Social Bias Frames: Reasoning about Social and Power Implications of Language

Weld Lucas Cunha
Prof.: Ricardo Sovat

What are Social Bias Frames?

Social Bias Frames is a new way of representing the biases and offensiveness that are implied in language. For example, these frames are meant to distill the implication that "women (candidates) are less qualified" behind the statement "we shouldn't lower our standards to hire more women."

The main aim for this dataset is to cover a wide variety of social biases that are implied in text, both subtle and overt, and make the biases representative of real world discrimination that people experience.

Available at: https://homes.cs.washington.edu/~msap/social-bias-frames/

The Dataset

The Social Bias Inference Corpus (SBIC) supports large-scale learning and evaluation of social implications with over 150k structured annotations of social media posts, spanning over 34k implications about a thousand demographic groups.

The dataset is split into train, validation, and test sets (75%/12.5%/12.5%).

train	validation	test
112900	16738	17501

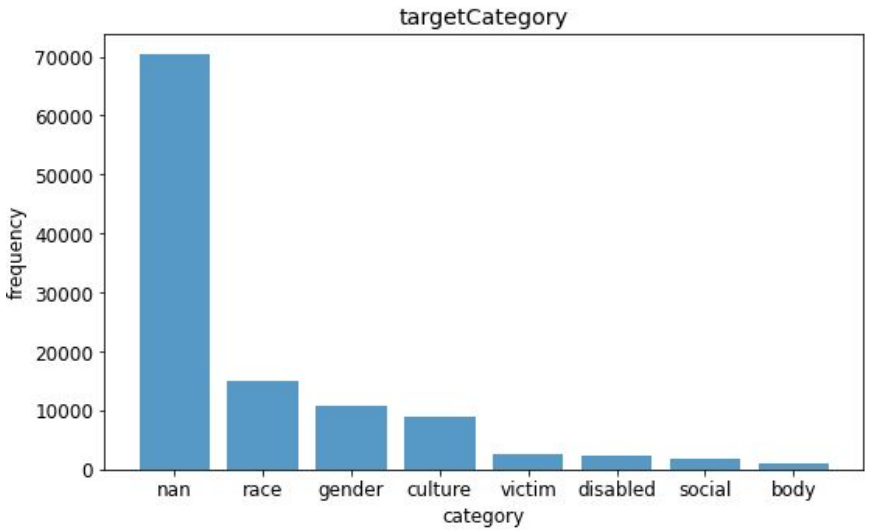
Supported Tasks and Leaderboards

- This dataset supports both classification and generation. Sap et al. developed several models using the SBIC.
- They report an F1 score of 78.8 in predicting whether the posts in the test set were offensive;
- F1 score of 78.6 in predicting whether the posts were intending to be offensive;
- F1 score of 80.7 in predicting whether the posts were lewd;
- F1 score of 69.9 in predicting whether the posts were targeting a specific group.

Data Fields

- ***whoTarget***: a string, '0.0' if the target is a group, '1.0' if the target is an individual, and blank if the post is not offensive
- ***intentYN***: a string indicating if the intent behind the statement was to offend. This is a categorical variable with four possible answers, '1.0' if yes, '0.66' if probably, '0.33' if probably not, and '0.0' if no.
- ***sexYN***: a string indicating whether the post contains a sexual or lewd reference. This is a categorical variable with three possible answers, '1.0' if yes, '0.5' if maybe, '0.0' if no.
- *sexReason*: a string containing a free text explanation of what is sexual if indicated so, blank otherwise
- *offensiveYN*: a string indicating if the post could be offensive to anyone. This is a categorical variable with three possible answers, '1.0' if yes, '0.5' if maybe, '0.0' if no.
- *annotatorGender*: a string indicating the gender of the MTurk worker
- *annotatorMinority*: a string indicating whether the MTurk worker identifies as a minority
- *sexPhrase*: a string indicating which part of the post references something sexual, blank otherwise
- *speakerMinorityYN*: a string indicating whether the speaker was part of the same minority group that's being targeted. This is a categorical variable with three possible answers, '1.0' if yes, '0.5' if maybe, '0.0' if no.
- *WorkerId*: a string hashed version of the MTurk workerId
- *HITId*: a string id that uniquely identifies each post
- *annotatorPolitics*: a string indicating the political leaning of the MTurk worker
- *annotatorRace*: a string indicating the race of the MTurk worker
- *annotatorAge*: a string indicating the age of the MTurk worker
- ***post***: a string containing the text of the post that was annotated
- ***targetMinority***: a string indicating the demographic group targeted
- ***targetCategory***: a string indicating the high-level category of the demographic group(s) targeted
- ***targetStereotype***: a string containing the implied statement
- *dataSource*: a string indicating the source of the post (t/...: means Twitter, r/...: means a subreddit)

Data Exploration



Samples:

- culture: There's a lot of wildfires around us and I noticed some ashes falling down at my Jewish friend's house I called to make sure his family reunion was going well (r/darkjokes)
- gender: what do you call bathrooms for transgender people ? the disabled bathroom (r/meanjokes)
- victim: What's the difference between hitting and touching. I've never hit a child. (r/darkjokes)

Objective

Train a Machine Learning Model that classifies the posts into one of the seven categories, based mainly on the post text:

- race
- gender
- culture
- victim
- disabled
- social
- body

Model 1: Sklearn Baseline (Logistic Regression)

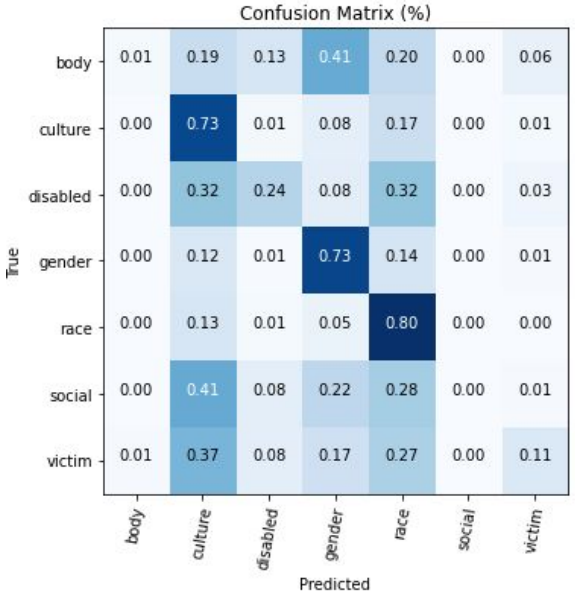
Descrição do modelo:

- Proprocessamento simples: TfidfVectorizer com ngram_range de 1 a 3 e 100 features.

```
vectorizer = TfidfVectorizer(strip_accents='unicode',
                             lowercase=True,
                             stop_words=stop_words,
                             max_features=max_features,
                             ngram_range=(1, 3))
```

- Modelos de ML (sklearn) com Cross-validation

```
Correctly classified......: 6264/9200
Accuracy (simple) ........: 0.68
Balanced acc.  ...........: 0.38
```



Model 2: Enhanced Model Sklearn (Random Forest)

Descrição do modelo

- Proprocessamento melhorado: Melhorias NLP, TfidfVectorizer com ngram_range de 1 a 3 e 500 features tfidf;

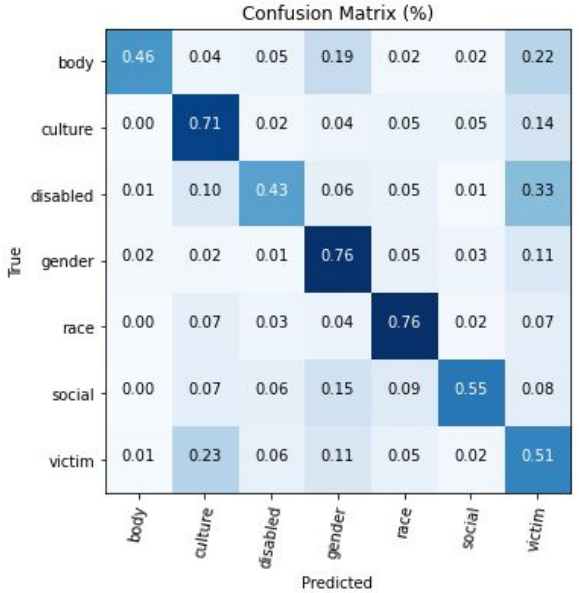
- Features adicionais além do texto do post: comprimento do texto e origem;

- Seleção de features com baixa variância;

- Reequilibrio das classes com SMOTE;

- Modelos de ML (sklearn) com Cross-validation.

```
Correctly classified......: 6571/9200
Accuracy (simple) ........: 0.71
Balanced acc.  ...........: 0.60
```



Model 3: Enhanced Model TensorFlow (Neural Network)

Descrição do modelo

- Proprocessamento melhorado: Melhorias NLP, TfidfVectorizer com ngram_range de 1 a 3 e 500 features tfidf;

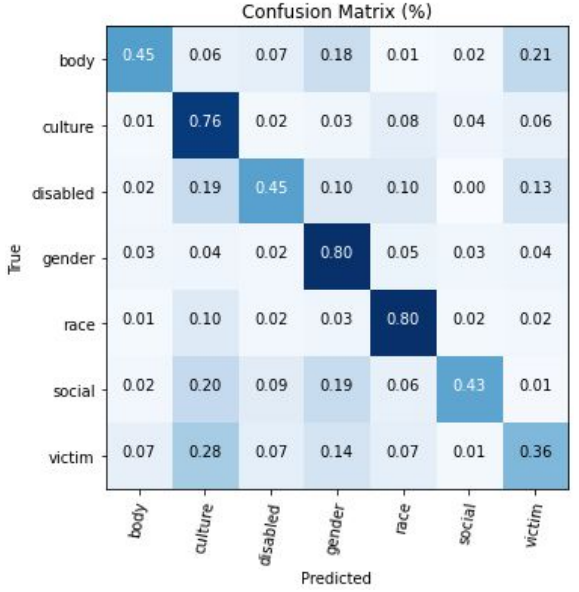
- Features adicionais além do texto do post: comprimento do texto e origem;

- Seleção de features com baixa variância;

- Reequilibrio das classes com SMOTE;

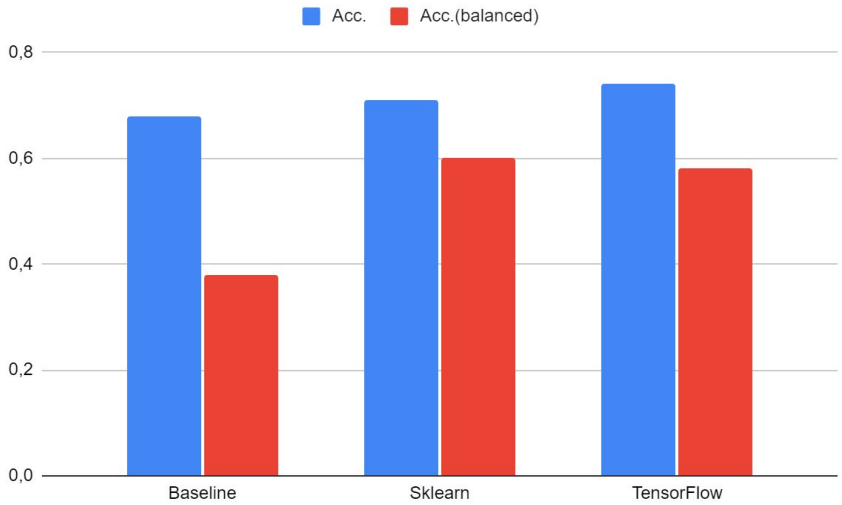
- Modelo de NN com Hold-out.

```
Correctly classified......: 6822/9200
Accuracy (simple) ........: 0.74
Balanced acc.  ...........: 0.58
```



Resultados e Conclusões

- Melhorias no pré-processamento, incluindo as features NLP e rebalanceamento das classes foi muito importante para melhorar a Acurácia balanceada.
- A rede neural não apresentou melhoria relevante se comparada ao modelo de machine learning convencional.
- A maior oportunidade de melhoria de desempenho está nos dados e não no modelo.



Muito Obrigado!