Volume 7, No. 1 | January–February 2024

The Journal of Robotics, Artificial Intelligence & Law

fastcase FULL COURT PRESS

RAIL
The Journal of Robotics, Artificial Intelligence & Law
Volume 7, No. 1 | January–February 2024

THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW (ISSN 2575-5633 (print) /ISSN 2575-5617 (online) at $495.00 annually is published six times per year by Full Court Press, a Fastcase, Inc., imprint. For customer support, please contact Fastcase, Inc., 729 15th Street, NW, Suite 500, Washington, D.C. 20005, 202.999.4777 (phone), or email customer service at support@fastcase.com.

Publishing Staff
Publisher: Morgan Morrissette Wright
Production Editor: Sharon D. Ray
Cover Art Design: Juan Bustamante

Cite this publication as:

The Journal of Robotics, Artificial Intelligence & Law (Fastcase)

A Full Court Press, Fastcase, Inc., Publication

Editorial Office

729 15th Street, NW, Suite 500, Washington, D.C. 20005
https://www.fastcase.com/

POSTMASTER: Send address changes to THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW, 729 15th Street, NW, Suite 500, Washington, D.C. 20005.

Articles and Submissions

Direct editorial inquiries and send material for publication to:

Steven A. Meyerowitz, Editor-in-Chief, Meyerowitz Communications Inc., 26910 Grand Central Parkway, #18R, Floral Park, NY 11005, smeyerowitz@meyerowitzcommunications.com, 631.291.5541.

Material for publication is welcomed—articles, decisions, or other items of interest to attorneys and law firms, in-house counsel, corporate compliance officers, government agencies and their counsel, senior business executives, scientists, engineers, and anyone interested in the law governing artificial intelligence and robotics. This publication is designed to be accurate and authoritative, but neither the publisher nor the authors are rendering legal, accounting, or other professional services in this publication. If legal or other expert advice is desired, retain the services of an appropriate professional. The articles and columns reflect only the present considerations and views of the authors and do not necessarily reflect those of the firms or organizations with which they are affiliated, any of the former or present clients of the authors or their firms or organizations, or the editors or publisher.

QUESTIONS ABOUT THIS PUBLICATION?

For questions about the Editorial Content appearing in these volumes or reprint permission, please contact:

Morgan Morrissette Wright, Publisher, Full Court Press at morgan.wright@vlex.com or at 202.999.4878

For questions or Sales and Customer Service:

Customer Service
Available 8 a.m.–8 p.m. Eastern Time
866.773.2782 (phone)
support@fastcase.com (email)

Sales
202.999.4777 (phone)
sales@fastcase.com (email)

ISSN 2575-5633 (print)
ISSN 2575-5617 (online)

Unpacking Averages: Searching for Bias in Word Embeddings Trained on Food and Drug Administration Regulatory Documents

Bradley Merrill Thompson*

In this article, the author explores how bias can creep into word embeddings by analyzing a model trained on what regulatory affairs professionals in industry and the Food and Drug Administration have written.

Often when we talk about bias in word embeddings, we are talking about such things as bias against race or sex. This article talks about bias a little bit more generally to explore attitudes we have that are manifest in the words we use about any number of topics.

Bias Evaluation Using Sentiment Analysis

There are many different ways to evaluate potential bias in word embeddings, but I did not want to do a survey article where I talked briefly about all of them. Instead, I thought I would pick just one approach for illustration. The one I picked is perhaps the simplest, which is an evaluation of the word embeddings using a model for positive versus negative sentiment. In other words, I am looking to see whether particular word embeddings have a largely positive or negative connotation.

If words that should be regarded similarly have significantly different sentiments or connotations, that would be evidence of bias. In other words, if the word "Black" as an adjective for people has a largely negative connotation while the word "white" as an adjective for people has a largely positive connotation, that would be some evidence that the embeddings, trained on what people have written, have absorbed from that training data a bias against Black people.

Robotics, Artificial Intelligence & Law / January–February 2024, Vol. 7, No. 1, pp. 9–18.

ISSN 2575-5633 (print) /ISSN 2575-5617 (online)

However, I am not going to use race as my example in the analysis below. For one thing, race is rarely discussed in the documents that I am going to examine—Food and Drug Administration (FDA) documents—apart from a handful of documents specifically on race. I will leave you to draw your own conclusions from that. Instead, I am going to look for bias in other topics.

Methodology

I wanted to keep it simple, so I will not use any of the cool, sophisticated, but complicated techniques that recently have been developed. Instead, for this first foray into the topic, I am going to use a methodology that has been around for a while because it is relatively simple to understand. In fact, I am shamelessly mimicking an approach used by Robyn Speer in her July 2017 piece entitled "How to Make a Racist AI Without Really Trying."[1] I have updated the methodology only slightly to account for changes in software libraries since she published her article.

The basic approach is to train an algorithm—specifically a classifier—to recognize the differences between positive and negative words. In my particular case, I chose to use a random forest classifier from sklearn because it has shown to be effective in this sort of analysis.

This is an exercise in supervised learning, meaning that the algorithm needs to be trained by being told which words are positive and which words are negative in a training set. For that, I used a list of words that have been labeled as either positive or negative that most researchers in this space use, data created by Hu and Liu which are available from Bing Liu's website.[2]

Here is the theoretical part. To recap, I have thousands of words that researchers have labeled as positive and thousands of words that researchers have labeled as negative. I also have thousands of these word embeddings through my prior machine learning work defined by 300 dimensions. The idea is that one or more of those 300 dimensions might correspond to positive versus negative connotations of the word. Thus, in theory, if I take the positive and negative words labeled as such from Hu and Liu, and if I represent those words using my 300-dimension word embeddings, I can train a machine learning classifier to spot positive versus negative words using the embeddings that I created from training on FDA

regulatory documents. In other words, the algorithm can learn which of the 300 dimensions in the word2vec model I created correspond to positive versus negative sentiment.

So that is what I did.

Validation

I like to document the uncertainty of any algorithm I use. I withheld about 10% of the data so I could test the algorithm on labeled data to see how well it did. I went into this assuming that I could perhaps get 90-95% accuracy from this exercise because that is what Robyn Speer in her original article achieved. But I could not. The best I could do was approaching 80% accuracy.

I spent a fair amount of time, for example, using grid search to experiment with different hyperparameters, and I also experimented with adding additional data for training purposes from other data sets. Oddly enough, adding more training data caused the performance to go down. Ultimately, I concluded that this was about the best I could do.

If you ask me why my performance was lower than what Robyn Speer achieved, she was analyzing some common word embeddings developed from training data such as Google News. In comparison, my training data were scientific and regulatory documents. While there are many differences between those data sources, at a high level I would say that regulatory professionals use fewer adjectives and adverbs in their writing. But adjectives and adverbs are the food of sentiment analysis. Without adjectives and adverbs, the algorithm has far less to go on in categorizing words as positive or negative. A sentence "the results were a score of 16" just doesn't give the algorithm much to go on as to whether those words are positive or negative.

Or maybe I am just bad at it. But in any event, not quite 80% accuracy was the best I could do. Keep that in mind.

Validation Through Visual Exploration

Another way to validate the appropriateness of the algorithm that is little bit less scientific and relies much more on the anecdotal and the visual. I wanted to see how the results look, so I decided to assess an entire FDA guidance document on this negative versus

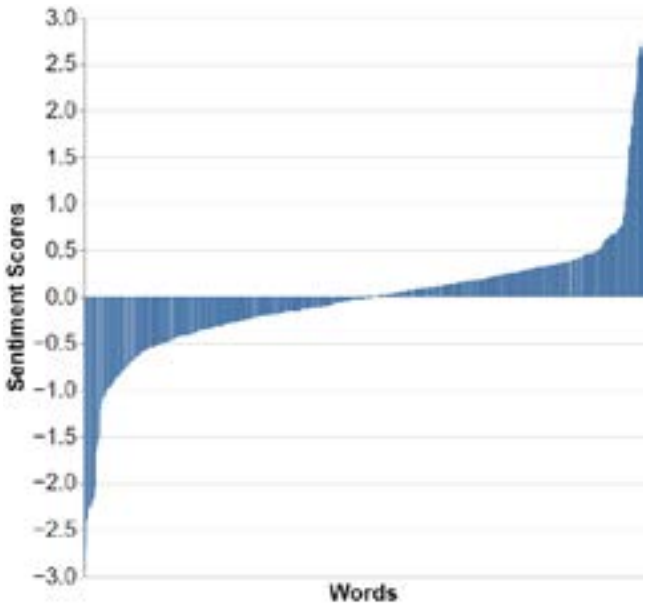
positive sentiment to see what it looked like. Obviously, an FDA document could be skewed positive or negative. I decided to go with one of the longer documents just because of regression to the mean. For a longer document, there should be positive and negative sentiment in the document. I therefore picked—not randomly—FDA's September 2022 guidance on "Clinical Decision Support Software."[3] I picked it because it was one of the longer guidance documents that I previously analyzed, so I already had much of the code written. That is probably not a great reason.

But I wanted to see what the distribution of words were on this positive versus negative scale, and I wanted to weight them by the frequency of the word used in the guidance. Figure 1 is what that frequency looks like in graph form.

I have sorted the words from most negative to most positive. I did not list the words on the x-axis because, well, there are thousands and you would not be able to read them.

Just eyeballing it, I think it looks pretty good. The words in the middle are obviously more neutral, and then you have the two extremes. The extremes look somewhat symmetrical. But I never trust my eye when I am trying to gauge, for example, how much

Figure 1. Sentiment of FDA's CDS Guidance



is positive versus how much is negative, so I thought I would run a simple calculation and take the average of all words used. That average turned out to be 0.0086400. You really could not expect much closer to zero. Thus, on the whole, in this particular guidance document, the negative sentiment words are pretty much in balance with the positive sentiment of words. I am not sure what that means but with my OCD tendencies I always like symmetry.

Truly Anecdotal Validation

Table 1

Words	Sentiment score
"medical professional"	0.2645
"attorney"	-0.2699

Table 2

Words	Sentiment score
"safe and effective"	1.8191
"adverse events"	-1.1846
"notification"	0.4109
"alarm"	-2.2675
"hospital"	-0.1968
"at home"	-0.3431

Table 3

Words	Sentiment score
"the device clinical trial was successful"	0.1353
"the device clinical trial was a failure"	-0.5736

Table 4	
Words	**Sentiment score**
"the device saved many lives"	-0.0079
"the device had few side effects"	-0.1665

Okay, so big picture it looks sort of reasonable although I have no objective evaluation of what the average of that particular guidance document should be. But why not look at a few individual words to see if their positive/negative scores seem intuitively correct. Remember that correct should be whatever we anticipate a group of FDA regulatory scientists thinking.

Table 1 compares two random words/phrases. Okay, well that seems about right.

I should point out that I really do not have any evidence to suggest that the absolute level of a sentiment score is all that accurate. As a result, what I focus on is comparing words to see if the comparisons ring true. So, Table 2 has a variety of comparisons I tried, selecting words that are somewhat common in FDA regulatory writing.

I always suggest that clients use the word "notification" instead of "alarm," and now I have the data to back that up.

It is kind of interesting that hospital is a bit negative, although certainly from a popular perspective people don't want to be at the hospital. But it is also interesting that "at home" is more negative through the eyes of FDA regulatory professionals.

One of the things that I noticed is that because of the way this algorithm is designed, evaluating sentences that include lots of words necessarily moves the score toward zero because the algorithm is just taking an average of the words, so with more words you regress to the mean. But even so, and even with many words the same, the algorithm does reasonably well with certain sentences, as shown in Table 3.

I also want to just remind you again that the algorithm is only about 80% accurate, so there are some results that caused me to scratch my head, like the ones in Table 4.

Remember too that the algorithm is just analyzing words, not sentences, so it does not catch the profoundly different meaning that words such as "not" or "few" convey in giving sentiment analysis to a sentence.

With those limitations, it seems as though the algorithm analyzing the sentiment of words created in the word embeddings is at least interesting if not somewhat accurate.

Results

What does all this tell us about whether there is bias in these word embeddings I created by training on FDA regulatory documents? Let's continue with a few comparisons in areas where I wondered if there might be bias.

Over the nearly 40 years I have been practicing FDA law, I get the sense that FDA regulatory professionals have sometimes strong opinions about the countries from which data are gathered. Let's look at Table 5 to see if there are any differences.

I need to start by observing that many of those differences are not statistically significant. We are dealing with some small numbers here frankly all clustered around neutral. But there are some stark differences, such as the difference between, say, Japan and Russia.

It is important to remember that the training data set is just the premarket review summaries as well as generally FDA guidance documents. There is really not much of that training set from

Table 5

Words	Sentiment score
"Mexico"	0.0783
"China"	0.0435
"United Kingdom"	0.1112
"Russia"	-0.5466
"France"	-0.1220
"Japan"	0.3084
"Foreign data"	-0.3008
"data"	0.1719
"foreign clinical trial"	-0.3959
"US clinical trial"	-0.1458

Table 6	
Words	**Sentiment score**
"recall"	0.0337
"warning letter"	-1.0746

Table 7	
Words	**Sentiment score**
"software"	0.4422
"hardware"	0.4782
"in vitro diagnostics"	-0.0399
"acupuncture"	0.2168
"pedical screw"	-0.2985
"ventilator"	0.2377
"infusion pump"	-0.1383
"minimally invasive"	-1.1198
"aid in diagnosis"	-0.2770
"pediatric"	-0.1878

enforcement or quality contexts, so this really wouldn't reflect FDA enforcement views.

I included the last four in Table 5 to show a more global level sentiment around foreign data versus data more generally. I honestly cannot explain why the U.S. clinical trial sentiment would be negative.

Now, look at Table 6. I thought I would assess the sentiment of some regulatory words.

I like the fact that recall is more neutral because it is simply the responsible action of a manufacturer to address the occasional but unavoidable quality issue, where warning letter has a decidedly negative connotation.

I then wanted to assess product words (Table 7) to see if there are connotations associated with different specific or even general categories of products.

Table 8	
Words	**Sentiment score**
"man"	-0.7500
"woman"	-0.2115

I will let you draw your own conclusions from those, but again, keep in mind, only 80% accurate and the magnitude of the actual scale has not been validated in any way. I really do not understand the "minimally invasive" result.

I could go on, but I will close with this. I mentioned above that I did not see much point in including race in this discussion because so few documents discuss it. I think more discussed is sex because sex has long been recognized as a factor that needs to be considered. Consider the sentiment scores for the sexes in Table 8. I will let you draw your own conclusions from that.

Conclusions

The whole point of this exercise is to illustrate that any word embeddings, because they are trained on human input, will have biases. That is true because no human being on earth is free from bias, so any machine learning model trained on that human input will have those biases.

We must be aware of those biases in all natural language processing, and more than that we must find them and then account for them. It often is impossible to remove them, but there are other coping mechanisms we have developed such as explicitly considering the existence of the bias.

In the future, I will dive deeper into this topic because I find it personally interesting, but it also is one that I think many companies need to consider on a more sophisticated basis.

Notes

* The author, a member of Epstein Becker & Green, P.C., counsels medical device, drug, and combination product companies on a wide range of Food and Drug Administration and Federal Trade Commission regulatory,

reimbursement, and clinical trial issues. He may be contacted at bthompson@ebglaw.com.

1. https://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/.
2. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon.
3. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software.