Multilingual model for holistic social bias detection

Anindita Ghosh
USC Viterbi School of Engineering
ghosha@usc.edu

Abstract

Neural models trained on large-scale corpora have tremendously aided research into open-domain conversation systems; yet, such corpora frequently pose different safety issues that severely impede the deployment of dialog systems in practice. Among all of these dangerous challenges, overcoming social prejudice is the most difficult since its detrimental impact on excluded communities is often communicated indirectly, necessitating normative reasoning and extensive study. This paper focuses on creating new biased conversations datasets in multiple languages for the research community to explore and use in their future research. The paper at the same time also targets a new ensemble model that detects the language of conversation and classifies it as biased / neutral or anti-biased based on the specific language models trained on the generated datasets.

1 Project Domain And Goals

Living in the modern world, where we all want equality, racial equality, bias is something that has been mitigating progress towards the goal of achieving equality. In order to live in a world free of biases, every individual needs to learn to be unbiased. People now know the importance of being unbiased and have developed a better mindset for it. But living in the modern world means living with technology. Not a day goes by where a human being does not use technology in his daily routine. But what if we as humans progress towards being unbiased but technology using machine learning or deep learning models remain biased? Would that still be progress towards a world free of bias?

This question would not have remained logical if asked in the previous years, however, the present is different from that of the past. With the continuous advancements in the field of technology, Machine learning models have become a part of every human's life. The techniques of Natural language processing are being used in many of these ML applications. Imagine, asking google to recommend a restaurant for dinner. Google would use a recommendation system using machine learning models to search through the reviews online and then suggest a subset of restaurants. But if those online reviews are biased in any way, then the decisions you make would also be biased. Similarly, a chatbot trained on some data would produce biased results if the data is biased. As a model designer, one would just preprocess the data, train the model on that data and produce the results. But no steps are taken to make sure that the model does not produce biased results.

As per the definition, NLP refers to the branch of computer science concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. So in order to induce human intelligence to a machine to make them distinguish semantic rules, its only NLP that can process and extract valuable information from huge amount of unstructured data. Use of NLP would help us understand the meanings of texts, any implicit observations or underlying sentiments associated.

The goal of our project is to detect if a dataset is biased or not. We would be working on the CDIAL-BIAS dataset for which we would be creating an ensemble model which would convert the dataset to the required language and would detect if the data is biased or not. The result would be either anti-bias, bias or neutral. The type of biases that we would be covering are race, region, gender and occupation and the type of languages covered are English, Chinese and Spanish, three of the most spoken languages in the world. When implemented successfully, applications involving textual data could use this to detect bias in any multilingual dataset.

2 Related Work

Most of the studies on bias detection are based on English and focus on few bias categories. There are few papers that have studied gender bias in Chinese language, especially on adjectives, which is a powerful tool to measure social conventions on male and female roles. Some papers focused on training the model to better learn the complete label (i.e. irrelevant, anti-bias, neutral, and biased) information in the first-stage prompt learning by employing a contrastive learning module to further regularize the text representation of the same labeled samples to the uniform semantic space. There are also papers that adopted an ensemble model approach, which combines five different pre-trained language models, and uses adversarial training and regularization strategy to enhance the robustness of the model. We will be expanding the CDIAL-BIAS dataset to a multilingual dataset that will additionally have the data in English and Spanish. This will help future researchers work on a larger multilingual dataset. We will also use an ensemble learning approach to learn on the new dataset.

3 Dataset

We will be using the CDIAL-BIAS dataset, a Chinese dataset which contains 28k conversations with detailed bias-related labels. To the best of our knowledge, this is the first well-annotated Chinese dialog social bias dataset considering both utterance level and context level. We will be creating new datasets in English and Spanish based on the benchmark CDIAL dataset and the total dataset will contain around 54k conversations.
Link to Dataset
The CDIAL-BIAS Dataset can be used as-is and does not require any pre-processing. Any pre-processing and text correction required for the new generated datasets would be taken care of during the dataset creation phase itself.

4 Technical Challenges

The idea of our project is to detect biases present in multilingual languages such as English, Chinese and Spanish. We are taking into account four different types of bias - Race, Gender, Region and Occupation. The project further aims at creating an Ensemble model that inputs a Sentence in any of the three languages, detects the language and then executes the corrresponding language model to detect and classify biases in the sentence.The project would also give us the interpretation and observations of biases present in the languages. Apart from that, we will be using a single dataset and converting the same dataset into different languages using GPT-3. One of the major problems while converting the same corpus to different languages is the nuances that each language has that could be lost. Each language has specific semantics translating which might not always be grammatically correct. Hence that could also bring more noise while trying to detect bias.

In the course syllabus, we have dealt with data in English till now. Also most of them were grammatically correct with good uses of punctuation or exclamation/interrogation marks to give us a good idea about the underlying sentiment that lies inside the text data. But the process of translating into English from other language and then processing it itself invites an amount of implicit risk because the usage of metaphors, colloquial patterns or order of words vary in different languages which cumulatively carry different meanings and not always necessarily translating them can possess the same ones. For example, while capturing details of users who supported the Russia-Ukraine war and circulated hate speeches, chat-bots used to analyze sentiments generated 2 different results when trained with 2 languages of the same channel @ukraineenglish on Telegram.

To be sure these challenges are addressed, we would repeatedly run the translated datasets with the same model and for skewed results we will manually try to induce biases to the texts generated after translating the corpus. This way the annotations would mostly be similar for all the datasets in different languages which then could be used for training and testing purposes. Also we would thoroughly apply the data cleaning and preprocessing steps like tokenizing and lemmatizing the data to yield the best results. This data-preparatory pipeline could also be enhanced later based on research and brainstorming ideas.

We present the aggregated precision, recall, and F1 score over all the classes of classifiers trained. The model is trained on utterance level and context-level to generate a weighted F1 score for the overall classifier.

References

[Sheng2019] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan and Nanyun Peng. 2019. *The Woman Worked as a Babysitter: On Biases in Language Generation*. University of Southern California, University of California, Los Angeles.

[Zhu and Liu2020] Shucheng Zhu and Pengyuan Liu. 2020. *Great Males and Stubborn Females: A Diachronic Study of Corpus-Based Gendered Skewness in Chinese Adjectives*. School of Information Science, Beijing, China.

[Zhao2022] Lucy Li and David Bamman. 2021. *Gender and Representation Bias in GPT-3 Generated Stories*. University of California, Berkeley.

[Yang2022] Aimin Yang, Qifeng Bai, Jigang Wang, Nankai Lin, Xiaotian Lin, Guanqiu Qin and Junheng He. 2022. *A Fine-Grained Social Bias Measurement Framework for Open-Domain Dialogue Systems*. Natural Language Processing and Chinese Computing, China.

[Zhao2022] Jishun Zhao, Shucheng Zhu, Ying Liu and Pengyuan Liu. 2022. *CDAIL-BIAS MEASURER: A Model Ensemble Approach for Dialogue Social Bias Measurement*. Natural Language Processing and Chinese Computing, China.

[Zhou2022] Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu and Helen Meng. 2022. *A Towards Identifying Social Bias in Dialog Systems: Frame, Datasets, and Benchmarks*. The Chinese University of Hong Kong, State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing China.