Sexist Stereotype Classification on Instagram Data

Team 15: Alok Debnath, Sumukh S, Neelesh Bhakt, Kunal Garg

TA: Pulkit Parikh

Abstract

Sexist stereotyping is a social phenomenon described as an over generalization of the attributes, behaviour, characteristics or features of a few people to their entire gender. In recent years, due to the rise in awareness campaigns, both on social media and in form of protests, it has come to light that sexism and sexist stereotyping is quite a common occurrence and it can have far reaching consequences on the victims and therefore must be curbed. The first step in curbing sexist stereotypes is to ask, "what makes a statement or comment sexist?" In this project we turn to Instagram as a source of comments and try to answer the questions from a machine learning perspective: "what features can be extracted that determine that a comment is sexist?" We scrape Instagram captions and comments associated with multiple hashtags and cleaned it up. We annotated the data in a binary fashion (sexist or not), and created an active learning model to annotate the rest of the captions. Our next task is to create a finer classification guidelines and use that to extract some relevant features from the same.

1 Introduction

"Guys are just stronger than girls", "Girls drive badly, make me a sandwich", these comments are considered unsavoury in contemporary society, as they promote stereotypes about certain genders based on generalizations that are demeaning and untrue. These comments are therefore classified as sexist stereotypes, and unfortunately these are quite prevalent in the social media landscape of today. While social movements bring these to light, we must do our part in order to take action on comments that perpetuate these stereotypes. In order to take action, however, we must have a metric to identify whether a comment is actually sexist or not, and this can become very tricky [1] given that:

1. Recognizing a sexist stereotype, by definition, requires real world information, which most machine learning algorithms do not actually have.
2. There are a combination of factors such as context of the comment, social considerations and so on.

However, the detection of sexist comments is quite crucial for a healthy and wholesome social media landscape which can be accessed by everyone without discrimination. Multiple social media outlets do a good job of using community guidelines in order to curb the presence and perpetuation of sexist content, among other discriminatory ideas. One of these sites and apps is Instagram, which is a social media site dedicated to posting and sharing photos with friends and followers and allows captions and comments on those photographs.

The ontological and contextual difficulties of data collection have been curbed according to contemporary literature by a thorough two-fold data creation mechanism on social media which allows for filtering of the relevant content, after which data is collected using the biases created by the data creation procedure [2]. Most studies focus on the idea of ambivalent sexism, an idea that points towards the sexism in everyday discourse [3]. There can be considered multiple perspectives of accounting for sexist concepts such as hate or offensive speech, body-shaming, misogynous language, mansplaining and so on, which adds a layer of difficulty to the problem at hand [4].

One of the notable problems with the method used by contemporary papers is the prevalence of Instagram's Community Guidelines, which are very strict and therefore reported captions, photographs and comments get taken down almost immediately. Therefore, the problem was approached from a slightly different perspective in order to accommodate for these challenges. The dataset presented as a part of this submission pertains to captions that are *about* sexism, sexist behaviour, misogyny, rather than being the comments themselves. The idea is that in extracting the relevant features, we can isolate the patterns within sexist comments and tweets, while also analyzing the feature of comments and captions which are on the subject, but not sexist themselves, in order to red flag the former and promote the latter. The idea here is that more people should be able to read the stories which are about sexism so that awareness can be spread on these topics.

So far, our task has been the collection and cleaning of data from various hashtags. The data collected from these hashtags was then binary classified into whether the comments were sexist (or about sexism) or not. A supervised active learning model was used to classify the dataset based on a small annotated seed dataset of 200 captions. The final dataset consists of X captions and comments.

Here on out, we will follow a two stage annotation mechanism based on some synthetic data that we provide as well, to differentiate comments about sexism from comments which are sexist along different lines such as role or attribute stereotyping. Then, we shall create a new model for an unsupervised analysis of the collected data, given an absence of tags, in order to answer the question "What makes a comment or caption sexist?"

2 Data Collection and Annotation

In this section, we explain the data collection mechanism for Instagram captions and comments.

2.1 Data Collection Mechanism

Using the `instagram-scraper` API, we collected data form 10 hashtags such as `#bloodymen`, `#boys`, `#everydaysexism`, `#girls`, `#guys`, `#manspalining`, `#metoo`, `#sexism`, `#sexist` and `#slutshaming`. We collected $10,100$ captions from each hashtag. However, the data can be multiple emojis, other hashtags, mentions, in various different languages and quite noisy. Therefore, the captions and comments were processed, based on language, removal of the noisy elements to get less noisy text.

From these, the final dataset of 6238 captions and comments was created.

2.2 Data Annotation Guidelines

The data was annotated based on a simple metric: Do these comments and captions pertain to sexism or not? A comment or caption is defined to pertain to sexism if it is either sexist itself (for example: "Women belong in the kitchen") or about sexism (for example: "I was told to shut up because women don't know science"). This binary classification was manually annotated for 200 captions and comments, with an inter-annotator agreement rate of 94% across four annotators.

We then used an active learning approach to annotate the rest of the captions and comments, in batches of 200. We follow a pool-based active learning mechanism. We define a pool of 200 new captions and mentions, over an $80-20$ split of the labelled data. We use entropy selection for the k most representative samples as entropy can be easily calculated when using text features such as tf-idf. We use tf-idf features for the texts (due to presence of excess noise), and an ensemble of classifiers for our task.

Using tf-idf features for now was done in order to interpretably determine the features which are most useful in the analysis of sexist stereotypes in the data. This is useful in order to supervise the annotation being done by the active learning framework.

3 Findings and Methods

Using this, we found that the active learning mechanism provides successively better accuracy (from approximately 74% to 88% over multiple pools and increase in the size of the dataset. We find also that the higher accuracy exploits basic correlations based on the TF-IDF inputs in order to tag the data, therefore the longer the text, the easier it is to tag it accurately. Furthermore, it is more difficult to clean data based on the current data extraction method.

In the future, we plan to train a model that provides classification for a multi-class setting. The annotation shall be made for the larger dataset and much like many SoTA studies, we use a small-Insta and big-Insta caption dataset with 695 captions and comments in the former, 5545 in the latter. We plan to use the GloVe embeddings and use that in a sentence classification network (without attention). The details of the implementation are still being discussed.

4 Link to Implementation

You can find the notebooks with the data at: `https://github.com/AlokDebnath/Sexist_Stereotype_Classification`.

5 The Road Forward

Oct 21	Major Project 2nd Deliverable (create a dataset, with the binary tags for each post, and have a baseline binary classifier)
Oct 22 - Oct 31	Create labels for the multi-class classifier, and work on classification of the sexist posts between attribute and role stereotype (can be worked on in parallel)
Nov 1 - Nov 7	3 way classifier - A single classifier for non-sexist stereotype, role stereotype and attribute stereotype
Nov 1 - Nov 5	Fine tune the models
Nov 6 - Nov 11	Compiling results, preparing report and presentation
Nov 11	Major Project 3rd Deliverable

Effort will be put in so that work is done well ahead of the above schedule, so that a simple GUI can be possibly implemented.

References

[1] Foster, M. D. (2015). Tweeting about sexism: The well-being benefits of a social media collective action. British Journal of Social Psychology, 54(4), 629-647.

[2] Fox, J., Cruz, C., & Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. Computers in Human Behavior, 52, 436-442.

[3] Jha, A., & Mamidi, R. (2017, August). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In Proceedings of the second workshop on NLP and computational social science (pp. 7-16).

[4] Anzovino M., Fersini E., Rosso P. (2018) Automatic Identification and Classification of Misogynistic Language on Twitter. In: Silberztein M., Atigui F., Kornyshova E., Métais E., Meziane F. (eds) Natural Language Processing and Information Systems. NLDB 2018. Lecture Notes in Computer Science, vol 10859. Springer, Cham