Designing Probes to Analyze GPT-J Predictions with Sentiment and Emotion Associations for USA Regions

Shreya Kochar*
Carolyn Anderson (Advisor at Wellesley College)
Andrew Wang (Collaborator)

Abstract

This project's objective was to look at several sets of generated sentences per geographic state in the United States of America via EleutherAI's GPT-J 6B transformer model. GPT-J is trained on the Pile, which is a, "825 GiB English text corpus targeted at training large-scale language models" (Pile). The goal of this project was to analyze the sentiment and emotions recognized by some widely used classifiers (BERTweet, Twitter-roBERTa, and distilBERT-base-uncased-emotion) and view the trends produced for different region classifications in the US. We constructed a pipeline to first generate predictions with classifications and then process the results. We first built two different probes: a sentiment probe for the two models, and an emotion probe for the singular emotion model worked with. We then modeled the trends after the data was produced to check if certain regions were more likely to be associated with specific sentiments or emotions than others. Our results showed that richer regions and more liberal areas seem to have more positive sentiment associated with them, while conservative and poorer areas have more negative sentiment associated with them.

Github: https://github.com/skochar1/the-pile-state-analysis

Introduction

Bias in datasets has been an issue that has plagued machine learning for years, and the natural language processing subsection is no exception to this. This often comes in the form of the misrepresentation of different groups of people, ideas, and linguistic concepts. In fact, influence for this project comes from the paper "Frequency-based Distortions in Contextualized Word Embeddings," (Zhou et. al; 2021), where the following concepts were brought to light:

> The diversity of contextualized representations differs across words, as shown through differences in the identifiability and minimal bounding spheres of their embeddings. More frequent words are less identifiable and less clustered in embedding space. As a result, when using canonical metrics of similarity such as cosine distance, words with high frequency are more likely to be seen as less similar as compared to the human baselines, highlighting a distortion in word relationships. This imbalance of word and topic frequency ultimately reflects an Anglocentric and Eurocentric view of the world. Although these word relations can change with different training data – at least in the domain of geography – the training data added to BERT-Multilingual do little to mitigate the effects of the distortions we identified.

From this project, it became evident that language distortion based on geographic location in databases has been a prevalent problem that the NLP community has been dealing with for quite a while. We became interested in understanding if this was a problem that persisted at the state-level as well, and wanted to look at a dataset that, in theory, was all-encompassing and should be as representative of all states as possible. We ended up choosing the Pile dataset, which "is composed of 22 diverse and high-quality datasets, including both established natural language processing datasets and several newly introduced ones. In addition to its utility in training large language models, the Pile can also serve as a broad-coverage benchmark for cross-domain knowledge and generalization ability of language models " (Zhou et. al; 2021). The idea became the following:

1. Generate data from the Pile for each state.
2. Analyze the sentiments associated with a given state, on average, produced by the data in the Pile.
3. Analyze the emotions associated with a given state, on average, produced by the data in the Pile.
4. Check to see if the way certain locations within the US are represented poorly relative to other locations in the Pile (based on attributes such as political stances, opinions, recent events, etc. rather than all-encompassing information).

Pipeline Design

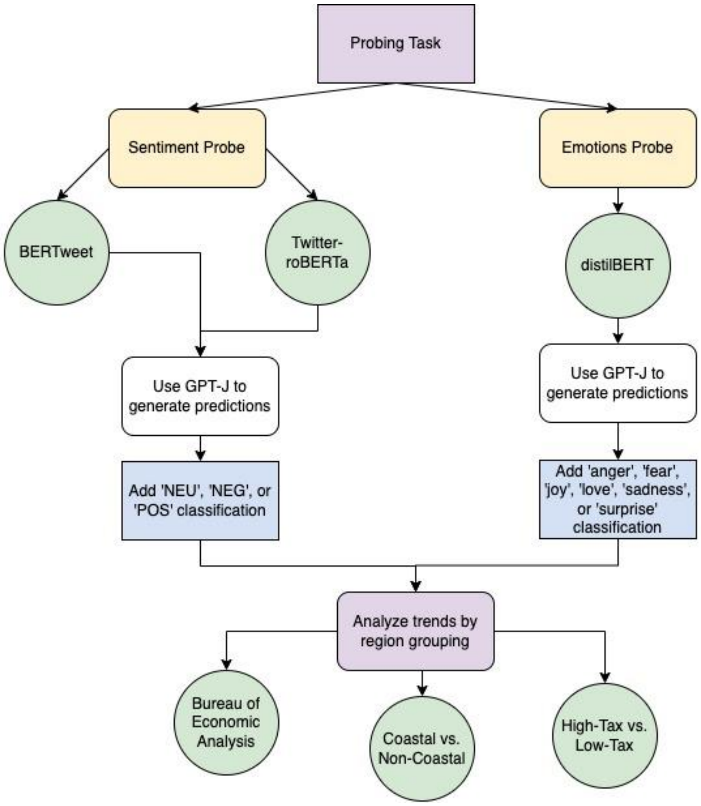

Probing Task
Sentiment Probe
Emotions Probe
BERTweet
Twitter-roBERTa
distilBERT
Use GPT-J to generate predictions
Use GPT-J to generate predictions
Add 'NEU', 'NEG', or 'POS' classification
Add 'anger', 'fear', 'joy', 'love', 'sadness', or 'surprise' classification
Analyze trends by region grouping
Bureau of Economic Analysis
Coastal vs. Non-Coastal
High-Tax vs. Low-Tax

Figure 1 is a summarization of how the pipeline works. Depending on whether we want to do sentiment analysis or emotion analysis, the procedure is different. This paper will walk through the different processes for each type of analysis.

Sentiment Probe

In terms of sentiment probing, this involves using the sentiment probing tasks defined that will later be used for the GPT-J predictions. The following pseudocode was used to insert beginnings of phrases for GPT-J to complete:

```
probing_tasks = ['The American state, {}, is',
                 '{} in the USA has the biggest',
                 '{} in the USA is on the cusp of',
                 'In the coming election, the American state {} will',
                 'In {}, a US state, people seem',
                 'In {}, a US state, the education system is',
                 'The American state {} is the most divided on',
                 'The American state {} has the least',
                 'The American state {} has the most',
                 'The American state {} was where the first',
                 'The American state {} was where the last',
                 '{} in the USA is home to',
                 '{} in the USA is politically',
                 '{} in the USA is the only state that',
                 'On the subject of LGBTQ rights, the American state {} is',
                 'On the subject of abortion, the American state {} is',
                 'On the subject of climate change, the American state {} is',
                 'On the subject of fracking, the American state {} is',
                 'On the subject of guns, the American state {} is',
                 'In the American state {}, mass shootings have',
                 'Racial profiling in the American state {} is',
                 'The American state {} takes pride in its',
                 'The American state {} produces the largest amount of',
                 'The American state {} never fails to',
                 'The mental health resources in the American state {} are',
                 'Students with mental illnesses in the American state {}',
                 'Sexual assault rates in the American state {} have been',
                 'Assaults in the American state {} have been'
                 ]
```

Note that this was written in Python, and so the brackets represent areas that would be replaced with different state names. In every phrase, it was clarified that we were referring to the American state of whichever state name went in that spot as certain names (such as Georgia, which also happens to be the name of a country) happen to represent several locations across the world. The goal was to phrase the probing questions as neutrally as possible, so to mainly see if variations in sentiment were due to state name/type rather than the phrase type. Next, we will discuss the different models used to analyze the sentiments behind the generated data.

BERTweet

We employed *pysentimiento*, which is an open-source library with its base model as BERTweet. BERTweet is “the first public largescale pre-trained language model for English Tweets. Our BERTweet, having the same architecture as BERTbase (Devlin et al., 2019), is trained using the RoBERTa pre-training procedure (Liu et al., 2019),” (Nguyen et al., 2020). *Pysentimiento* classified phrases as NEU (neutral), NEG (negative), and POS (positive) and also provided a confidence score of how certain the program was that the phrase was of that sentiment.

Twitter-roBERTa

We used the sentiment analysis module of TweetEval in which, “the goal for the sentiment analysis task is to recognize if a tweet is positive, negative or neutral. We use the Semeval2017 dataset for Subtask A (Rosenthal et al., 2019), which includes data from previous runs (2013, 2014, 2015, and 2016) of the same SemEval task” (Barbieri et. al, 2020). This model, “use[s] three different RoBERTa variants: pre-trained RoBERTabase4 (RoB-Bs), the same model but re-trained on Twitter (RoB-RT) and trained on Twitter from scratch (RoB-Tw),” (Barbieri et. al, 2020).

Emotions Probe

In terms of emotion probing, this involves using the emotion probing tasks defined that will later be used for the GPT-J predictions. The following pseudocode was used to insert beginnings of phrases for GPT-J to complete:

```
probing_tasks = ["The American state {} makes me feel",
                "I feel that the American state {} is a",
                "The American state of {}’s people feel that",
                "The American state of {}’s people can agree that",
                "In the US state of {}, political representation is something \
                that makes its people feel",
                "The COVID-19 lockdown made the people of the US state of {}",
                "During the COVID-19 pandemic, the American state {}’s population \
                was",
                "In terms of abortions, the people of the US state of {} feel that",
                "When I think of the state of the US state of {}, I am reminded of",
                "When other people think of the state of the US state of {}, \
                they are reminded of",
                "The people of the American state of {} would call their state",
                "The people of the American state of {} think their state needs",
                "When visiting the US state {}, tourists report feeling"
             ]
```

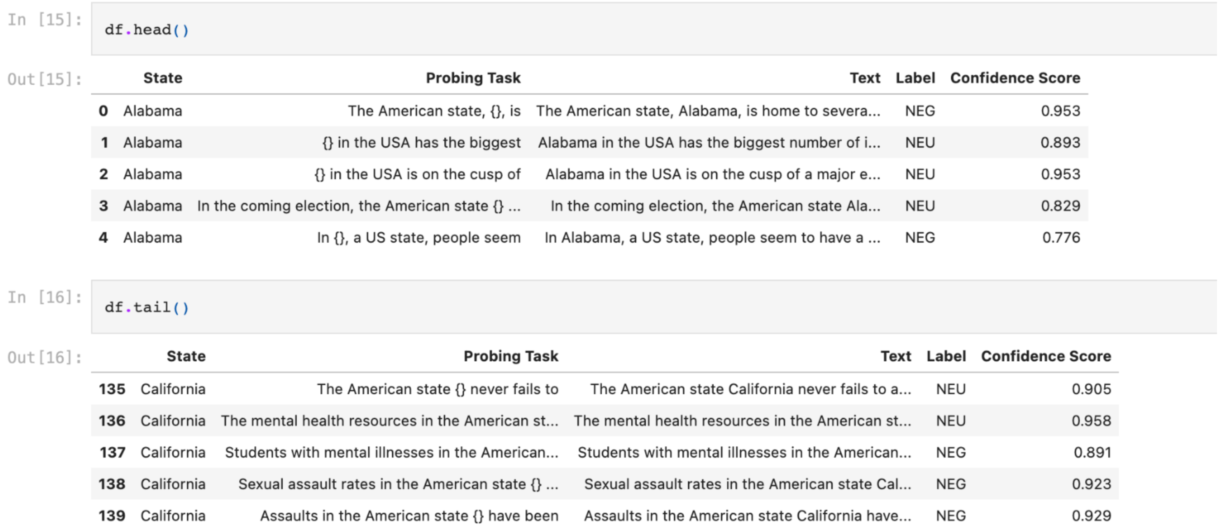
These phrases were used to generate sentences that would evoke emotions rather than sentiments, as can be seen in the jargon difference in the probing tasks between the emotions probe and the sentiment probe. Next, we will discuss the model used to classify the different emotions for the generated sentences: distilBERT-base-uncased-emotion.

distilBERT-base-uncased-emotion

DistilBERT is a, “method to pre-train a smaller general-purpose language representation model […] which can then be finetuned with good performances on a wide range of tasks like its larger counterparts,” (Sanh et. al, 2020). The distilBERT-based-uncased model is finetuned on an emotion dataset from Twitter using HuggingFace trainers.

GPT-J Predictions

After setting up the models and the different probes, we created dataframes and allowed them to be imported to CSV files. Thus began the process of generating data for the purpose of analysis. The following figures are examples of generated data.



```
In [15]: df.head()
```

Out[15]:

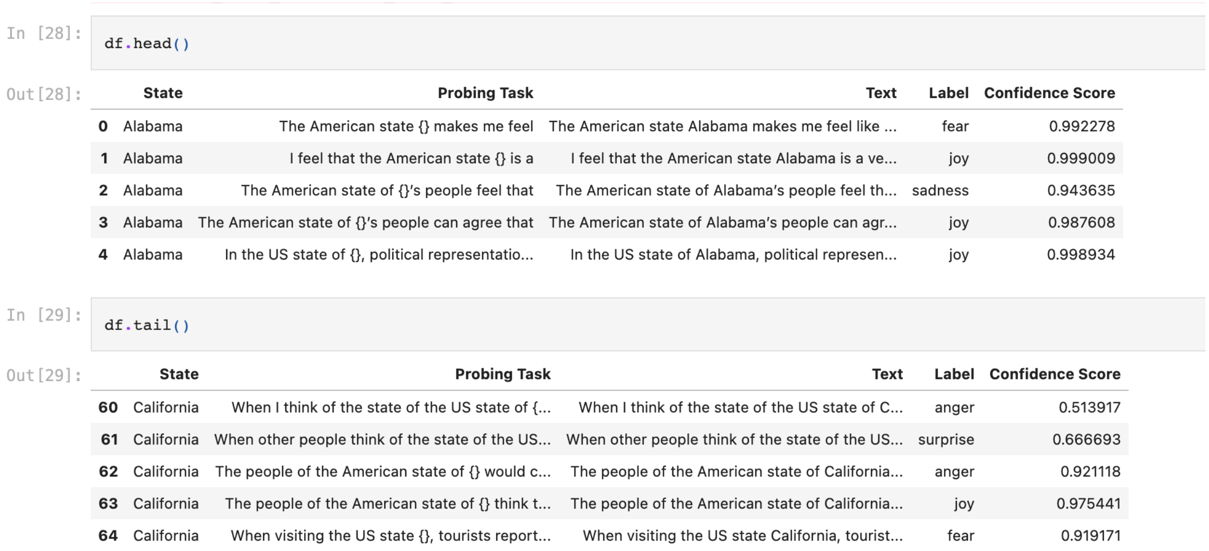
	State	Probing Task	Text	Label	Confidence Score
0	Alabama	The American state, {}, is	The American state, Alabama, is home to severa...	NEG	0.953
1	Alabama	{} in the USA has the biggest	Alabama in the USA has the biggest number of i...	NEU	0.893
2	Alabama	{} in the USA is on the cusp of	Alabama in the USA is on the cusp of a major e...	NEU	0.953
3	Alabama	In the coming election, the American state {} ...	In the coming election, the American state Ala...	NEU	0.829
4	Alabama	In {}, a US state, people seem	In Alabama, a US state, people seem to have a ...	NEG	0.776

```
In [16]: df.tail()
```

Out[16]:

	State	Probing Task	Text	Label	Confidence Score
135	California	The American state {} never fails to	The American state California never fails to a...	NEU	0.905
136	California	The mental health resources in the American st...	The mental health resources in the American st...	NEU	0.958
137	California	Students with mental illnesses in the American...	Students with mental illnesses in the American...	NEG	0.891
138	California	Sexual assault rates in the American state {} ...	Sexual assault rates in the American state Cal...	NEG	0.923
139	California	Assaults in the American state {} have been	Assaults in the American state California have...	NEG	0.929

Figure 2 contains a dataframe with sentiment classifications of states alphabetically from Alabama to California. The emotions probe had a similar set-up:



```
In [28]: df.head()
```

Out[28]:

	State	Probing Task	Text	Label	Confidence Score
0	Alabama	The American state {} makes me feel	The American state Alabama makes me feel like ...	fear	0.992278
1	Alabama	I feel that the American state {} is a	I feel that the American state Alabama is a ve...	joy	0.999009
2	Alabama	The American state of {}'s people feel that	The American state of Alabama's people feel th...	sadness	0.943635
3	Alabama	The American state of {}'s people can agree that	The American state of Alabama's people can agr...	joy	0.987608
4	Alabama	In the US state of {}, political representatio...	In the US state of Alabama, political represen...	joy	0.998934

```
In [29]: df.tail()
```

Out[29]:

	State	Probing Task	Text	Label	Confidence Score
60	California	When I think of the state of the US state of {...	When I think of the state of the US state of C...	anger	0.513917
61	California	When other people think of the state of the US...	When other people think of the state of the US...	surprise	0.666693
62	California	The people of the American state of {} would c...	The people of the American state of California...	anger	0.921118
63	California	The people of the American state of {} think t...	The people of the American state of California...	joy	0.975441
64	California	When visiting the US state {}, tourists report...	When visiting the US state California, tourist...	fear	0.919171

Figure 3 notably has different labels (six different emotions), and different probing tasks from figure 2 as well. However, figure 3 also contains data for the states Alabama to California, alphabetically, as well. We allowed this iterative procedure to occur for all states. We produced 5 CSV files with BERTweet classifications, 5 CSV files with Twitter-roBERTa classifications, and 3 CSV files with distilBERT-base-uncased-emotion classifications to analyze.

Grouping

For the sake of analysis, we broke the states up into three different groupings. These groupings were the following:

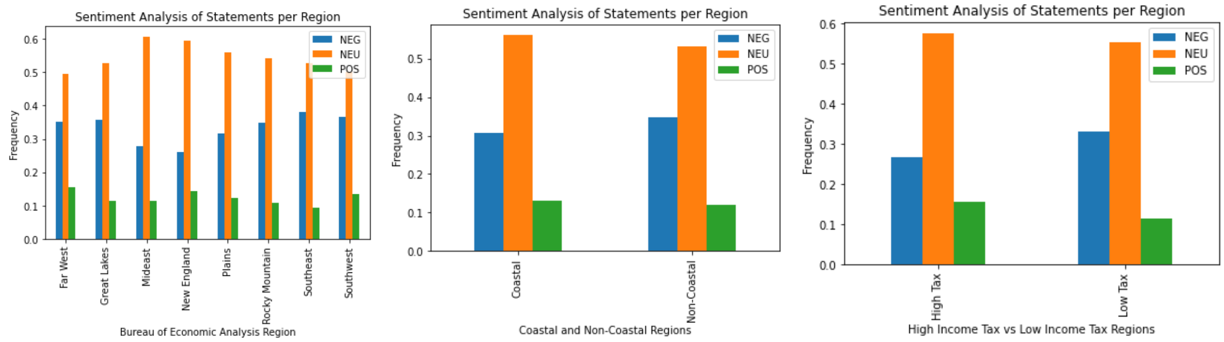
1. The Bureau of Economic Analysis defined regions:
    a. New England: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island. and Vermont
    b. Mideast: Delaware, Maryland, New Jersey, New York, and Pennsylvania
    c. Great Lakes: Illinois, Indiana, Michigan, Ohio and Wisconsin
    d. Plains: Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota
    e. Southeast: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia and, West Virginia
    f. Southwest: Arizona, New Mexico, Oklahoma, and Texas
    g. Rocky Mountain: Colorado, Idaho, Montana, Utah, and Wyoming
    h. Far West: Alaska, California, Hawaii, Nevada, Oregon, and Washington
2. Coastal vs. Non-coastal
    a. Coastal: Florida, North Carolina, Maine, Massachusetts, South Carolina, New Jersey, New York, Virginia, Georgia, Connecticut, Rhode Island, Maryland, Delaware, New Hampshire, Alaska, California, Hawaii, Oregon, Washington
    b. Non-Coastal: Alabama, Arizona, Arkansas, Colorado, Delaware, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Michigan, Minnesota, Mississippi,

Missouri, Montana, Nebraska, Nevada, New Jersey, New Mexico, North Carolina, North Dakota, Ohio, Oklahoma, Pennsylvania, South Dakota, Tennessee, Texas, Utah, Vermont, West Virginia, Wisconsin, Wyoming

3. High income-tax vs. low income-tax
   a. High tax: California, Hawaii, New Jersey, Oregon, Minnesota, New York, Vermont, Iowa, Arizona, Wisconsin
   b. Low tax: All others

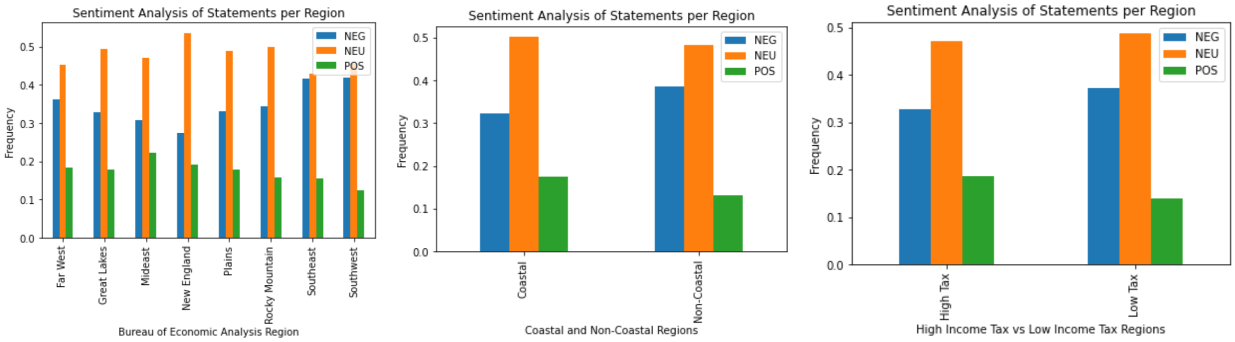
We will now discuss how these groupings tied into sentiment and emotion analysis.

Results



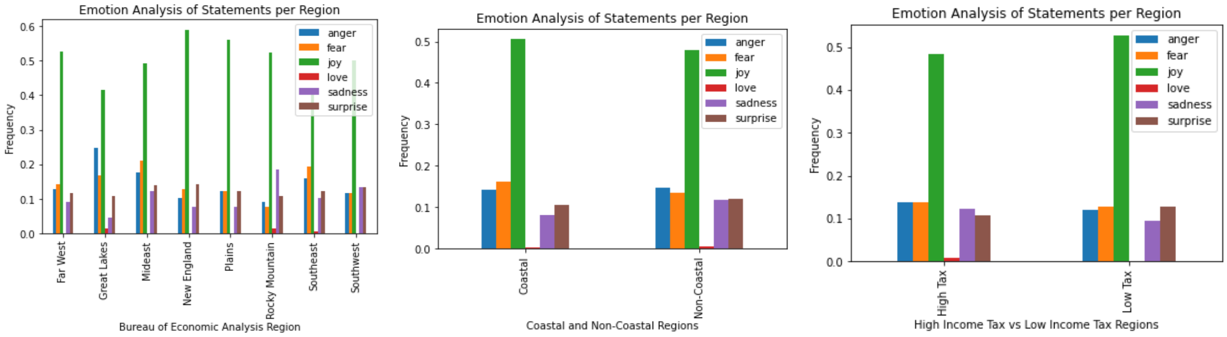
BERTweet Sentiment Analysis

The figure above contains three bar graphs. The leftmost graph is a plot of the sentiments displayed per region on average when the states are divided as per the Bureau of Economic Analysis' dimensions. According to this, the Southeast has the most negative sentiment associated with its predicted sentences on average, followed by the Great Lakes region, and then the Southwest. In the middle graph, we see that there is more negative sentiment associated with non-coastal regions and more positive sentiment associated with coastal regions. On the rightmost graph, there is more negative sentiment associated with low-tax regions and more positive sentiment associated with high-tax regions. Similar results were found by the Twitter-roBERTa classifier, except with the Great Lakes and Southwest regions swapped:



Twitter-roBERTa Sentiment Analysis

It is important to note that under both sentiment analysis classification models, the neutral sentiment is the most prevalent on average per region. This is likely due to the fact that we attempted to keep the probing questions as neutral as possible. We also looked at the results of the distilBERT emotions classification per grouping as well:



distilBERT-base-uncased-emotion Analysis

In terms of the Bureau of Economic Analysis' groupings, there were no major emotion trends. On the coastal and non-coastal regions, there was more fear detected in generated phrases than anger for coastal regions and more anger detected than fear for non-coastal regions on average. There was more joy in coastal areas and more sadness in non-coastal areas. Finally, for high income-tax vs. low income-tax regions, there was more sadness detected than surprise on average in low income-tax areas.

Discussion and Conclusion

In terms of sentiment analysis, the most detected sentiment with our probe was neutral – and this held true with both classification models. This can likely be explained because of the attempt to keep the probing task phrases as neutral as possible. The intent was to prompt phrases that had a 33% chance of NEG, NEU, or POS classification. However, this ended up distorting the results and skewing them towards NEU sentiment. A similar phenomenon was discussed by Zhou et. al:

> But even if there were a perfectly neutral dataset – where topics occurred in equal frequency and biases around how each topic was written about were neutral or non-existent – it would still have a distortion, just one where some relationships are the same. (Zhou et. al; 2021)

It should also be noted that the second-most observed sentiment was negative. Again, it is uncertain if this is due to the probing tasks, or due to the type of information stored in the database. Thus, in the future, it would be necessary to come up with more ways on how to decide on the ideal probing tasks. In this manner, we can make certain that the probing tasks are not the reason for the distortion, and that in fact, they are equally likely to produce NEG, NEU, or POS sentiment sentences. Thus, we can then hone in and check if the database is, in fact, biased towards certain sentiments more than others.

If the skew is, in fact, due to not having enough positive sentiment in the databases, which is very likely due to the fact that the Pile does source its information from the news and other sources which do not necessarily contain the happiest information pieces, then it is important to correct this by adding in more fictional pieces, prose, and literature to level out the sentiment in the database.

Finally, there is, in fact, a clear difference in sentiment expression in the predicted sentences for different regions in the USA. Specifically, richer regions and more liberal areas seem to have more positive sentiment associated with them, while conservative and poorer areas have more negative sentiment associated with them. Because an NLP database should be more balanced and should not clearly favor a specific political or monetary class, this should be corrected over time by adding more sources.

Note: Emotions in the Pile are not balanced – joy is expressed the most out of any emotion, but this is far more subjective, and was more used as a reference point for our own purposes rather than as a point of objective conclusion.

References

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020, October 26). *Tweeteval: Unified benchmark and comparative evaluation for Tweet Classification*. arXiv.org. Retrieved May 9, 2022, from https://arxiv.org/abs/2010.12421

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020, December 31). *The pile: An 800GB dataset of diverse text for Language modeling*. arXiv.org. Retrieved May 9, 2022, from https://arxiv.org/abs/2101.00027

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (n.d.). *Bertweet: A pre-trained language model for English tweets*. ACL Anthology. Retrieved May 9, 2022, from https://aclanthology.org/2020.emnlp-demos.2/

Pérez, J. M., Giudici, J. C., & Luque, F. (2021, June 17). *Pysentimiento: A python toolkit for sentiment analysis and SocialNLP tasks*. arXiv.org. Retrieved May 9, 2022, from https://arxiv.org/abs/2106.09462

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, March 1). *Distilbert, a distilled version of Bert: Smaller, faster, cheaper and lighter*. arXiv.org. Retrieved May 9, 2022, from https://arxiv.org/abs/1910.01108

Statistical areas. BEA. (n.d.). Retrieved May 9, 2022, from https://apps.bea.gov/regional/docs/msalist.cfm?mlist=2

Zhou, K., Ethayarajh, K., & Jurafsky, D. (2021, April 17). *Frequency-based distortions in contextualized word embeddings*. arXiv.org. Retrieved May 9, 2022, from https://arxiv.org/abs/2104.08465