pancreatic cancer	2.309	11.5	2.070		(I)	heavy drinking	pancreatic cancer	1.624	11.5	1.528	
		Alzheimer's disease	1.968	25.3	2.070				Alzheimer's disease	3.557	25.3	1.528	
		periodontal disease	1.715	7.09	2.070				periodontal disease	2.965	7.09	1.528	
		flu	0.941	6.39	2.070				flu	0.868	6.39	1.528	
		(alopecia areata)	2.736	0.60	2.070				(alopecia areata)	3.159	0.60	1.528	
		(bone fracture)	1.374	0.00	2.070				(bone fracture)	1.893	0.00	1.528	
		(nosebleeds)	1.363	0.00	2.070				(nosebleeds)	1.962	0.00	1.528	
	alcohol abuse	pancreatic cancer	1.921	11.5	1.755			(long distance running)	pancreatic cancer	0.048	11.5	0.922	
		Alzheimer's disease	4.211	25.3	1.755				Alzheimer's disease	0.509	25.3	0.922	
		periodontal disease	4.055	7.09	1.755				periodontal disease	0.021	7.09	0.922	
		flu	0.693	6.39	1.755				flu	0.448	6.39	0.922	
		(alopecia areata)	4.522	0.60	1.755				(alopecia areata)	1.415	0.60	0.922	
		(bone fracture)	1.852	0.00	1.755				(bone fracture)	1.509	0.00	0.922	
		(nosebleeds)	1.933	0.00	1.755				(nosebleeds)	1.262	0.00	0.922	

Table 3: Results by method (γ).

Type	PhraseX	PhraseY	$\theta_{relevance}$	result	Type	PhraseX	PhraseY	$\theta_{relevance}$	result
(III)	life expectancy	proportional to GDP	0.145		(X)	(constant epidemic damages)	increasing in deaths from epidemic damages	0.811	
		(inversely proportional to GDP)	0.033				(decreasing in deaths from epidemic damages)	0.748	
		(not correlated to GDP)	0.079				(constant deaths from epidemic damages)	0.915	
	healthy life expectancy	proportional to GDP	0.168			increasing in industrial accidents	increasing in deaths from industrial accidents	0.913	
		(inversely proportional to GDP)	0.059				(decreasing in deaths from industrial accidents)	0.877	
		(not correlated to GDP)	0.150				(constant deaths from industrial accidents)	0.759	
(VIII)	child labor	not decreasing	0.061	FN		(decreasing in industrial accidents)	increasing in deaths from industrial accidents	0.811	
		(decreasing)	0.062				(decreasing in deaths from industrial accidents)	0.888	
		(not increasing)	0.090				(constant deaths from industrial accidents)	0.706	
		(increasing)	0.111	FP		(constant industrial accidents)	increasing in deaths from industrial accidents	0.751	
		(constant)	0.110				(decreasing in deaths from industrial accidents)	0.712	
	child hunger	not decreasing	0.146	FN			(constant deaths from industrial accidents)	0.857	
		(decreasing)	0.148		(XI)	death of many babies	increasing	0.133	
		(not increasing)	0.187				(decreasing)	0.114	
		(increasing)	0.177				(not increasing)	0.129	
		(constant)	0.188	FP			(not decreasing)	0.112	
	child mortality	not decreasing	0.195	FN			(constant)	0.064	
		(decreasing)	0.206			death of many children	increasing	0.129	
		(not increasing)	0.207				(decreasing)	0.102	
		(increasing)	0.217	FP			(not increasing)	0.116	
		(constant)	0.126				(not decreasing)	0.092	
(IX)	world population	will just increase	0.203	FN			(constant)	0.076	
		(will rapidly increase)	0.221	FP		death of many adults	increasing	0.169	
		(will just decrease)	0.130				(decreasing)	0.114	
		(will rapidly decrease)	0.127				(not increasing)	0.158	
		(will keep constant)	0.210				(not decreasing)	0.141	
(X)	increasing in natural disasters	increasing in deaths from natural disasters	0.876	FN			(constant)	0.052	
		(decreasing in deaths from natural disasters)	0.819			death of many old people	increasing	0.156	
		(constant deaths from natural disasters)	0.665				(decreasing)	0.122	
	(decreasing in natural disaster)	increasing in deaths from natural disasters	0.775				(not increasing)	0.138	
		(decreasing in deaths from natural disasters)	0.882	FP			(not decreasing)	0.130	
		(constant deaths from natural disasters)	0.639				(constant)	0.031	
	(constant natural disaster)	increasing in deaths from natural disasters	0.646		(XIII)	risk of death from cancer	increasing	0.082	
		(decreasing in deaths from natural disasters)	0.577				(decreasing)	0.033	
		(constant deaths from natural disasters)	0.838				(constant)	0.001	
	increasing in epidemic damages	increasing in deaths from epidemic damages	0.939			risk of death from Alzheimer's disease	increasing	0.064	
		(decreasing in deaths from epidemic damages)	0.840				(decreasing)	0.018	
		(constant deaths from epidemic damages)	0.785				(constant)	0.001	
	(decreasing in epidemic damages)	increasing in deaths from epidemic damages	0.866			risk of death from heart disease	increasing	0.095	
		(decreasing in deaths from epidemic damages)	0.936				(decreasing)	0.063	
		(constant deaths from epidemic damages)	0.754				(constant)	0.006	

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[2] Mandelbaum Amit and Shalev Adi. 2016. Word embeddings and their use in sentence classification tasks. arXiv:1610.08229

[3] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Measuring semantic similarity between words using web search engines. *www* 7, 2007 (2007), 757–766.

[4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[5] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. arXiv:1803.11175

[6] Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—A survey. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–37.

[7] Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training Tasks for Embedding-based Large-scale Retrieval. In *Proceedings of the 8th International Conference on Learning Representations*. Ethiopia.

[8] Rudi L Cilibrasi and Paul MB Vitanyi. 2007. The google similarity distance. *IEEE Transactions on knowledge and data engineering* 19, 3 (2007), 370–383.

[9] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 670–680.

[10] Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised Sequence Learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Montreal, Canada, 3079–3087.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805

[12] Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 122–133.

[13] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. 1986. Distributed Representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA, USA, 77–109.

[14] Sun Kim, Nicolas Fiorini, W John Wilbur, and Zhiyong Lu. 2017. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. *Journal of biomedical informatics* 75 (2017), 122–127.

[15] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning*. PMLR, Beijing, 1188–1196.

[16] Yuhua Li, Zuhair A Bandar, and David McLean. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering* 15, 4 (2003), 871–882.

[17] Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria, and Eneko Agirre. 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems* 119 (2017), 186–199.

[18] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. ACM, Melbourne, VIC, Australia, 1751–1754.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781

[20] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. ACL, 1532–1543.

[21] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv:1802.05365

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[23] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Hong Kong, China, 3980–3990.

[24] Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems* 34, 2 (2019), 76–81.

[25] Hans Rosling, Ola Rosling, and Anna Rosling Roennlund. 2018. *Factfulness: Ten Reasons We'Re Wrong About The World - And Why Things Are Better Than You Think*. Sceptre.

[26] Ola Rosling, Anna Rosling Rönnlund, and Hans Rosling. 2005. Gapminder Download the data. https://www.gapminder.org/data/.

[27] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[28] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*. 18–22.

[29] Peng Wang, Bo Xu, Jiaming Xu, Guanhua Tian, Cheng-Lin Liu, and Hongwei Hao. 2016. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing* 174 (2016), 806–814.

[30] Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Dominican Republic, 10837–10851.

[31] William Yang Wang. 2017. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada, 422–426.

[32] Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, Marina Del Rey, CA, USA, 637–645.

[33] Tomáš Zemčík. 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, 1 (2021), 361–367.

[34] Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, Seattle, Washington, USA, 1393–1398.