GAYTWITTER: AN INVESTIGATION OF BIASES TOWARD QUEER USERS IN AI AND NATURAL LANGUAGE PROCESSING

Ricardo Saucedo
Faculty Mentor: Kevin Scannell, Ph.D.
St. Louis University

Abstract
Natural Language Processing (NLP) has gained attraction for its universal applications and importance in decision making in AI technology. Research has revealed that Google's NLP API holds biases toward certain words; for example, it deciphers "homosexual" as holding a negative sentiment. This investigation focuses on applying NLP strategies to the queer virtual community, colloquially known as GayTwitter, to further investigate biases. Tweets from users of GayTwitter were comprised into a dataset to build, train, and test a sentiment analyzer. This sentiment analyzer employs Word2Vec, a NLP/AI technology, in conjunction with t-SNE technology which produce word embeddings of the tweets in the corpus. From this point alone, creating a unique dataset with hopes of reducing the bias found within NLP models showed a promising trajectory for formulating a method of mitigating biased AI technology.

Keywords: AI, Natural Language Processing, Word2Vec, t-SNE, Google, GayTwitter, Twitter

Introduction

As technology continues to progress, artificial intelligence (AI) and its many compartments persevere in filling the role of many traditional techniques. In its many components, NLP is quintessential for AI to operate efficiently, accurately, and ethically to provide proper and effective service. However, current NLP models contain biases against individuals of certain racial groups, sexual identities, and genders which inherently affects them negatively in one way or another[1] (Caliskan-Islam, Bryson, & Narayanan, 2016). These biases come from the programmers who have engineered these models, despite the fact that the following appears in the ACM Code of Ethics and Professional Conduct in Principle 1.4: "Be fair and take action not to discriminate" ("Code of Ethics," 2016). Although associations or guidelines that aim to prevent unethical biases to be programmed, engineers and recent innovations have continued implementing these biased models into greater and larger products, such as AI.

Considering that AI is around us everywhere now, NLP and the Internet of Things have taken meta-data collection to a new level. Natural language processing is an area of computing that is still in research; however, its purpose is to understand and manipulate natural text to create clearer understandings in computing systems. The issue behind this investigation involves the bias toward queer users and their sexual orientation being analyzed in a negative manner, primarily due to other AI technologies, but considerably NLP, too (Wang, Kosinski, 2017; Thompson, 2017). Without considering what further research of improving mechanics of NLP models, the societal impacts of biased NLP are also a field of research that needs attention; considering that NLP models can further

[1] Google against queers, machine learning on prison bias, and gender bias articles

analyze structures of sentences, books, or "tokens," this makes these models the perfect tool for constant monitoring and investigating into inequalities, biased incidents. In my final analysis of NLP, its influence once combined with AI, and its prospective direction in further years to come, I propose the following research hypothesis: Since bias toward minority populations exists in models which inherently influence AI technology to take harmful or negative action, what biases are instilled toward the queer community, and what solutions are there to fix them?

Literature Review

With the help of my mentor's work on NLP models, Andrew Thompson's article on Motherboard guided the initial direction of our investigation, as it posed further intriguing questions to what exactly these biased NLP models looked like. Thompson's findings allude to bias that has made its way into AI, and provides examples of Google's Cloud Natural Language API doing exactly just that; utilizing the sentiment analyzer component of the API, Thompson reveals that "jew" and "homosexual" carry a negative sentiment value of about -0.2 and -0.5 respectively on a scale from -1 to 1 (Thompson, 2017). However, these results do not only affect religious groups or attacks one for their sexual identity, researchers all over have found multiple NLP models and artificial intelligence models to hold a bias in one sense or another toward gender and race as well. To exemplify the power of bias in artificial intelligence, (Angwin et. al. 2016) investigated biased prison systems which utilized face-recognition AI to denote black defendants as more likely to be at "risk" of recommitting a crime than their white counterparts (2016). While this issue deals with image processing more so than NLP, a similar study which utilized enhanced image-processing technology managed to classify one's sexual orientation from a single image at an 83% success rate (Wang, Kosinski, 2017).

In combined efforts to find a solution on how to mitigate these biases, AI researchers and legal teams have together observed that these models train and "learn" how to come become intelligent, the data in which they learn from is whatever is easiest to attain; furthermore, reinforcing that no matter how diverse creators are, or how perfect any algorithm may be, feeding in biased data will nonetheless produce biased results (Levendowski, 2018). Levendowski, a practicing lawyer, suggests copyright laws as a reason programmers and innovators are halted in bias mitigation strategies, referencing Google's decision to not publicly release the Google News corpus (2018). Without possibilities of gaining access to proper training data, it remains in question whether de-biasing NLP models is a possibility given that proper data needed for training is secured or costly to acquire, leaving algorithms behind existing AI/NLP models to operate under discrete and evidently biased measures.

Methods

Working on formulating an approach on how to strategically examine these biases in NLP models, I knew that I wanted to gather and build a corpus utilizing Twitter data: more specifically, *Gay Twitter* data. *Gay Twitter*, a colloquially known sub-community of Twitter, involves queer identifying individuals that share the space online to discuss their culture, struggles, identities, and the freedom to speak in their own vernacular. While various NLP technologies exist, and are easily accessible to the public, we chose to implement a *Word2Vec* model in this investigation for its efficiency as a word embedding processor, creating vectors from "tokens" gathered

from the corpora it is trained on (Caliskan-Islam, Bryson, & Narayanan, 2016).

For the sake of not gathering or utilizing large corpora which have been used in previous studies, a collection of tweets were gathered using a Python library called tweepy[2] granting access to Twitter API and the tweets of users a part of Gay Twitter. This way, we are testing unique, relevant, and fresh data. After compiling these into a CSV (comma separated value) file, it is crucial to ensure that all tweets pulled from Twitter API be in one encoding (i.e UTF-8). Next, this CSV file can be used in a NLP model: for this investigation, the Word2Vec model was utilized for training and testing the word embeddings. Lastly, an adaptation of Ahmed Besbes' Twitter Sentiment Analyzer (2017) was used to determine the accuracy of our model; Besbes' model implements use of the Word2Vec model, along with Keras, a high-level neural network API that runs on top of TensorFlow will be training our sentiment analyzer. After building the NLP model and training it on Besbes' provided corpus of 1,600,000 labeled tweets, and testing on tweets pulled from Gay Twitter associates, an analysis of most similar tokens and accuracy of the sentiment analyzer can be drawn.

For visualization purposes, Appendix A and Appendix B show visuals as to how the word vectors become reduced down into a two-dimensional plane by implementing a t-SNE (t-Distributed Stochastic Neighbor Embedding) technique (Van der Maatens, 2018). This prize-winning technology is applied to large, real-world datasets, hence why this strategy has proved itself to be efficient and simple in its presentation.

[2] http://www.tweepy.org/

Results

Running through Besbes' adaptation of his Sentiment Analyzer, observing trends in selected "tokens" and the accuracy of the overall model, we were met with some alarming results. To no surprise, biases toward specific tokens used as slurs against the queer community (i.e gay, fag, and lesbian) showed similarity in context to societally-negatively viewed terms:

Most Similar: Besbes' Model						
gay	'dumb' 0.664	'lame' 0.654	'retarded' 0.604	'creepy' 0.596	'rude' 0.587	'silly' 0.560
fag	'douchebag' 0.625	N/A	N/A	N/A	N/A	N/A
lesbian	'slut' 0.659	'musician' 0.625	'retard' 0.621	'gentleman' 0.619	'douchebag' 0.618	'stalker' 0.6

Figure 1. Word similarities, in context, to other tokens in Besbes' corpora.

As shown in Figure 1, the results of the NLP embedding relays informative insight into what exactly these models are synthesizing behind the scenes whenever tokens are fed through it; overall, Besbes' NLP Sentiment Analyzer scored 85% accuracy.

Onto our own model - the *Gay Twitter* NLP model - running the exact same tests as our first trial.

Most Similar: Gay Twitter Model						
gay	'straight' 0.876	'white' 0.825	'person' 0.723	'racist' 0.716	'lesbian' 0.711	'queer' 0.71
queer	'black' 0.92	'young' 0.898	'trans' 0.891	'Black' 0.862	'brown' 0.853	'women' 0.834
trans	'Black' 0.931	' queer' 0.891	'young' 0.889	'women' 0.878	'color' 0.853	'Latinx' 0.822

Figure 2. Word similarities, in context, to other tokens in our Gay Twitter corpora.

As shown in Figure 2, different tokens came back with much more promising results, inclusive results in respects to the queer community, thus providing greater insight into what the mechanics of the Word2Vec model does under the hood. Unfortunately, an inconclusive accuracy percentage of the sentiment analyzer was not determined in this trial due to complications in readability and processing of tweets.

Discussion

In reflection of my methods, there were occurrences of human error, limitations in my access to data, and challenges in understanding the models I was working with. While we made an adaptation of Besbes' Sentiment Analyzer for tweets, we came across the error in translations and encodings of the tweets in the corpora; similarly, in the second trial, ensuring all tweets pulled from Gay Twitter were all in UTF-8 encoding consumed more time than anticipated. Having access to Standard Twitter API through the Tweepy library for Python granted us limited access to Gay Twitter user's tweets, capping off at the latest 3240 tweets from the user. Thus, in comparison to Besbes' corpus of 1.6 million labeled tweets, my corpus was insignificant in size, alluding to the skewed values of similarity in tokens. However, the results from Figure 2 equated the token 'gay' being closest to 'straight,' showing the word embedding strategies making connections between two sexual orientations, different than what we observed in the first trial. Lastly, failing to come up with an accuracy score of the Sentiment Analyzer for our second trial leaves out a significant piece to this investigation; given time constraints, the time it would have taken to fix those errors would have required a few more days of work.

Despite the fall backs, this investigation did produce beneficial results and advanced our understanding of what influences a unique and corpora can have on our Word2Vec model once trained and processed. Considering the limitations, these results can still conclude a trajectory path for what is to be expected in years to come for AI and what improvements can be made. Further investigations into the realm of robotics AI and NLP seem to be on the come up, for recent innovations such as Sophia by Hanson Robotics and Erica the prospective news anchor in Japan will soon be blazing the path of where AI is heading today. Making sure to reduce or eliminate bias entirely in as many NLP models will be crucial for these new artificial intelligent species, as they will be interacting with members of society not only in person, but through data available through the world wide web.

Conclusion

In a final analysis of our investigation, it has been concluded that NLP models hold a bias in sentimental values for certain tokens, thus affecting AI to discriminate against those individuals. In this investigation, we were specifically focused on Gay Twitter, a queer community which resides online at Twitter.com. When considering that secondary data online is readily accessible and available to most users, it leads to greater inquiries as to what AI may use, or even abuse, with this type of data. As AI is composed of many other technologies other than NLP, this one field of research cannot lose its momentum, just as it is important for other sub compartments of AI to not lose their momentum either. In the context of this project, finding better methods to clean up and format tweets for readability, creating a larger corpus, and working on a machine of quicker computing capabilities can really push forward the results to show further inquiries, biases, or observable trends when working with diverse and unique datasets.

References

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Machine Bias. Retrieved from https://www.propublica.org/article/machine-bias-risk assessments-in-criminal-sentencing

Besbes, A. (2017). Sentiment analysis on Twitter using word2vec and keras. Retrieved from https://ahmedbesbes.com/sentiment-analysis-on twitter-using-word2vec-and-keras.html

Caliskan-Islam, A., Bryson, J., & Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. Retrieved from arXiv: 1608.07187.

Code of Ethics. (2016). Retrieved July 25, 2018, from https://ethics.acm.org/code-of-ethics/

Horner, E. (2007). Queer identities and bisexual identities: What's the difference? In B. A. Firestein (Ed.), Becoming visible: Counseling bisexuals across the lifespan (pp. 287–311). New York, NY: Columbia University Press.

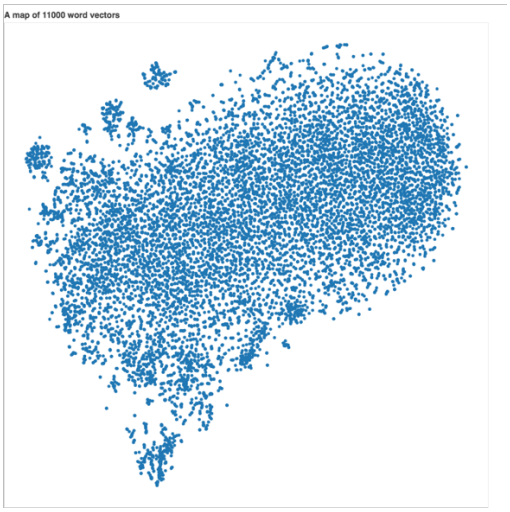
Levendowski, A. (2018). How copyright law can fix artificial intelligence's implicit bias problem. Washington Law Review, 93(2), 579-630

Thompson, A. (2017). Google's Sentiment Analyzer Thinks Being Gay Is Bad. Retrieved from https://motherboard.vice.com/en_us/article/j5jmj8/google artificialintelligence-bias

Van der Maaten, L. (2018). T-SNE. Retrieved from https://lvdmaaten.github.io/tsne/

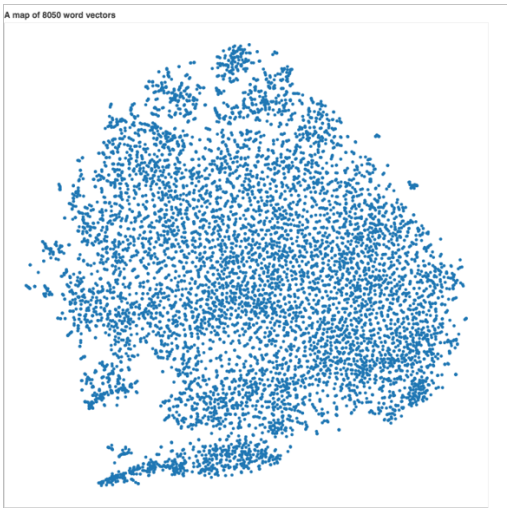
Wang, Y., & Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Retrieved from osf.io/zn79k

Appendix A
Besbes' Word Vectors Reduced to a Two-Dimension Plane



Appendix A shows what *Word2Vec* paired with *t-SNE* technology can do conjointly in visually representing words or "tokens" found in the large corpus. The points plotted on the map indicate similar surrounding words which are of the same context. In this sampling, an average of $n = 11{,}000$ words were plotted every computation.

Appendix B
GayTwitter's Word Vectors Reduced to a Two-Dimension Plane



Appendix B shows what *Word2Vec* paired with *t-SNE* technology can do conjointly in visually representing words or "tokens" found in the corpus. The points plotted on the map indicate similar surrounding words which are of the same context. In this sampling, an average of $n = 8{,}050$ words were plotted for every computation. Despite this corpus having a smaller sample size, the distribution and concentration varies in comparison to Besbes' corpus, alluding to continual investigations and work.