I. Introduction

Written text is one of the most oldest forms of media that exists. To learn and interact with the world, it is impossible to not encounter text. One of the downsides of text is that it takes time and energy to read and even more time and energy to analyze and examine text and writing. With the rise of Machine Learning, it is easier now more than ever to spend less time reading and analyzing data, lending more time to understanding the results and outcomes of data.

As someone of the Faith and a computer scientist, I am privileged to be able to not only have the skills for NLP, but also be able to apply to more a minority of media having to do with Christianity and the faith. As the subject of Christianity is a minority with in all public media there are defiantly bias that has been the result of being a minority. This largely comes from the fact the when media or content is shared within the public realm, it is to show something negative as a way to disprove or show the negative of a religion. This coverage of Christianity often focuses on scandals, controversies, or political conflicts rather than the positive contributions of Christian communities. Unless one is of the faith, tradition or simply a believer, they are not interested in the positive as they don't believe in it.

Within the article *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle,* this is what's referred to as historical bias. As NLP or any ML model is fed with data, having to do with Christianity or any other religion, since it is mostly negative the models will given only portrayals limiting them to see and learn to positives of Christianity.

This existence of bias is a great hinderance when it comes to not just Christianity or other religious texts, but also effecting both gender and race. This makes evident the need for both

mitigations and effective ways to detect bias. This will not only benefit these distinct groups but also benefit the study and practice of NLP throughout all other aspects.

II. Defining Bias

Before diving into the mitigations and detection of bias. To understand bias, one has to understand the importance of the two attribute protected and stereotyped properties. The article *Discrimination Bias Detection Through Categorical Association in Pre-Trained Language Models,* the author spells out the foundation of understanding bias. A protected property is the representation of each minority or small representation of a given group such, as race, gender or in our case religion. A stereotyped property is the expression in which the manifestation of discrimination comes from. An example of this can be someone's profession, human-describing adjectives, positive and negative verbs, just to spell out a few.[1]

These two properties can lead to bias. The bias starts to exists as the two properties, protected and stereotypes, start correlating with one another. Some examples of this can be found everywhere. Through gender classes such as hard and salon worker are generalized to often be as feminine and construction and landscaping has more of a masculine connotation to it.

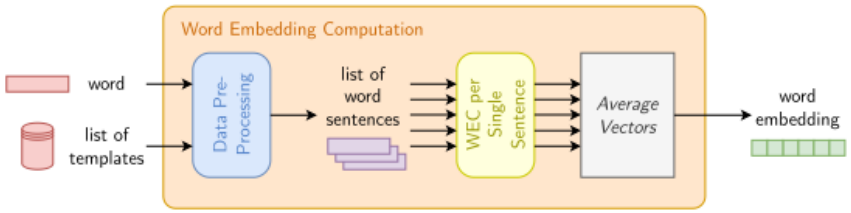
As it relates to Christianity, there a big difference in the way different churches worship. Within different traditions of the faith there is a difference in the style of worship. his bias arises when a person's religious tradition is linked to assumptions about their approach to faith, rather than recognizing the diversity of beliefs and practices within each group. Such stereotypes can lead to misunderstandings, divisions, and even discrimination within Christian communities.

III. Mitigating Bias

With the nature of how news is presented and the discussions of what to present, it is apparent to consider and evaluate what is given. Through evaluation one can consider

mitigations and solutions. The article reference above, layouts out the need for additions to be made one a data set after an evaluation was preformed.

The process for the a possible mitigation largely involved word embeddings. Starting off my taking a list words and applying them to various templates the goal was to give examples of the same words being used in not just the stereotypical context, but also presented in other possible contexts. In the article, an example of a template given was “I have a <adjective> neighbor”[2], and the possible words given were, ‘pacifist’, ‘loyal’, ’terrible’, and ‘criminal’. Each of the words would ten be applied into the template as three separate pieces of data. The diagram below, illustrates this process being preformed:



Considering the impossibilities of representing real world examples, this comes closer to accomplishing the task. Further more the study and process of the data addition were done at and random occurrence where the templates and the selected words were paired. Factoring out any chance of bias through the someone picking the words and template on their own.

further more once this has been accomplished, the templates, which are referred to as sentences, are now ready for a machine learning model to be trained. Through this there will be associated word embedding or vectors, that can then be used to discover any existing amounts of bias.

IV. Detecting Bias

With one established mean to work on mitigating bias, there is still the existence of bias within any machine learning model including a NLP model. Being able to detect bias within NLP is just as. Important as any amount of mitigation preformed.

A. The Embedding Process

The study conducted used ML models from the BERT family. BERT stands for Bidirectional Encoder Representations from Transcoders. Developed by Google, the models takes into account the text, documents, and converts each one into its own unique vector, that then can be use4d and compared to other vectors. BERT uses the tools of deep neutral networks as part of its structure and functionality.[3]. The models used were BERT, DistilBERT, RoBerta, and ELECTRA. The model BERT is the base model and the other three models are variations of the BERT model.

Once ready the documents are first split up into words and stored as tokens. These tokens are words or parts of a word that is recognized by one of the models within BERT, itself. Through the use of BERT and the models, words embeddings are formed. The resulting outcome are separated into two categories of protected and stereotyped embedding that function just like the respected properties intruded earlier. A derived set of protected embeddings given in the article are ‘christian’, ‘muslim’, ‘church’, ‘bible’, ‘mosque’, and ‘quran’. From this an example set of stereotyped embeddings include, ‘kind’, ‘lovely’, and ‘aggressive’[4].

B. Categorical Associations

focusing in on specific word embedding groups, protected and stereotyped, there is now two sets of data that both signify two different purposes and can be used to determine bias. Each group is a representation of how the model used has encoded the sentences. The protected

embedding demonstrate how the model learned and encoded the protected property. Whereas, the stereotyped embeddings demonstrate the variation in spatial differences compared to the protected embeddings, exposing bias through prejudice links between the properties.

Once the embedding are formed, they are used their distinctive forms of protected and stereotyped and classed using a machine learning classifier such as a Linear Support Vector Model, SVM. This will then give the possible distinctions between male vs. female or Christian vs. Muslim. The final matrix is an organized set of all possible data where the rows are the possible stereotyped categories(i.e. positive, neutral, negative), and the column are the predicted protected classes (e.g. Christian or Muslim)[5]. Furthermore, bias exists where ever there are high values within any of the cells. Take the following tables as an example of this illustration[6].

P_{prot}	P_{ster}	Language Models BERT	DistilBERT	RoBERTa	ELECTRA
gender (51 *female*, 51 *male*)	profession (30 *female-lean.*, 30 *male-lean.*)	33.5 %	17.8 %	39.2 %	31.9 %
gender (51 *female*, 51 *male*)	profession (11 *female-lean.*, 49 *male-lean.*)	34.8 %	36.2 %	39.9 %	40.9 %
gender (51 *female*, 51 *male*)	profession (20 *female-lean.*, 20 *balanced*, 20 *male-lean.*)	44.6 %	32.9 %	48.5 %	43.2 %
gender (51 *female*, 51 *male*)	profession (236 *high-salary*, 237 *low-salary*)	12.8 %	11.5 %	4.6 %	10.3 %
nationality (20 *british*, 20 *hispanic*)	adjectives (120 *positive*, 120 *negative*)	17.0 %	5.3 %	3.6 %	21.3 %
nationality (20 *british*, 20 *hispanic*)	verbs (43 *positive*, 41 *negative*)	9.9 %	11.6 %	4.7 %	9.0 %
nationality (20 *british*, 20 *hispanic*, 20 *russian*)	adjectives (120 *positive*, 120 *negative*)	11.0 %	6.1 %	4.2 %	7.8 %
religion (20 *christian*, 17 *muslim*)	adjectives (120 *positive*, 120 *negative*)	13.9 %	12.2 %	2.8 %	34.2 %
religion (20 *christian*, 14 *jewish*, 17 *muslim*)	adjectives (120 *positive*, 120 *negative*)	27.1 %	4.5 %	18.0 %	42.2 %
religion (12 *buddhist*, 20 *christian*, 14 *jewish*, 17 *muslim*)	adjectives (120 *positive*, 120 *negative*)	22.0 %	20.9 %	6.0 %	40.9 %

C. Evaluation Metrics

To be able to evaluate the data, a metric has tone computed to make conclusion. For evaluation there are multiple different metrics to choose form such as Chi Squared, but CramÈr V's metric is the most flexible when it comes to handling the most variation in the size of

documents, this has to do with the mathematical properties[7]. There is multiple steps in which take place to complete the metric.

The foundations of the models is the mean squared error of both the observed frequency:

$$Of(p;s) = \frac{|w_{p;s}|}{|W|}$$

as this considers the protected and stereotyped embeddings together. The other metric needed is the expected frequency, entails the predicted and original classes do not affect each other:

$$Ef(p;s) = \frac{|W_p|}{|W|} \times \frac{|W_s|}{|W|}$$

These are then both used to create a standard values between and protected embedding and stereotyped embedding:

$$MSE = \sum_{\substack{p \in P \\ s \in S}} \frac{Ef(p;s) - Of(p;s))^2}{Ef(p;s)}$$

The MSE value is further used to completed the CramÈr V's Metric on each score to normalize it in an range from 0 to 1:

$$V = \sqrt{\frac{MSE}{n \times min(|S| - 1'|P| - 1)}}$$[8]

D. Results

The study surrounding all of this examined the results, using the four different models above, to determine the likelihood of bias within NLP models on gender, race and religious specific text. The metric of all four of the models used confirmed the existence of bias especially within gender roles within the category of professions. Resulting specifically in gender bias.

Some of the further results suggest the that nationality bias, however, was found to be weaker, especially in RoBERTa, which shows low correlation values. Finally, Religion bias was also observed, with models like BERT and ELECTRA showing stronger biases toward Muslim people with negative associations.

VI. Conclusion

Natural language Processing deems itself useful for all areas of study as it can give results of accusations or ideas of text efficiently. Sometimes NLP can be the most sensitive to bias based on its reliance on data formed by the stereotypes and prejudices enveloped naturally into society. By focusing further on mitigations and detection of bias with the world of Natural Language Processing, we can work towards a more ethical and sustainable technology that will help the world.

VII. References

[1] Michele Dusi, Nicola Arici, Alfonso Emilio Gerevini, Luca Putelli, and Ivan Serina. 2024. “Discrimination Bias Detection Through Categorical Association in Pre-Trained Language Models.”, 4

[2] Dusi, 4.

[3] Simha, Anirudha. “Understanding TF_IDF for Machine Learning”. (Capital One, PDF), https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/, 2.

[4] Dusi, 5.

[5] Dusi, 7.

[6] Dusi, 9.

[7] Dusi, 7.

[8] Dusi, 7.

All equations came from the same citation above.