Social Bias Frames
NLP Course Project & Project Work

Simone Mele, Matteo Periani, Gian Marco Baroncini and **Giuseppe Mantineo**
Master's Degree in Artificial Intelligence, University of Bologna
{simone.mele, matteo.periani2, gianmarco.baroncini, giuseppe.mantineo }@studio.unibo.it

Abstract

Language wields significant influence, capable of offending sensibilities, perpetuating stereotypes, and conveying social biases. The core of this study delves into the challenge of pinpointing socially inappropriate phrases and comprehending their often implicit meanings. In an an era where communication assumes growing importance, verbal precision attains increasing significance. Eradicating even the most subtle forms of discrimination becomes imperative. As cited in the original study (Sap et al., 2020), the statement "we shouldn't lower our standards to hire more women" does not explicitly employ offensive language. Nevertheless, it subtly implies that women are less qualified than men. Inspired by these motivations, we strived to replicate and even exceed the performance attained in the previously cited work.

1 Introduction

This works examines the crucial matter of detecting social biases and offensive stereotypes within the digital domain. The creation of technological solutions adept at identifying and understanding these harmful occurrences becomes an indispensable requirement.

Previous approaches to identifying harmful statements relying primarily in a straightforward toxicity classification (e.g., (Founta et al., 2018); (Davidson et al., 2019)). In light of this, we follow the approach proposed by (Sap et al., 2020) to go beyond the conventional classification of statement, trying to identify whether the post is targeted towards a specific group and the possible implied statement.

The resulting task is divided into two subproblems: a categorical classification of the post and a free-form text generation.

We use two different type of architectures: a Decoder-Only, following the work of (Sap et al., 2020) and an Encoder-Decoder, which tries to solve limitations of the first one. Also, we improve quality of collected data, by semantically reducing annotated target groups.

We find that these models are effective at classifying whether a post is offensive or is targeting a groups, while struggle on generating explanation about the offense: they exhibit a lack of deep understanding and reasoning.

2 System description

Following the work of (Sap et al., 2020), we cast our work as a hybrid classification and language generation task. We explore two different architectures based on a Decoder-only and an Encoder-Decoder.

Given a post, it is pre-processed and tokenized to be provided as input to our models, which produce as output five binary values (offensive, intentional, sex, vs_group, in_group), and two free-form texts (group and stereotype).

2.1 Decoder-Only Model

Our approach is highly based on the one presented in (Sap et al., 2020). We use GPT2 (Radford et al., 2018) and we add two task-specific vocabulary items for each of our five classification features, each representing the negative and positive values of the class (e.g., for offensiveness, `[offY]` and `[offN]`). Then, we transform classification features into binary values by assigning 1 if has a values of "yes", "probably yes" or "maybe", and 0 otherwise.

Given as input a tokenized post followed by a separator token

$$w_{P_1}, w_{P_2}, \ldots, w_{P_N}, \texttt{[SEP]}$$

the model generates the annotation as follows[1]:

$$w_{off}, w_{int}, w_{sex}, w_{grp}, \texttt{[SEP]},$$
$$w_{G_1}, w_{G_2}, \ldots, \texttt{[SEP]},$$
$$w_{S_1}, w_{S_2}, \ldots, \texttt{[SEP]}, w_{ing}$$

where w_{G_i} are the tokens representing the group and w_{S_i} the stereotype.

At the last layer, the model projects the embedding into a vocabulary-sized vector, which is turned into a probability distribution over the vocabulary using a SoftMax layer.

The model is trained by concatenating the input with the expected network output. We minimize the weighted cross-entropy of the output tokens averaged across number of non-ignored tokens (T_i) and batch size (N):

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t_i=1}^{T_i} w_{y_t} CE(x_{i,t}, y_{i,t}) \quad (1)$$

where w_{y_t} is the weight of the vocabulary token y_t. During training, no loss is incurred for variables that cannot take values due to earlier variable values (e.g., there is no targeted group for posts marked as non-offensive) and for the 5-th generated token, i.e. the first generated `[SEP]`.

At inference time, generated tokens are restricted to the valid choices (e.g. `[offY]` and `[offN]` for the first generated token, `[SEP]` for the 5-th, etc.). Generation is stopped after having generated first 5 tokens if post labeled as non-offensive or non targets a group. Otherwise, it is stopped after the generation of the in_group token.

2.2 Encoder-decoder model

Forward-only attention and the generative-only nature of a decoder-only architecture motivate us to try a different approach. We use BART (Lewis et al., 2019), a relatively small encoder-decoder model which shows natural language comprehension capabilities.

The tokenized post is provided as input to the encoder, where the bidirectional attention allows to contextualize each token with all others.

To classify the post, a pooled representation of the post, obtained by averaging the last layer's output of the encoder across the sequence dimension, is projected into a 5-sized vector by a two-layers MLP network (linear + tanh + linear), which is turned into 5 probabilities using a Sigmoid layer.

If the post is classified as offensive and the offense targets a group, the decoder is inputted with a start token `[START]` and outputs group and stereotype tokens, separated by a separator token and ending with an EOS token:

$$w_{G_1}, w_{G_2}, \ldots, \texttt{[SEP]},$$
$$w_{S_1}, w_{S_2}, \ldots, \texttt{[EOS]}$$

For the classification task, we minimize the KL divergence between the outputs of the MLP network ($x_{i,c}$) and the true classification labels ($y_{i,c}$)

$$\mathcal{L}_{\mathcal{CLS}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} D_{\mathrm{KL}}(y_{i,c} \| x_{i,c}) \quad (2)$$

where N is the batch size, and $C = 5$ is the number of classification features. No loss is incurred for variables that cannot take values due to earlier variable values.

For the generative task, group and stereotype annotations are tokenized and the decoder is fed with the start token concatenated with the expected output. We minimize the weighted cross-entropy introduced in the previous section (1). Posts annotated as non-offensive or that do not target a group do not contribute to the generative loss.

3 Data

Models are trained on the Social Bias Inference Corpus (SBIC) (Sap et al., 2020) encompassing 150k structured annotations of social media posts referencing over 34k implications regarding a thousand demographic groups. The SBIC dataset comprises five categorical label, ranging from [0, 1] for the classification task:

- Offensive: if the text is offensive [no, maybe, yes]
- Intentional: if the offense is intentional [no, probably no, probably yes, yes]
- Sex: if the the text has sexual content [no, maybe, yes]
- vs group: if the offense targets a group, e.g. a minority, or an individual [no, yes]
- in group: if the writer belongs to the targeted group [no, maybe, yes]

[1] We do not use a start token as we believe it is meaningless due to forward-only attention.

model	offensive *57.9% pos. (val)*	intent *51.6% pos. (val)*	sex *9.1% pos. (val)*	vs_group *64.4% pos. (val)*	in_group *8.7% pos. (val)*
Validation set					
GPT2	0.824	0.832	0.583	0.786	0.315
BART-bce	**0.845**	**0.849**	0.731	0.832	**0.370**
BART-kl	**0.845**	0.847	**0.743**	**0.836**	0.347
Test set					
GPT2	0.831	0.839	0.588	0.813	0.179
BART-kl	**0.860**	**0.860**	**0.727**	**0.850**	**0.293**

Table 1: Classification result per model. Bart outperforms GPT2.

The dataset also includes two free-text generation labels:

- group: the targeted group of the post
- stereotype: an explanation of the involved stereotype in the post

The dataset may also include information about the post's annotator, although this information is not consistently present.The dataset is highly imbalanced, with significantly more posts annotated as intentional or offensive compared to sex.

To enhance data quality, we implement a substantial data cleaning phase. We manually remove incorrectly annotated posts and semantically aggregate group labels (e.g. *muslims*, *muslim folks* and *muslim people* groups are mapped to *muslim people*). We believe that a smaller, more focused set of group labels would improve data quality and network accuracy.

We employed UAE-Large-V1 (Li and Li, 2023), which according to MTEB (Muennighoff et al., 2023) is one of the best embedding model for text similarity. Following specific rules, we halved the domain of text generation labels for targeted groups.

In addition, posts are preprocessed: html-escaped strings are replaced with the corresponding utf-8 strings, url are removed, emojis are replaced with their shortcodes, authors are removed from Reddit posts.

Since multiple annotations exist for the same post, we obtain another dataset by aggregating them and calculating the mean of the classification features and the set of groups and stereotypes.

4 Experimental setup and results

We train each model for three epochs on three different seed values, $42, 1234, 2023$, averaging results in order to test the performance of the models. Each model is trained for 3 epochs using AdamW (Loshchilov and Hutter, 2019) as optimizer. A linear learning rate scheduler is used, where the maximum value is $1e$-5 with a warm up fraction equal to 0.2 for GPT2 and 0.1 for BART. Gradients are clipped to 1.0.

GPT2 Embedding matrix is expanded with classification tokens (i.e. `[offY]`, `[offN]`, `[intY]`, etc.) and with `[SEP]` and `[PAD]` tokens. We initialize added tokens embedding using a Gaussian with mean and standard deviation of the pre-train embedding vector plus some small noise (Hewitt, 2021).

The weight of the Cross-Entropy are set to the inverse of the (binary) frequency for the classification tokens and to 1 for others.

At the end, GPT2 achieves satisfactory results both on validation and test set. It gets more than 80% F1 score, across all the seeds, on predicting whether a post is offensive or not and more than 80% on Rouge and BLEU of group prediction. Although these good results, it struggles on predicting if the author belongs to the targetted group and the stereotype associated to the post. More results in Tables 1 and 2.

BART We experiment by using two different classification losses: Binary Cross-Entropy and Kullback-Leibler divergence. KL-divergence shows slighlty better overall performances on validation set (Tables 1 and 2), although they are equivalent on classification metrics.

To binarize predicted classification features, we tune 5 different thresholds over the validation set with grid search on values $\{0.1, 0.11, ..., 0.9\}$.

Bart achieves slightly better results compared to GPT2, across the three seeds, scoring more than 84% on predicting whether a post is offensive or not and 89% on both Rouge and BLEU score on the targeted group. However, even this model struggles on predict whether the speaker is in the targeted group or the stereotype. More details can be found in Tables 1 and 2.

model	targeted group			stereotype		
	Rouge-L	BLEU	WMD	Rouge-L	BLEU	WMD
Validation set						
GPT2	0.873	0.876	0.13	**0.562**	**0.6**	**0.518**
BART-bce	0.884	0.886	0.118	0.521	0.585	0.543
BART-kl	**0.894**	**0.895**	**0.107**	**0.526**	**0.591**	**0.536**
Test set						
GPT2	0.539	0.597	0.39	**0.574**	**0.617**	**0.496**
BART-kl	**0.560**	**0.618**	**0.368**	0.531	0.602	0.530

Table 2: Generation result per model. Bart outperforms GPT2 on group generation, while GPT2 does better on stereotype predictions. Differences on group results between validation set and test set are due to groups reduction, non applied to test set.

Differences on group results between validation set and test set are due to data cleaning. Indeed, we only aggregate group labels on train set and validation set.

5 Discussion

Classification Both models achieve high scores on the prediction of offensiveness, intentionality and group targeted. Comparing the performance of GPT2 with those of BART, we can highlight that they are very similar, with slightly better results of the latter, especially on the prediction of sexual content.

The comparison of our result on test set with those in (Sap et al., 2020) shows that our BART gets better results on all the classification features by getting an improvement between 5 and 15 points on F1. Surprisingly, sex score is lower; we believe that increasing the number of training epochs will improve it.

Moreover, by comparing conditional probabilities on each classification features given the the post is offensive,

H	True	GPT2	BART
intentional	0.91	1	0.999
sex	0.15	0.26	0.14
vs_group	0.71	0.84	0.76
in_group	0.02	0.03	0.02

Table 4: conditional probabilities given the offensiveness of the post, i.e. $P(H|E = \texttt{"offensive"})$

we find that:

- models always label an offensive post as intentional. This limitation can be attributed to the challenging nature of the task, which can be difficult even for humans. Additionally, the limited availability of unintentional samples, inconsistency in annotations, and relatively small model sizes hinder the ability to generalize this complex concept.
- GPT2 assigns an higher probability than the true one to a classification feature if the post has been labeled as offensive. We assume that this behavior is more prevalent in the aforementioned model due to its auto-regressive nature.

Generation On generation tasks, we find that our two models obtain similar performances: BART performs slightly better on generating targeted groups, while GPT2 is better on stereotypes.

On the validation dataset, groups are aggregated by similarity, while on the test set, they are left untouched. The cleaning procedure seems effective as we obtain very good results on validation groups, keeping the output semantically intact. We believe that worse results on the test set highlight a weakness of the used metrics. Indeed, even if according to WMD our models outperforms the one of (Sap et al., 2020), we reach consistently worse results according to Rouge and BLEU. Therefore a better embedding matrix on WMD or a different metric that focuses more on the semantic meaning may lead to improved results.

By looking at Table 3, we see that most of the "errors" are not genuine mistakes, as the models predict the correct group in a syntactical different manner ((a), (d)). Additionally, they tend to extract groups from the post and struggles when they need to infer them (e). Also, they are bad when stereotype needs previous knowledge ((a) and (c)).

It's worth noting that predicted implications follows a fixed structure: <*predicted group*> *are* <*ad-*

	Post	Predicted Group	Reference Group	Predicted Implication	Reference Implication
(a)	Why colored soap create white bubble? Because Jews are white	Jewish people	Jewish folks	Jews are inferior	Jewish people deserved to be victims of Hitler
(b)	What's black and blue, and doesn't like sex? The kid in my trunk.	Black people	Rape victims	Black people are criminals	Trivializes harm to victims
(c)	Does God drink? Because I heard that if you go to his house in Texas, you'll get free shots.	Mass shooting victims	Christian folks	Trivializes harm to victims	Christians don't drink
(d)	What's the difference between / and a milk cow? There isn't one - Muslims keep doin' em both. Edit: Fixed the punch line, it's too late for this crap haha	Muslim people	Muslim folks	Muslims are terrorists	Has sex with animals
(e)	Last thing we want is them moving off the reserves.	black people	native american/-first nation folks	black people are worthless	native americans should not live off of reserves

Table 3: Model error on predicting targeted group or the stereotype

jective>. When the group or the stereotype cannot be easily extracted from the post, they generate the most frequent group (e) and/or the relative biased adjective (d) in the dataset. Post (b) is a clear example of how models cannot understand humor. The generation of the "*black folks*" group is influenced by the presence of the word "*black*" in the post. Indeed, what comes out from results is that there are some words that drift the classification or the generation towards mistake. When models encounter words as 'gay', 'black', 'women', 'muslim', etc... they tend to classify the post as offensive, even if it is not, and to generate text related to those words. We find no evidence of reasoning during the generation phase.

6 Conclusion

Our work explores the potential to classify and interpret potentially harmful or offensive sentences. Our models are effective in predicting the offensiveness of a post and in identifying groups and stereotypes. Although lack of reasoning, our models outperform the ones introduced in the paper of the dataset, predominantly on the classification, thanks to the Encoder-Decoder architecture. The generation of groups and stereotypes are caught only when they are explicit, i.e. can be extracted from the post, or when they are based on well-known biases (e.g. "muslims are terrorists").

We hypothesize that the limited size of the models contributes to their difficulty in dealing with this intricate task. In light of the challenges posed by limited model size, larger models may offer a viable solution for improving results in free-text generation. Furthermore, advanced pre-processing techniques could be employed to refine the quality of the training data. By meticulously removing typographical errors, and rewriting slang words and abbreviations, we can provide models with better data, enhancing their ability to learn, understand and generate meaningful text. In addition, pre-training models on a comprehensive resource like Urban Dictionary can further improve their understanding of slang and informal language usage.

7 Links to external resources

- Github repo
- Dataset

References

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

John Hewitt. 2021. Initializing new word embeddings for pretrained language models. https:/nlp.stanford.edu/~johnhew//vocab-expansion.html.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.