Bias in COVID Disinformation

Truong Pham
Nupur Singh
Department of Computer Science
Illinois Institute of Technology

November 30, 2021

Abstract

Bias is deeply rooted in human culture and it affects how we think and interpret the world. A person may rely on their biases to form their thoughts and communicate them, thus they can unknowingly inject those biases into their language. When interpreting language data, Machine Learning methods will also learn those biases as they can only understand words by reading them in context. Using this theory, we investigate the bias that can form in fake news articles, which are sometimes made to support popular biases, as opposing to real news articles, which are more based on facts.

1 Problem Statements

1.1 Bias in Languages

Bias is a inclination to view or project feelings on a particular subject. Bias can be considered as a projection of our thoughts and consciousness into the world to fill them with meanings and explanations.

AI is becoming part of every industry, and it is boosting production, increasing efficiency and speeding up the decision process in every aspect of life. As human, though we can use all our senses (see, hear, listen, touch, taste and smell) to communicate but majority (90%) of the time we use language as a communication medium, and this makes NLP a vital part of AI. Verbal communication is one of the most powerful means through which discriminations are perpetrated and reproduced, and writing is and was the most potent way verbal communication was spread across time and space. Let's consider that language itself is Neutral and unbiased, then the problem arises with the imbalance in the way language represents data. Many a time its not intentional e.g. Relating 'She' with 'Nurse' and 'He' with 'Technician' in documents depicting the society 50 years back is actually a fact. It was the time when majority of female were found in the occupation of Nurse while male were found in Technology. The writer had no intention of being sexist but still the data holds it. But though unintentional if this data pass to the Word-Embeddings it will learn that the pair 'she ' and 'nurse' is more prominent than 'he' and 'nurse', and thus the word 'nurse' will be accommodated in the feminine space of Word-Embeddings.

The default assumption is that computation, deriving from mathematics, would be pure and neutral, providing AI a fairness beyond what is present in human society. Instead, concerns about machine prejudice are now coming to the fore—concerns that our historic

biases and prejudices are being refeed in machines. As We keep trying to build AI systems which act as close to humans the question arise is, isn't it prone to come with the fault humans has? If we consider Word-Embeddings as a digital translation of human thoughts which are reflected in literature, it is bound to have all good and bad qualities human possesses.

There are many techniques introduced to overcome this, like

- diversity among AI developers, to address insensitive or under-informed training of machine learning algorithms, and
- collaboration between engineers and domain experts who are knowledgeable about historical inequalities etc .

While all of these strategies might be helpful and even necessary, they could not be sufficient. Machine prejudice gets derived so fundamentally from human culture that it is not possible to eliminate it through strategies such as the above. The fact that it is rooted in language makes prejudice difficult to address. We propose prejudice must be addressed as a component of any intelligent system learning from our culture. It cannot be entirely eliminated from the system, but rather must be compensated for. Detecting bias and being aware of its presence provide human and machine an opportunity to device solution to deal with it.

1.2 Covid Disinformation

Novel coronavirus (SARS-CoV-2), also known as COVID-19, has claimed over half a million of lives world wide in the span of over a year. The severity of this pandemic is comparable to the flu pandemic in 1918. COVID-19 possesses the ability to infest new victims at an incredible speed, which leads to high number of cases world wide without effective treatments being found. This has created an opportunity for misinformation to quickly spread in social media, as well as traditional media, which led the World Health Organization (WHO) to warn of an on-going "infodemic" or an overabundance of information—especially misinformation regarding the pandemic.

Misinformation about COVID-19 has proliferated widely on social media, ranging from the peddling of fake cures such as gargling with lemon or salt water and injecting yourself with bleach to false conspiracy theories that the virus was bioengineered in a lab in Wuhan or that the 5G cellular network is causing or exacerbating symptoms of COVID-19. These misinformation might seem blatantly false but some fake news can even come from sources that can be considered informed with some scientific evident to justify them. For example, President Trump and Brazilian President Jair Bolsonaro had falsely claimed that hydroxychloroquine is a treatment for COVID-19. Although the harms and benefits of hydroxychloroquine as a potential treatment are indeed being studied, there is currently no scientific consensus on its effectiveness. Therefore, deciding whether or not a piece of information is true could be a complicated mater.

To aid the fight against harmful misinformation during the pandemics, we try to gain a better perspective on the semantics of COVID-19 fake news language. Based on the fact that human language contains implicit biases regarding specific topics [1], we extend the research to find the biases that fake news possess when the topic is about COVID-19. Fake news, by definition, cannot be based on concrete facts from scientific researches. Therefore, the spread of fake news must rely on the fallible understandings of its readers. Just like how

prejudice is integrated into our language through the influence of our cultures, the erroneous information that we willing accept also comes from the wrongly conceived intuition that we formed regarding on how the world operates. Prejudice and false intuition are just two sides of the coin called bias. Based on the success regarding the extraction of cultural biases in language, we hope to also extract the biases that are ingrained in misinformation.For example the virus being detected in Wuhan, China made the news call the Corona Virus as 'Chinese virus'. Across the social media a asian bias has being detected even in the fake news.

"If you get corona virus from **Chinese food** the simple cure is to gargle bleach."

"Good news **Wuhan's corona virus** can be cured by one bowl of freshly boiled garlic water. Old Chinese doctor has proven its efficacy. "

Figure 1: The above posts are sample from our data set.

These clearly shows that in time of Pandemic people are associating the virus with anything related to China like "getting the virus from Chinese food". Intentional or not these data if fed to word embedding it will associate the Virus as Asian/Chinese/Wuhan. In our project we have tried ti bring out the Asian bias in our dataset (consisting of twitter feeds on Covid-19).

2 Proposed solution

2.1 Covid Fake News Dataset

The fake and real news contexts in for this experiment are taken from the MM-COVID dataset which contains multi-lingual articles about Covid-19 news crawled from multiple sources that labeled them fake and news along with the social media interaction regarding those articles from Twitter [2]. In the scope of this experiment, we will only be using the English articles.

The dataset got their labels from from the fact-checking websites, and then retrieve the source content from these websites. The labels are collected from Snopes and Poynter where the domain expert and journalists review the information and provide the factchecking evaluation results as fake or real. Additional news pieces with the COVID-19, Coronavirus and SARS-CoV-2 are also collected from several official health websites. Social media like Facebook, Twitter, Instagram, etc, and blogs and traditional news are also used as sources for the news articles.

2.2 Word2Vec

Word embeddings in NLP taks used to be treated as atomic units where there is no notion of similarity between words. The work of Mikolov, et al. [3] revolutionized embedding representation of words by giving them multiple degree of similarity. Similar words are closer to each other like *France* and *Italy* since they are both names of countries. Even more so, the words *biggest* is close to *big* the same way smallest is close to *small*, enabling us to

use queries on the embeddings like: What is the word that is similar to *small* in the same sense as *biggest* is similar to *big*? The most surprising result of the paper is that we can do calculations on the embeddings. For example, we have 4 embeddings for 4 words *King*, *Queen*, *man*, *woman*, then we can perform calculations on those embeddings to get from King to Queen: *vector(King) - vector(man) + vector(woman) = vector(Queen)*. For this project, we are using the Genism Python package to run the model in the paper

After the revolution in Word2Vec, more models with increasing efficiency of the same type were born. One of them is the Global Vectors for Word Representation (GloVe) model which ultilize the coocurrence matrix of words to achieve state of the art results in NLP tasks [4]. This is also the model of choice in the original bias in language paper [1].

2.3 Bias Test

To measure bias in language, we use the distances between the embeddings which contains the relative meanings between the words when the embeddings are extracted using the Glove and Genism-W2V. The test is called Word Embedding Association Test (WEAT) [1] which measure the statistics based on the distance of nouns and their attributes to measure bias. Consider two sets of target words (e.g., programmer, engineer, scientist, ... and nurse, teacher, librarian, ...) and two sets of attribute words (e.g., man, male, ... and woman, female ...). The null hypothesis is that there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words. The permutation test measures the (un)likelihood of the null hypothesis by computing the probability that a random permutation of the attribute words would produce the observed (or greater) difference in sample means.

In formal terms, let X and Y be two sets of target words of equal size, and A,B the two sets of attribute words. Let $\cos(\vec{a}, \vec{b})$ denote the cosine of the angle between the vectors$\vec{a}$ and $\vec{b}$.

- The test statistic is:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

  where

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} cos(\vec{w}, \vec{b})$$

- The effect size is:

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{std - dev_{w \in X \bigcup Y} s(w, A, B)}$$

This test only indicates bias and more research must be done before deciding if the bias in question exists like the IAT.

3 Implementation details

3.1 Data processing

The dataset downloaded was a huge file of 1.21gb. It was a json file but the content was formatted as JSON Lines. JSON Lines is a convenient format for storing structured data

that may be processed one record at a time. It works well with unix-style text processing tools and shell pipelines. But in our case parsing this was challenging as it was not proper json. As Json lines are basically a valid json structure in each line i.e. like having multiple json in one file, the option was to read each line at a time as text from the file and then parse it as a json and either work with it as is or generate a new file with the needed fields. This was cumbersome. We wanted a format that allow us to read and parse the complete file at once. To achieve this we tried two options , first was writing the content to a new file and save with the file extension '.jsonl' and secondly format the existing file content to a parse able json format.The second approach worked for us.

To convert a JSON Lines file like "¡anyfile¿.jsonl" to one-line valid JSON, you only need to add a [at the beginning and a] at the end, and replace all but the last newline with a comma. In our case we already had the comma present among the each line so we just added a [at the beginning and a] at the end. We can approach to do this

- Can use ordinary line-processing tools, like AWK to do this.
  awk 'BEGIN {printf "["} {printf("%s,",$0)} END {print $0 "]"}'
- Use Java/JavaScript/python to rewrite the file with the starting and ending square braces.
- Simply open the file in text editor and include the starting and ending square braces and save the file.(For our case this was little troublesome as the file was very huge 1.21gb)

The original format of the faulty data is in Figure 2 while the fixed format is in Figure 3.


```
{
    "_id":{"$oid":"5f8911702a4301b0368d10f9"},
    "agency": "webMD",
    "claim":"Coronavirus Is a Breeding Ground for Conspiracy Theories",
    "label":"real",
    "lang":"en",
    "news_id":"webMD-2102578004"
},

{
  "_id":{"$oid":"5f8911702a4301b0368d112a"},
  "news_id":"webMD-9632391802",
  "agency":"webMD",
  "claim":"Worldwide Number of COVID-19 Cases Over 1 Million",
  "label":"real",
  "lang":"en",
  "news_id":"webMD-2102578004"
}
```


Figure 2: Original File

Since most of the fields in the dataset are in dictionary form, we must process them in json format to extract meaningful information from them. In the end, all extracted data is in string or date-time form if their information is provided in their data entry. Otherwise, the missing data is set to be *np.nan* or special float. The texts of the articles required further parsing to remove the hyperlinks, special characters, and punctuation, and we also set all the words into lowercase. However, we predict that we might need commas and periods in the future to tokenize the setences and train the emebeddings with higher accuracy so we create two separate sets of article data, one with commas and periods, and one without any commas and periods.

```
[{
    "_id":{"$oid":"5f8911702a4301b0368d10f9"},
    "agency": "webMD",
    "claim":"Coronavirus Is a Breeding Ground for Conspiracy Theories",
    "label":"real",
    "lang":"en",
    "news_id":"webMD-2102578004"
  },

  {
    "_id":{"$oid":"5f8911702a4301b0368d112a"},
    "news_id":"webMD-9632391802",
    "agency":"webMD",
    "claim":"Worldwide Number of COVID-19 Cases Over 1 Million",
    "label":"real",
    "lang":"en",
    "news_id":"webMD-2102578004"
  }]
```


Figure 3: Formatted File with starting and ending square braces

Even though the dataset has a field for the language of the articles, the labeling is not entirely correct. Therefore, we need to use Natural Language Processing packages, *langdetect* and *nltk*, to find English articles which are mutually classified as English articles by the two packages. First we use *langdetect.detect* function to classify the language of the articles and retain only the English articles. After that, we run *nltk.wordpunct_tokenize* on the articles and check for the existance of those tokens in the English corpus of nltk in lowercase form. If only one token in an article is not in the nltk lowercase English corpus, we classify the article as not English and remove it from the dataset. This hard condition is necessary to ensure that we will train the embeddings on all English words. We gathered 3764 real articles and 1186 fake articles, the number of resulting articles can vary -+5 articles each class since *langdetect.detect* is non deterministic.

3.2 Embeddings

The embedding are critical for the implementation of this experiment. The original paper uses pre-trained Glove embedding to test for bias in the context of general culture opposing from our double language contexts of fake and real news. Therefore, we must inject fake and real news contexts into the word embeddings, creating two separate word embedding sets for fake and real news.

We use 2 self supervised models to learn the word embeddings Genism-Word2Vec and Glove since they are both very popular models. For the Genism-Word2Vec model, we can finetune the model on the pretrained embeddings. We train a set of 50 dimensional embeddings on all words that appear in the dataset. The window for the context words is set to 7 and we train for 100 epochs. For the Glove model, we use a window of size 15 to calculate the concurrence matrix to train for 15 epochs with the model, resulting in a set of 50 dimensional embeddings for every words that appear at least 5 times in the dataset. The whole process is run using the official Glove github (**For simplification, the resulted embeddings trained from Glove is moved to the designated files for real and fake embeddings so we can easily access them when we use WEAT**)

To test our Embedding we used the vector generated by Genism-Word2Vec for predicting the authenticity of the news in data. We ran two tests :

- 1st We generated the input vectors using Genism-Word2Vec trained on the complete dataset and

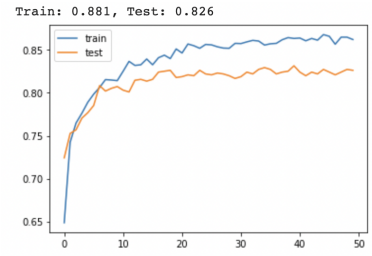

Figure 4: Genism-Word2Vec trained with complete data

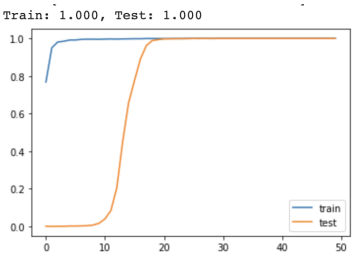

Figure 5: Genism-Word2Vec Trained on only real data

- 2nd We generated the input vectors using Genism-Word2Vec trained on only the real news data-subset.

Both the models gave pretty good result though we had not very large number of words to train from.

Figure 5 shows the accuracy of the one trained on real data. Our test data was of length 950 out of which 712 are real. Hence vectors from the embedding trained on dataset having only real news gave 100% accuracy. This test helps us to represent the $\frac{\partial W}{\partial X}$ i.e. change in word embedding due to change in the document/data. As the data have different context the feature vector for each word have different values.

3.3 Word Embedding Association Test (WEAT)

To implement WEAT, we must account for both fake and real news contexts when calculating the bias as oppose to a singular context. Therefore, we devise a new way to set up the target and attribute words when they switch between fake and real news like in Figure 6. Each set of target words will be accompanied by 2 sets of attribute words just like in the original paper. However, since we want to find the bias when a subject is talked about in fake news rather than real news, each set of words will have 2 sets of word embeddings which correspond to their appearances in fake and real news. In total, instead of 4 sets of word embeddings for 4 sets of words for each bias test, we have 6 sets of words for 3 sets of words. The way to choose the null distribution will also be changed. Since the bias of question is the switch of semantics between fake and real news, we must be sure that the differences that arise in the relationship between the target words and their attributes arise from bias and not because of the context switch. Therefore, the null distribution must represent the regular change in the target-attribute relationship in the context switch. We can do this by choosing random words to be the target words while maintaining the attribute words.

In new WEAT we have 2 statistics:

1. The statistic:

$$s(X_{real}, X_{fake}, A_{real}, B_{real}, A_{fake}, B_{fake}) = \sum_{x \in X_{real}} s(X_{real}, A_{real}, B_{real}) - \sum_{x \in X_{fake}} s(X_{fake}, A_{fake}, B_{fake})$$

Where: X_{real}: the real news embeddings of targets
X_{fake}: the fake news embeddings of targets
A_{real}: the real news embeddings of attributes 1
B_{real}: the fake news embeddings of attributes 1
A_{fake}: the real news embeddings of attributes 2
B_{fake}: the fake news embeddings of attributes 2

2. The effect-size:

$$\frac{\text{mean}_{x \in X_{real}} s(x, A_{real}, B_{real}) - \text{mean}_{x \in X_{fake}} s(x, A_{fake}, B_{fake})}{std - dev_{w \in X_{null}} s(x, A, B)}$$

The way the words are chosen are very adhoc. For the attributes, I simply choose words that exist in both fake and real news and have a close relationship with the attribute they represent (not necessary the target), although these attribute words exist in the same article with the target words. The attributes are chosen represent the biases against and for the targets.

For Asian-Health bias:

- Targets = asia, asian, wuhan, china, chinese
- Attributes 1 = danger, harm, risk, infectious
- Attributes 2 = healthy, safe, asymptomatic

For Vaccine-AntiVaccine bias:

- Targets = vaccine
- Attributes 1 = harm, danger, dangerous, risk, distress
- Attributes 2 = safe, effective, useful, healthy, great

4 Results and discussion

We choose to test for 2 biases in the fake and real news which are Asian-Health bias and Vaccine-AntiVaccine bias and the results are shown in Table 1. We can see that the embed-

Embeddings	Bias	p-value	effect-size
Glove	Asian-Health	< 0.05	0.135
	Vaccine-AntiVaccine	< 0.05	-2.0
W2V	Asian-Health	< 0.05	0.098
	Vaccine-AntiVaccine	0.4-0.6	-2.0

Table 1: Results of WEAT

dings of both model agrees on the existence of the Asian-Health bias, suggesting that there is a change in attitude when asian related targets are being mentioned in fake as opposed to real news regarding health problems. Even though the pandemic affects everyone in the

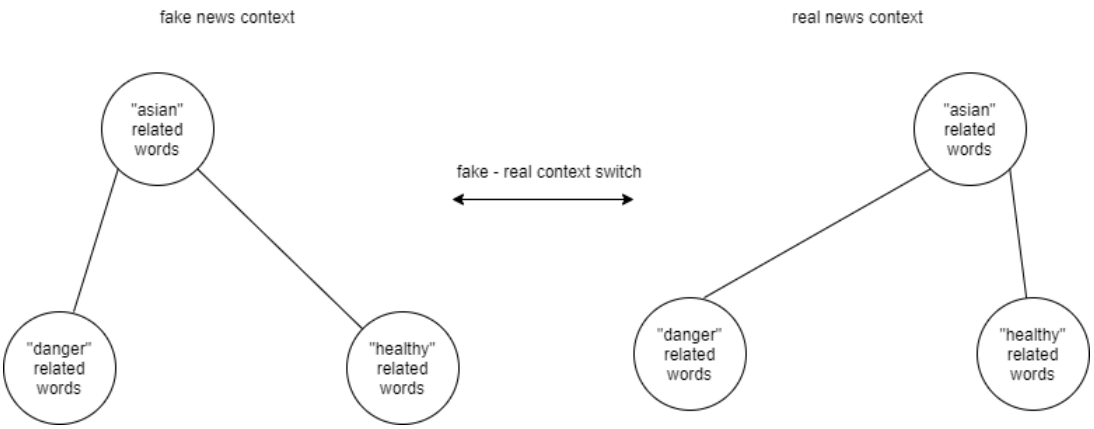

Figure 6: A pictorial example of what we hope to find. This is not an accurate representation of a real bias. The change in distance can also be only because of different context and not because of bias.

world, fake news might try to paint Asian people in a more dangerous way or as unreasonable health hazard. This can also explain why there is only a small effect size when it comes to this bias since the differences in opinion might be very small but noticeable. We can clearly see from Figure 7 that the words that belong in the same set of words are closer together which partly support our decision in selecting them for their respective word set.

The second bias that we test for is the Vaccine-AntiVaccine bias which is a much debated medical topic before the pandemic started. However, only the Glove embeddings accept the bias while the W2V embeddings completely fails to reject the null hypothesis. The date of the news articles in our dataset are from 2020, when the vaccine for COVID-19 was still in the distant future. This can explain why there are much divisive when vaccine is mentioned in fake and real news since the vaccine topic was not as sensational in 2020 as current time in 2021. The embeddings also cluster together according to their word set as we can see in Figure 8.

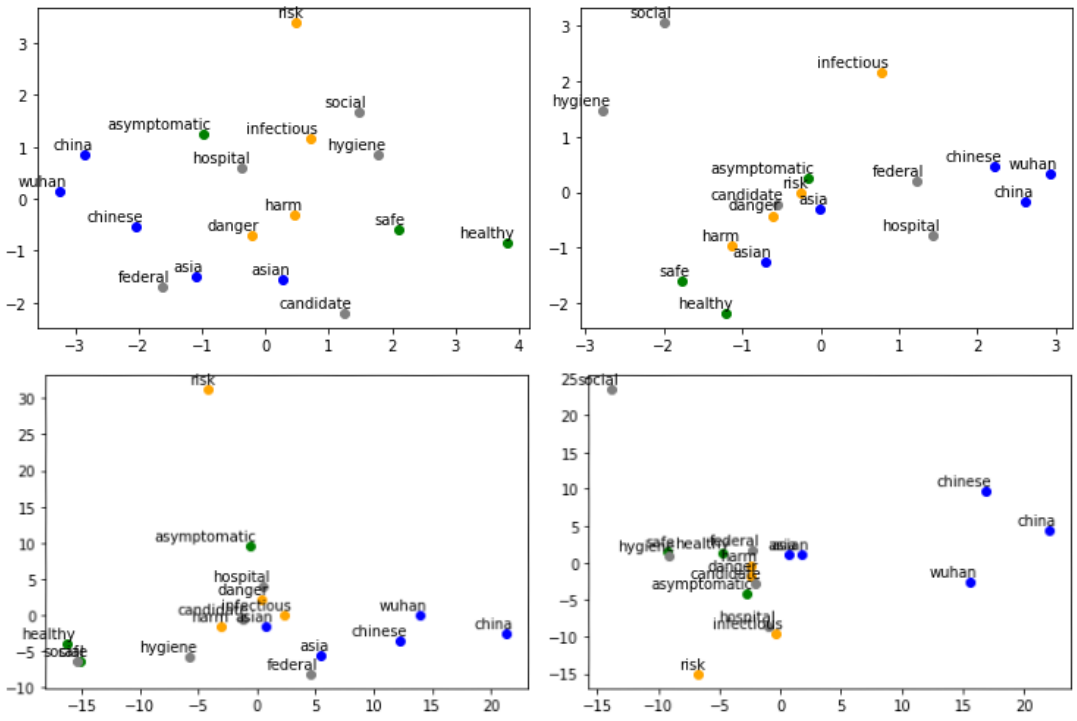

Figure 7: The embeddings of the target and and attributes of the Asian-Health bias arranged from left to right, top to bottom: Glove embeddings of real news, Glove embeddings of fake news, Genism-Word2Vec embeddings of real news, W2V embeddinsg of fake news

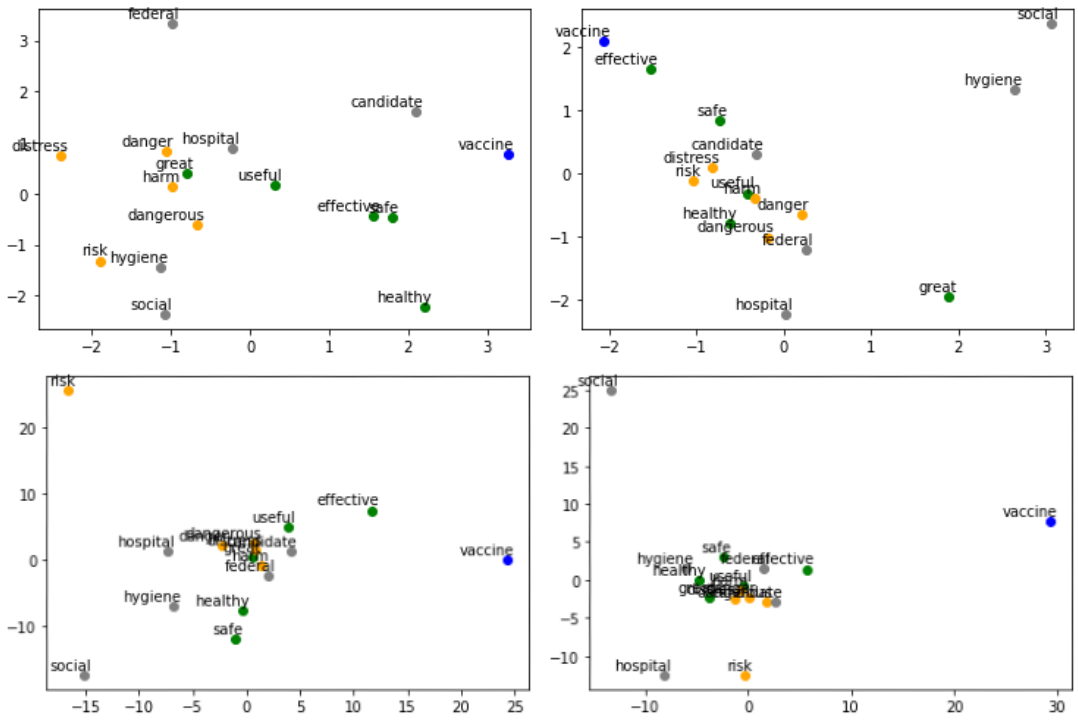

Figure 8: The embeddings of the target and and attributes of the Vaccine-AntiVaccine bias arranged from left to right, top to bottom: Glove embeddings of real news, Glove embeddings of fake news, Genism-Word2Vec embeddings of real news, W2V embeddinsg of fake news

References

[1] A. C. Islam, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora necessarily contain human biases," *CoRR*, vol. abs/1608.07187, 2016. [Online]. Available: http://arxiv.org/abs/1608.07187

[2] Y. Li, B. Jiang, K. Shu, and H. Liu, "MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation," *CoRR*, vol. abs/2011.04088, 2020. [Online]. Available: https://arxiv.org/abs/2011.04088

[3] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162