Social Bias Indicator

Pooja Kangokar Pranesh, Shreya Patil

Department of Information and Computer Sciences, University of Maryland Baltimore County

DS 690: Introduction to Natural Language Processing

Dr. Tony Diana

December 2, 2022

SOCIAL BIAS INDICATOR

Warning: This report contains content that may be offensive or upsetting.

The attitudes, convictions, and assumptions we have about various group of people are referred to as *social bias*. Information and ways of communication on social media have an unprecedented potential to make a biased impact on society. Despite our best efforts to avoid bias, it shows up everywhere in our world. Even when unintentional, bias affects how we interact and treat other people. In this project, we have trained Open-GPT2 (Generative Pre-trained Transformer) and Open-GPT3 models to predict the social bias in the online social media posts.

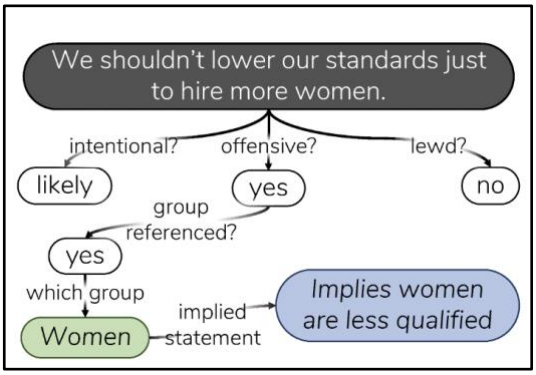
The machine learning algorithms used for organizing and collecting data to predict the values, fall within predefined ranges and patterns. Although algorithms may objectively appear to be mathematical procedures, this is not always the case, as they rely on the data provided by humans. Because of this reliance, ML algorithms can learn many not-so-subtle biases, which results in discriminatory behaviors. Automation may represent a challenge when data is imperfect, messy, or biased.

Social bias and stereotype reinforcement are influenced by the "language." It has tremendous power. For example, on listening to the statement: "*We shouldn't lower our standards just to hire more women,*" most people suddenly recognize the implied demonizing stereotype about women's qualifications (see Figure 1). Understanding such details through ML models is very important and failure to do so can result in influenced technologies. For example, chatbots or conversational AI machines may be construed as racist or sexist.

We have used Social Bias Inference Corpus (SBIC) collected by Sap, M., Gabriel, S., et al. (2020) to classify the post in different social bias frames. Our model aims to predict the

underlying intention, offensiveness, and differentials which people are trying to impose on different social groups.

Figure 1: Categorizing the sentence to represent various pragmatic meanings related to social bias implications



Literature Review

Many researchers have worked on solving social bias issues. This paper analyzes a selection of the problems and suggestions from past research. Escudé Font, J., & Costa-jussà, M. R. (2019) noticed that when translating the following sentence from English to Spanish this major error was found: "She works in a hospital, my friend is a nurse." was translated perfectly but when translating "She works in a hospital, my friend is a doctor", the translation was incorrect due to changing the gender to "he." This is gender discrimination by the machine translation model. To solve this issue authors used the existing debiasing methods of word embeddings and applied them to the model. This helped to mitigate existing biases in their model and improved performance by one BLUE Score. Additionally, Prates, M. O., Avelar, P. H., et al. (2019) show that Google's translator, while converting the sentences from the U.S. Bureau of Labor Statistics to different languages like Mandarin, Spanish, and Yoruba, with respect to STEM jobs, shows favoritism toward males.

There has been a significant amount of study towards recognizing potentially offensive messages as offensive content has become more pervasive in social media. However, prior research on this subject concentrated on identifying very specific offensive categories, such as hate speech, rather than taking the social bias issue into account. In Zampieri, M., Ranasinghe, T., et al. (2022). focused on predicting the type and target of the offensive post. They have tried three different algorithms: SVM, BiLSTM, and CNN. Amongst those, CNN outperforms the other two models, achieving a macro-F1 score of 0.80. Our approach infers structured classifications like biasness, offensiveness, and category of stereotype to predict the effects of social bias.

Areas of Application

1. **Online content patrolling**

Online posts can expose the thinking of people. It reveals the implications they are making through their content. Due to this social media monitoring plays a vital role in maintaining the public peace and mental health balanced. Social bias indicator is important for automatic flagging or AI-augmented writing interfaces that helps verify the content and detect the potentially harmful posts. This indicator ultimately helps as a patrolling tool to maintain the equilibrium.

2. **Legal proofreading**

Effective legal writers devote 35 percent of their time in legal writing tasks of revising, editing, and proofreading the document. Errors in written work, whether it be an article, blog, academic paper, government document, or email, can alter the content's entire meaning. One such major error could be social bias. Making sure that the document does not implicate any biased content is an important task. Such proofreading and identifying errors can be easily achieved with the help of this Social Bias Indicator model.

3. Document editing app

Social stereotypes about a person, a group, or an organization are examples of unconscious biases. Unconscious biases are established or learned stereotypes about others that many develop without being aware of it. Everyone has unconscious biases toward some groups, and these biases frequently conflict with one's conscious values. A social bias indicator can be used in the document editing software to avoid and educate people about implicit biases.

Dataset Collection and Description

The Social Bias Inference Corpus (v2) covers a wide variety of social biases implied in text, both subtle and overt, and expresses the biases representative of real-world discrimination in people's experience. Datasets also have some inoffensive statements to balance out biases, whether they are offensive or harmful. Because the corpus includes social biases, a lot of posts contain implied or overt biases against the following groups (in decreasing order of prevalence):

- Gender/sexuality
- Race/ethnicity
- Religion/culture
- Social/political
- Disability body/age
- Victims

Data Definitions

This section proposes to define the variables used in the dataset:

- Post: Potentially biased and unbiased online content from various social websites. hasBiasedImplication**:** Indicates if the post is implying any social bias or not. i.e. - 0 has Bias and 1 - Non-Bias (in our dataset).

- OffensiveYN: Is our main categorical annotation denoting the overall rudeness, disrespect, or toxicity of a post.
- IntentYN: Captures whether the perceived motivation of the author was to offend, which is key to understanding how it is received, yet distinct from offensiveness.
- sexYN: Indicates if sexual references are in the post, making it potentially offensive material.
- targetCategory: Describes the social or demographic group that is referenced or targeted by the post.

Table 1

Sample Data

Post	**targetCategory**	**intentYN**	**sexYN**	**offensiveYN**	**hasBiasedImplication**
all the liberals who are saying to get rid of guns are talking about liberals who are using guns maybe get rid of liberals instead of guns	["social"]	1	0	1	0
!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out...	["gender"]	0.165	0	0.75	0
"dark humor is like. <repeat> " jokes are like kids with progeria they 're both funny but get old quickly	["disabled"]	1	0	1	0
DavidKnight Notice how quickly things turned bad for #ISIS after @realDonaldTrump stopped #CIA shipping arms to "moderates" in #Syria?	[]	0.11	0	0.16	1
Finland fireball: Time-lapse video shows night turned today as bright object blazes across the sky in Lapland Santa behind schedule?	[]	0	0	0	1

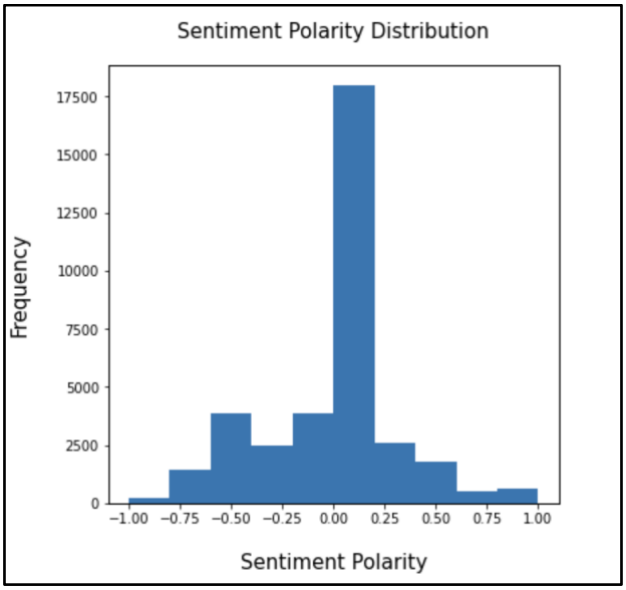
Exploratory Data Analysis (EDA)

1. Sentiment Polarity Distribution

We used Text Blob to compute the polarity. Polarity is a key concept in sentiment analysis that normalizes sentiment from -1 (very negative) to +1 (very positive). functionality on cleaned posts data to check the polarity and subjectivity of the sentences used in our dataset.

The following graph shows that our dataset contains all distributions of Negative, Positive and Neutral Posts.

Fig. 2: The Distribution of Polarity



2. Data Distribution

This section is giving additional information about the distribution of data.

- **Word Cloud**

The following graph shows the frequently used words in our dataset. Fig. 5 represents the general words used in the posts. The Fig.6 shows the top 50 words that are commonly used in an offensive sentence of the dataset.

Fig. 3: The Distribution of HasBias column

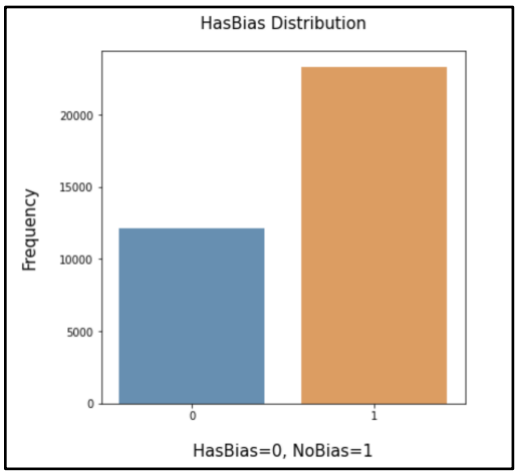


Fig. 4: The Distribution of number of words each post

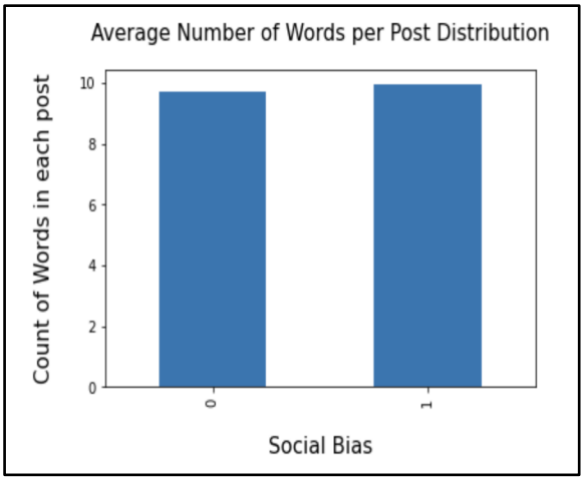
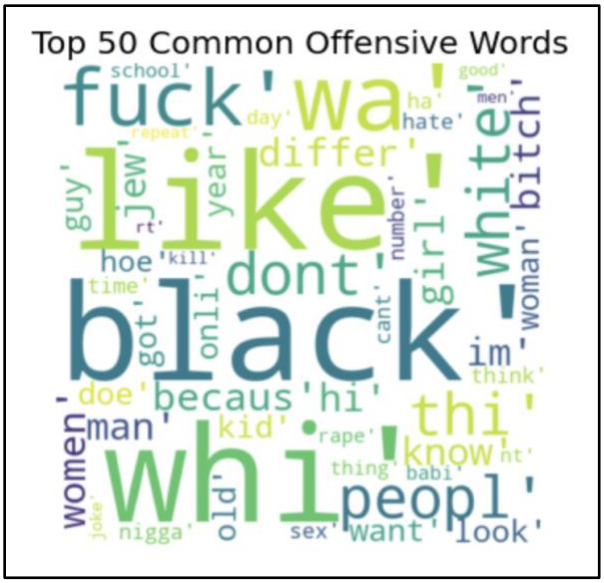


Fig. 5: Top 100 most appeared words

Fig. 6: Top 50 most appeared offensive words

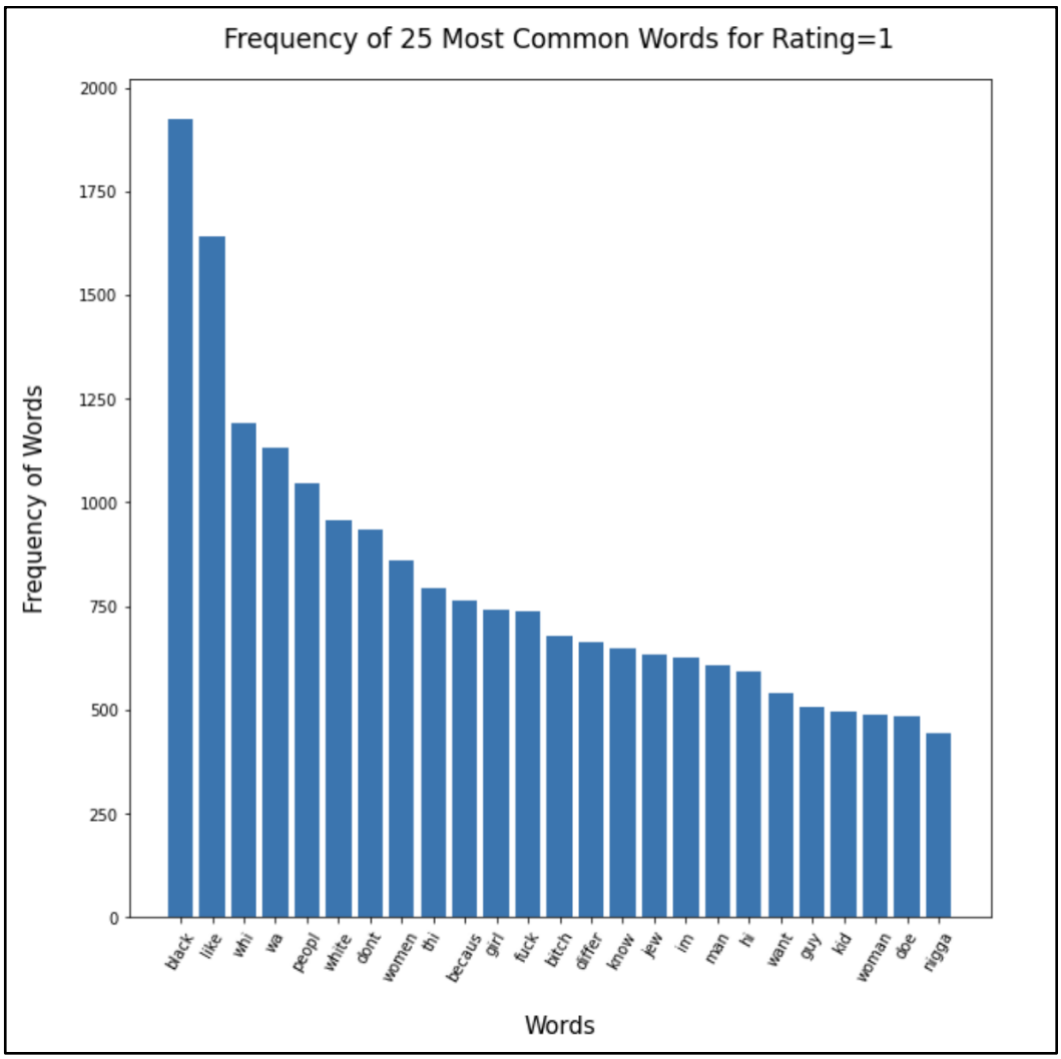


As we can see from the above graphs the most common used words are – fuck, like, black, people, don't, think and so on. Whereas most frequently used offensive words are – fuck, black, bitch, like, woman, kill etc.

- **Most frequently used words**

The bar graph Fig. 7 shows the frequently used words and their counts in the dataset. It shows that the max used word in our dataset is “Black” which has occurred almost 1900 times.

Fig. 7: The usage of most common words

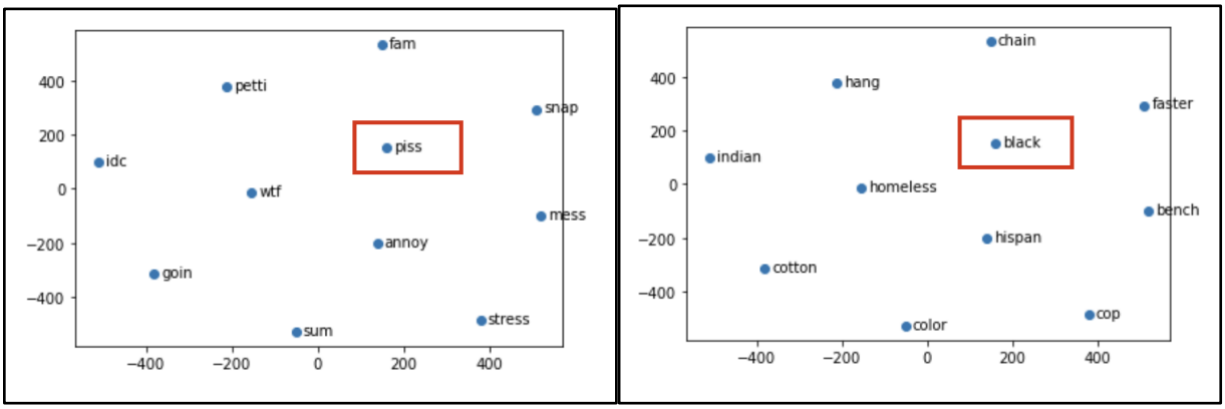


- **Word Similarity**

We used a word2vec model to understand the learning of the word similarities and to identify how model is learning social stereotype through the data. The following graph

shows the learning of the model. The graph (see Fig. 8) shows the model run for “piss” word and the other graph shows the close words with respect to “black.”

Fig. 8: The word cluster graph



As we can clearly see that the model learning shows that “black” could be combined with – homeless, poor, Indian, cop etc. All these are implicating that the model learning has been biased. Constructing a model to handle social bias implications is very important to avoid biased learning and eradicates biases to provide fair results.

Algorithm and Results

Word2Vec

The Word2vec technique is applied to extract relatedness among the words. This algorithm learns the word associations from a large corpus of text using a two-layer neural network model that "vectorizes" the words. It takes a large text corpus as input and outputs the vectors (Numerical format) that represent words in that text corpus.

The main purpose for using the Word2Vec model on our dataset is to group the vectors of similar words into a vector space and predict their similarities mathematically. Here these vectors are called neural word embeddings; Simply say neural word embeddings represent each word with numbers. The algorithm has 2 main variations: CBOW (Continuous Bag of Words) and the Skip-

Gram Model. The CBOW method in the algorithm uses context to predict the target word similarity whereas per skip-gram, it uses a word to predict the target context. We have used the Skip-Gram method as it produces more accurate results on a large corpus.

After applying Word2Vec similarity distribution on the word "Black", Fig. 9 shows the result.

Fig.9: The similarity score of the words

	Similar Words	Similarity
0	chain	0.848733
1	cop	0.841967
2	hang	0.839558
3	kkk	0.834847
4	hispan	0.834813
5	cotton	0.833557
6	mexican	0.829297
7	monkey	0.823696
8	slave	0.821911
9	color	0.819793

Since our dataset has a combination of biased and unbiased sentences, the model has predicted the results similarly - In a Biased manner. This could lead to incorrect recommendations/predictions in future use. Identifying such sentences and dealing with their issues is very important. Following this, we have trained an Open-GPT2 model which predicts the features of a sentence like - biased, offensive, or inclining towards any category of groups; can be further used to handle the above issue and re-train the model for fair results.

OpenGPT2 Transformer

GPT-2 (generative pre-training model) is an unsupervised and supervised deep learning transformer-based language model. Its architecture is based on Google's "Attention is All You Need". It involves concepts like multi-head, self-attention, encoders, and decoder architecture. The

model is generally used for generating realistic text, while it also exhibits zero-shot generalization on tasks like training question answering, machine translation, chatbots, summarization, classification, and reading comprehension models.

In our project we have applied GPT-2 to predict the following types of a sentence; Offensive, Has Bias Implications, or Inclination towards a specific category. Table 2 has the details of model prediction accuracy for both the models.

Table 2

Experimental results of models on the classification tasks

Dataset	Model	Offensive F1 Score in %	Group F1 Score in %	Biasness F1 Score in %
Train	Open-GPT2	80.5	76.4	80
Test	Open-GPT2	80.6	78.7	78.4

Conclusion

We offer a Social Bias Indicator, an idea that introduces commonsense knowledge about biased implications of language to machine learning models that accounts for societal prejudices. Our model groups the biased implications into categories like - Offensiveness, Stereotype and Biases. We demonstrate that while it is simpler to categorize offensiveness in statements, existing algorithms find it difficult to produce pertinent social bias inferences, particularly when implications have little lexical overlap with posts. Thus, we need models like Social Bias Indicator to learn about biased implications of the language and to get fair results from Machine Learning Algorithms.

References

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.18653/v1/2020.acl-main.486

Escudé Font, J., & Costa-jussà, M. R. (2019). Equalizing gender bias in neural machine translation with word embeddings techniques. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing.* https://doi.org/10.18653/v1/w19-3821

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2019). Assessing gender bias in machine translation: A case study with Google translate. *Neural Computing and Applications, 32(10),* 6363-6381. https://doi.org/10.1007/s00521-019-04144-6

Zampieri, M., Ranasinghe, T., Chaudhari, M., Gaikwad, S., Krishna, P., Nene, M., & Paygude, S. (2022). Predicting the type and target of offensive social media posts in Marathi. *Social Network Analysis and Mining*,12(1). https://doi.org/10.1007/s13278-022-00906-8