Studying Sexism/Misogyny on Instagram

Aditya Agarwal, Aniruddha Deshpande, Nimmi Rangaswamy
aditya.agarwal@research.iiit.ac.in, aniruddha.d@research.iiit.ac.in, nimmi.rangaswamy@iiit.ac.in
20161104, 20161058, IIIT Hyderabad

Abstract

This project of Toxicity Detection in Instagram aims to study and help detect instances of cyber bullying in photo-sharing networks, with an eye on developing early-warning mechanisms for the prediction of posted images vulnerable to attacks. The study was narrowed down to the domain of Detecting and predicting Misogyny and Sexism on the Instagram platform. We plan to correlate the features of the person(s) in a given Instagram post to that of types of Sexist/Misogynistic Comments on that particular post. This will be done using a network which is trained to take in the Image Features and predict what type of sexist comment the post may be prone to. Thus we would be correlating features such as BMI, Age, Race, Emotion and Clothing with various types of Sexist Classes such as Body Shaming, Slut Shaming, Sexual Harassment etc. Our approach to this task is explained further in detail in this report.

1 Introduction

No app is more integral to teens' social lives than Instagram. According to a recent study by the Pew Research Center, 72 percent of teens use the platform, which now has more than 1 billion monthly users but according to a recent Pew survey, 59 percent of teens have been bullied online, and according to a 2017 survey conducted by Ditch the Label, a non profit anti-bullying group, more than one in five 12-to-20-year-olds experience bullying specifically on Instagram. The reason for this is the velocity and size of the distribution mechanism that allow rude comments or harassing images to go viral within hours. Like Twitter, Instagram makes it easy to set up new, anonymous profiles, which can be used specifically for trolling. Most importantly, many interactions on the app are hidden from the watchful eyes of parents and teachers, many of whom don't understand the platform's intricacies.

Instagram has pages like Four Year Party and College Nationwide boast tens of thousands of followers interested in getting an inside peek into the college lifestyle. Four Year Party has over 81,500 followers and emphasizes the tagline, "We are not here for the credits, just for the parties!" and College Nationwide encourages its 57,600 followers to check out "Hot chicks and rad pics." With such a broad reach of a very specific demographic, the influence of these two pages is something to consider. But what kind of influence do these pages have? Four Year Party and College Nationwide share photos and short videos to tens of thousands of college-aged people on a daily basis. These posts receive thousands of likes, comments, and shares. However, the messages these pages are promoting are problematic in that they reinforce the racial-ized and gendered inequalities present in US culture. Through a textual analysis of the two Instagram pages, hegemonic themes emerge. These themes are as follows: objectification of female college students, submissiveness of female college students, and emphasis on a young, white collegiate experience.

A similar incident in India that occured recently was the leaking of certain group chats like "Boys Locker Room" which had extremely sexist, offensive and vulgar comments that objectified and treated women as sexual objects. These comments were not limited to just adults, even underage and teenage girls were victimized.

Thus we aim to detect instances of misogyny/sexism through analysis of media content which is an important and challenging task, as the connection between an image and its context is unclear. Thus, we planned to deal with the Image and the Comments separately and then combined the two to study the correlations that emerge when an image is posted with the sexism classes that we identified.

2 Development Stages

Stage - 1: Crawling Instagram and Making a Dataset
Stage - 2: Feature Vector Construction and Deep Learning models
Stage - 3: Observations on Dataset and Model Evalutation
Stage - 4: Regression Networks for Predicting Sexist Comments

3 Crawling Instagram and Making a Dataset

The first task was to crawl Instagram and make a Dataset of images. Since the domain of the project was narrowed down to the correlation of Image features of an Instagram post to that of sexist comment labels for a given post, we decided to refer to the paper [1] which claimed that younger women tend to be targets of Sexist Hate Speech. Amongst them, women public figures with higher public status and fame are potentially even bigger targets to Misogynistic Hate speech.

Several papers were also looked into in order to categorise the type of Sexism/Misogyny which can be observed in the comments. Out of these, two papers - [2], and [3] , stood out the most and involved annotation schemas that made use of in-depth categories to classify Sexism. [2] also confirmed the fact that Publicly feminist influencers were heavily prone to Misogynistic Hate Speech.

Figure 1: Examples of Categories Influencer's Chosen for our Dataset: Body Positive Influencers - ***@curvesbecomeher***, Fashion Bloggers - ***@komalpandeyofficial***, Influencers who preach Women's Rights - ***@salonichopraofficial***, Actors - ***@radhikaofficial***, Non-Indian Influencers - ***@hannahwitton***, Miscellaneous Influencers - ***@ind0ctrination***

Based on these papers we curated a dataset consisting of 50 Instagram influencers's top 100 most commented posts and their respective comments (about 1000 per post) were gathered. The Following sections describe the criterias and sub-categories based on which we chose the influencers for this dataset. The posts were limited to Images, therefore video posts weren't considered for this task. Instagram also has a feature which lets the user upload multiple photos per post, and this feature was taken into consideration while downloading the images and therefore, the set of comments on that post could be applied to every single image present in that given post.

3.1 Influencer's Chosen for the Dataset

Based on the above claim from [1] we curated a list of 50 Instagram influencers whom we thought are prone to such hate speech. A lot of them also tend to be the type of influencers who shared their feministic thoughts frequently and in fact had gotten famous as influencers for being open when it comes to discussing issues involving feminism.

Figure 1 summarizes the categories of the influencers we thought would be prone to Sexist Hate speech:

3.1.1 Figures who preach Body Positivity:

These types of influencers are the ones who share their thoughts on the topic of asserting that all people deserve to have a positive body image regardless of how society and popular culture view ideal shape, size, and appearance. Some of the influencers in this category also tend to fall under the category of Plus sized models. We think these influencers are prone to hate speech related to Body Shaming and perhaps even Slut Shaming.
Example: ***@petite_and_curvy***

3.1.2 Fashion Bloggers / Models:

These influencers as the name suggests discuss fashion trends, and involve in fashion modelling. We think that their interest in fashion trends may end up being considered as obscene to some and hence will tend to have comments related to Slut Shaming and Objectification.
Example: ***@komalpandeyofficial***

3.1.3 Influencers who discuss Women's Rights / Sex Positivity:

These influencers discuss various aspects of feminism, calling out misogynistic ideals of the society. Some of them

also discuss Sex positivity which involve topics supporting LGBTQ, Polyamory, Asexuality, etc. These influencers are prone to several categories of Sexism involving complete Discredit of their ideas, Threats of Violence, Asserting Dominance, Sexual Harassment, Slut shaming and Stereotype.
Example: *@salonichopraofficial*, *an actress turned Instagram influencer who heavily preaches women's rights equality involving topics like "Free the Nipple Movement."*

3.1.4 Actresses and Models:
Actresses and Models have been known to be victims of criticism and hate speech towards maintaining their body and hence are prone to Body Shaming. Since actressing and modelling goes hand in hand, the categories of Slut shaming, Objectification and Sexual Harassment also come into play. Age based misogyny also has been frequently observed which may even be used to discredit the fame an actress has garnered overtime.
Examples: *@sunnyleone, @malaikaaroraofficial*

3.1.5 Non-Indian Influencers:
The dataset isn't limited to just Indian influencers as we wanted to bring the racial discrimination aspect one can find on Instagram comments as well. Also note that all the categories mentioned aren't mutually exclusive.
Example: *@hannahwitton*

3.1.6 Miscellaneous Influencers:
These influencers include other various types of influencers which may not fall in the above categories. One of the influencers who falls in this category that we would like to emphasize on is ***@ind0ctrination*** who by profession is a doctor but blogs about her life as a Transgender who identifies herself as a woman. Influencers like her are heavily prone to discredit and perhaps even body shaming.

Thus, after the above analysis, we had the top 100 posts and the respective comments for each post for 50 Instagram influencers as our dataset.

4 Feature Vector Construction and Deep Learning Models
This section is divided into three subsections, namely Image Feature Extraction, Comment Label Classification and the Correlation of the above two. We combined the above two subsections to obtain the correlation betwen Image Features and the Comments Labels.

4.1 Image Feature Extraction
This subtask is responsible for extracting features of the person(s) within the image of an Instagram post. Video based Instagram posts were ignored while making the dataset. Instagram also has a feature of adding multiple images to a given post, and the task of image feature extraction was done for each of these images. We chose some basic features of a person common to every image and essentially common to every person. We plan to make use of the following image features of the person(s) in a post:

- Age
- Ethnicity
- Facial Emotion
- Clothing Features
- Body Mass Index

Our decision for choosing these features stems from the real world intuition of what aspects of a person are usually targeted by a certain Sexist comment. Racist discriminations originate from Ethnicity of a person. Aspects of Sexism like Sexualisation, Slut shaming usually come from what a person wears. Weight of a person is a common target for a Body shaming comment. Age of a person could be essential in predicting a person's role - younger people tend to be students, elder women tend to be in a working/motherhood role. Aspects like Hostile Work environment, Role stereotyping, take the focus on the role of a person in society. Age works like a good intermediate representative of that role. These correlations are further explained in Section 5. Following sections explain the subsystem's responsible for Image Feature extraction.

4.1.1 Age, Ethnicity, Facial Emotion
We used a lightweight facial analysis framework written in Python called Deepface to analyze the demographic features of a person(age, emotion and race). It is also capable of predicting Gender of a person but our dataset contains people belonging to only the feminine gender. This DeepFace model was developed by Facebook Research and it's design is based on the paper [4].

The model is trained on over 4 million facial images belonging to about 4000 identities. The paper revisits both the alignment step and the representation step in the conventional pipeline of the modern face recognition that consists of four stages: **detect** $\rightarrow$ **align** $\rightarrow$ **represent** $\rightarrow$ **classify** by employing explicit 3D face modeling in order to apply a piecewise affine transformation, and derive a face representation from a nine-layer deep neural network. This deep network involves more than 120 million parameters using several locally connected layers without weight sharing, rather than the standard convolutional layers.

Thus using this model we were able to analyze the Age, Ethnicity and Facial Emotion of any human given an image. There were situations wherein the face of the person in an image is blocked by some object. DeepFace is unable to predict demographic features in such situations. In such cases, default values were assigned using the images wherein DeepFace could actually make predictions. Following criteria was used to assign default values:

- **Default Ethnicity:** Most frequently occurring ethnicity for a given user's images.
- **Default Facial Emotion:** Most frequently occurring emotions for a given user's images.
- **Default Age:** Average Age of the user based on already predicted age values.

4.1.2 Body Mass Index

To calculate the BMI, a pretrained model based on the medium article [5] was used. It was trained on a range of 4000 images, each of a different individual, taken from the front of the subject. The BMI of each training sample was calculated from the subject's height and weight (BMI is weight in kg divided by the squared height in meters). The image was then preprocessed by adding a margin of 20% i.e. the added margin captured features like the upper forehead, the ears, and the neck, which were useful to the model in predicting BMI. The image was also augmented to increase the size of the training set.

The model was created using the Keras ResNet50 class. The ResNet50 architecture was chosen so the weights generated by an extensively trained age classifier from the age and gender project could be used in transfer learning and also because ResNet (residual network) architectures are known to be good models for facial image recognition. Transfer learning was done to take advantage of the weights from the age classifier network as these should have been valuable for detecting lower-level features of the face to be used in predicting BMI. The age network was given a new linear regression output layer (outputting a number representing BMI) and was trained using MAE as a loss function.

It must be noted that this model is heavily dependent on the facial angle of the subject. There were situations wherein the face of the person in an image is blocked by some object. Similar to the DeepFace model, the BMI model is unable to predict BMI values in such situations. In such cases, default values were assigned using the images wherein the model could actually make predictions. Following criteria was used to assign default values:

- **Default BMI:** Average BMI of the user based on already predicted BMI values.

4.1.3 Clothing Features

This feature extraction required a lot of training and effort as there was no PreTrained model available. Thus, in order to extract clothing features of the person in a post, we trained a model based on the Medium article [6]. The article made use of a subset of the richly annotated **Deep Fashion Dataset** [7] and trained a classifier over it. The original Deep Fashion Dataset contained over 800,000 diverse fashion images ranging from well-posed shop images to unconstrained consumer photos. It is annotated with rich information of clothing items. Each image in this dataset was labeled with 50 categories, 1,000 descriptive attributes, bounding box and clothing landmarks. Deep Fashion also contained over 300,000 cross-pose/cross-domain image pairs. The Deep Fashion dataset can be summarized using Figure 2.

The Dataset used here had 290000 images and 46 classes and hence we had to use Transfer Learning because we were running Out of Memory. To overcome this we used the ResNet50 model that was pre-trained with ImageNet, but we did not train all layers in this model from scratch. After freezing the earlier layers which represent low-level features as weights such as line detector and pattern detector, we trained the layers which represent higher level features -more specific to data- by optimizing the loss function with low learning rate.

Top 2 clothing label predictions were used in the final image feature vectors.

4.2 Comment Label Classification

For the comment label classification, we had to come up with certain classes of Sexism/Misogyny. We studied the AMI (Automatic Misogyny Identification) Dataset [2] that involved the classes - *discredit, stereotype, objectification, sexual harassment, threats of violence, dominance, derailing.* We initially planned to use a model trained on this dataset to classify the comments on Instagram posts, but the dataset wasn't available openly. This meant that we would have to annotate the comments manually based on these labels which would be a very long and tedious task. Thus we found an alternative dataset used in the paper [3]. This paper is authored by two students at IRE Lab, IIIT Hyderabad. They provided us with the dataset for training but it cant't be made public without the authors' permission. The classes involved in this paper are described in Figure 3.

To the dataset used in [3], we added Non - Sexist comments which were manually annotated over the comments observed for the user ***@evyan.whitney***. This was done because the original dataset did not contain Non - Sexist comments. Data processing was done using standard procedures as per [8]. Emoticons and Emojis were converted into text as well.

We then used FastText to train a classifier on this dataset to classify all comments of each post of each user. FastText is a library for learning word embeddings and text classification created by Facebook's AI Research (FAIR) lab. The model allows to create an unsupervised/supervised learning algorithms for obtaining vector representations of words for any custom dataset, and the same can be understood from [9] and [10]. Facebook makes available pretrained models for 294 languages. If the classified comment label for a given comment is Non - Sexist, then the comment is classified as Non - Sexist, and if the comment Label is sexist then the top 3 label predictions are chosen.

4.3 Final Feature Vector Creation

In order to create the final dataset for classification and correlation study, we combine the image features retrieved using the above explained subsystems for each image in the following manner:

x (Image Features): [age, race, emotion, bmi, clothing1, clothing2]

Here, age and bmi are positive integral and float values (upto third decimal) respectively. Remaining features are categorical features (These are based on the categories that each of the above Image Feature extraction models

Figure 2: Deep Fashion Dataset - Clothes Categories

Category	Description
Role stereotyping	Socially constructed false generalizations about certain roles being more appropriate for women; also applies to such misconceptions about men
Attribute stereotyping	Mistaken linkage of women with some physical, psychological, or behavioral qualities or likes/dislikes; also applies to such false notions about men
Body shaming	Objectionable comments or behaviour concerning appearance including the promotion of certain body types or standards
Hyper-sexualization (excluding body shaming)	Unwarranted focus on physical aspects or sexual acts
Internalized sexism	The perpetration of sexism by women via comments or other actions
Pay gap	Unequal salaries for men and women for the same work profile
Hostile work environment (excluding pay gap)	Sexism encountered by an employee at the workplace; also applies when a sexist misdeed committed outside the workplace by a co-worker makes working uncomfortable for the victim
Denial or trivialization of sexist misconduct	Denial or downplaying of sexist wrongdoings
Threats	All threats including wishing for violence or joking about it, stalking, threatening gestures, or rape threats
Rape	FBI's expanded definition of rape
Sexual assault (excluding rape)	Any sexual contact without consent; unwanted touching
Sexual harassment (excluding assault)	Any sexually objectionable behaviour
Tone policing	Comments or actions that cause or aggravate restrictions on how women communicate
Moral policing (excluding tone policing)	The promotion of discriminatory codes of conduct for women in the guise of morality; also applies to statements that feed into such codes and narratives
Victim blaming	The act of holding the victim responsible (fully or partially) for sexual harassment, violence, or other sexism perpetrated against her
Slut shaming	Inappropriate comments made about women 1) deviating from conservative expectations relating to sex or 2) dressing in a certain way when it gets linked to sexual availability
Motherhood-related discrimination	Shaming, prejudices, or other discrimination or misconduct related to the notion of motherhood; also applies to the violation of reproductive rights
Menstruation-related discrimination	Shaming, prejudices, or other discrimination or wrongdoings related to periods
Religion-based sexism	Sexist discrimination or prejudices stemming from religious scriptures or constructs
Physical violence (excluding sexual violence)	Domestic abuse, murder, kidnapping, confinement, or other physical acts of violence linked to sexism
Mansplaining	A woman being condescendingly talked down to by a man; also applies when a man gives an unsolicited advice or explanation to a woman related to something she knows well that she disapproves of
Gaslighting	Sexist manipulation of the victim through psychological means into doubting her own sanity
Other	Any type of sexism not covered by the above categories

Figure 3: Description of Categories of Sexism used in Dataset

support) and their numerical values are assigned based on the enumeration as per the Tables 3, 4, 5 in the Appendix of the paper. clothing1 and clothing2 represent top 2 clothing type predictions.

Example: *x = [27, 4, 2, 29.561, 17, 39] represents an Image having features - [age: 27, race: indian, emotion: happy, bmi: 29.561, clothing1: tee, clothing2: dress].*

Corresponding to the x values, ie. the Image features, there are the y values which incorporate the Sexist comments. As explained in Section 4.2, FastText was used to retrieve top 3 sexist comment predictions for each comment on a post. Later frequencies corresponding to each Sexist Label type were calculated for a given post and were used to calculate the following weighted feature vector:

y (Sexist Comments Feature Vector): A weighted feature vector of size 21 where each value represents the percentage of comments belonging to a given sexist label for that post.

The percentages were scaled to 1 instead of scaling to 100. Within the vector of length 21, the positions of each weight value (ie. percentage) correspond to the index Table 2 enumerated in the Appendix. The weight is calculated using the following equation.

$$y[index] = no\_of\_comments(comment\_class[index]) / total\_no\_sexist\_comments$$

Example: *Consider a y value of length 21. Here consider the index value = 13. This index corresponds to the Mansplaining class as per the Appendix. Consider for the given post, we have 20 comments under the Mansplaining category and a total of 50 sexist comments. Therefore:*

$$y[13] = no\_of\_comments(comment\_class[index]) / total\_no\_sexist\_comments$$
$$y[13] = no\_of\_comments(Mansplaining\ Class) / total\_no\_sexist\_comments$$
$$y[13] = 20/50 = 0.4$$

These x and y values formed the final feature vector dataset. These were then used to study the correlations explained in Section 5.2. These values were also used to train the regression models used in Section 6 to create the final classifier. This classifier then predicts the type of Sexist comments a given image is prone to.

5 Observations on the Dataset and Model Evaluation

This section of the Report describes the observations we made on the Final Dataset after the Feature Extraction Steps that were carried out based on the section described above and describes the final Regression Model we used. Each of the subsystems in our methodology in this study were independent of each other and as described were even trained on different datasets. The DeepFace architecture was trained on a dataset of 4 million images curated by Facebook Research, The BMI prediction model was trained on a dataset which isn't publicly available, Clothing Features Extraction Model is trained on DeepFashion dataset, and finally the comments for each of the post were classified based on the dataset used in [3] using fasttext architecture.

Neither of these models ever made use of the Instagram user images for training as for that they will have to be annotated based on the annotation schemas of each of the above mentioned training datasets. This led to some misclassifications of an image and even under-classifications for some certain categories.

5.1 Dataset Distribution

If all the images were in fact manually annotated for the task, then all of the separate image feature extraction models could have been combined into a single CNN based architecture capable of predicting all the Image features simultaneously and not serially and hence independently. Parallely, an LSTM would be playing its role in embedding and classifying each comment into custom sexism comment classes. This methodology inherently brought in class based biases that seep in because of both the subsystem training models and the inaccurate predictions of these subsystems on the Instagram User Images.

Figure 4 explains the distribution of the dataset with respect to each of the Demographic, Clothing, BMI classes, and the Sexist Comment Classes after the step of feature extraction for both the images and the comments for the training dataset. Following observations about the bias and the inaccuracies in classifications were made:

5.1.1 Demographic Features:

Instagram being an image based social media platform tends to have images of either of the two categories - Happy and Neutral. Most users don't tend to share an image with other emotions and same is the case for the users included in our dataset. This can be observed in the graph distribution for emotions as well. This led to an observation that there will be no correlation between the emotion of the person and the type of sexist comment the photo receives. This was an expected behaviour as usually no sexist comment is based off of the emotion of the person in the photo.

The **Age classifier** on the other hand had a bias of classifying people into age groups of 30-35. This sort of inaccuracy is especially confirmed in the case of user ***@gretathunberg*** who is 17 years of age but is classified as a person in their 30s. The most inaccurate of all was the **Race prediction model** which showed a higher bias towards classifying people of Indian ethnicities into other classes. Inherently Indian users with a lighter complexion were heavily classified as belonging to the White Race and the ones with a darker complexion were classified under the Latino Hispanic race that can be seen in Figure 4. This is the reason why we see a higher distribution in both of these classes irrespective of the fact that the user list consists mostly of Indian Instagram influencers.

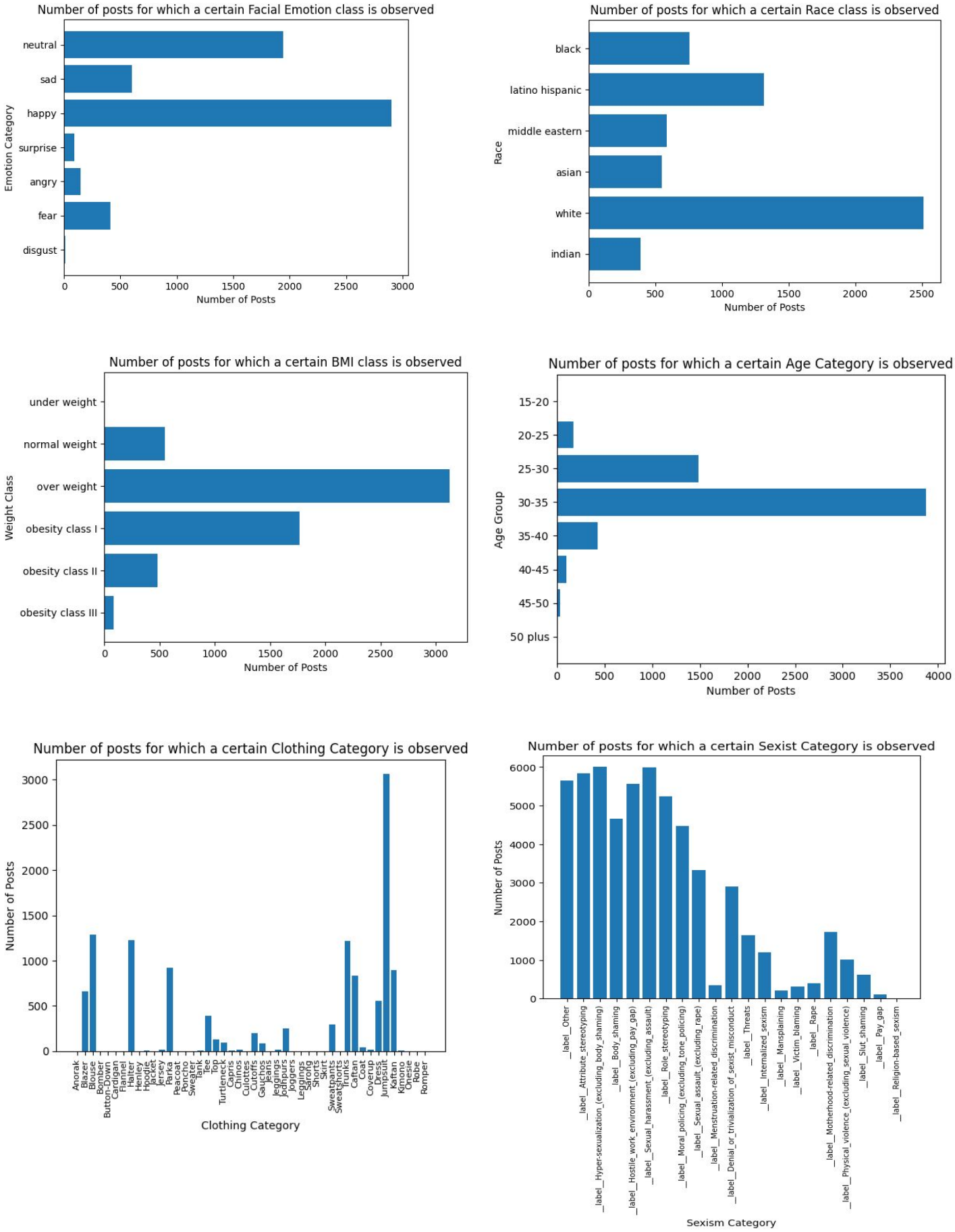

Figure 4 - Distribution Statistics of the Dataset: Facial Emotion Distribution, Racial Distribution, Weight wise distribution, Age wise distribution, Clothing type distribution, Sexist comment distribution (In order left-to-right top-to-bottom)

5.1.2 Clothing Features:

Due to hardware limitations the Clothing Feature dataset could only be trained up to a 60% training accuracy. On top of this the inaccuracies due to the novelty of the Instagram Images towards the model brought in its own limitations. Most Instagram images and selfies tend to be waist level which we think brought in the misclassifications as the lower half of the outfit wasn't visible. Due to this a lot of images were classified into the Jumpsuit/Dress categories as can be seen in Figure 4 as it is a difficult task for the classifier to predict the outfit type just with the waist to collar level outfit features.

A full size image was classified fairly well into a category close enough to the original outfit. A common example of such a classification was classifying a Bikini outfit photo into DeepFashion categories like Blouse and Trunks. This is why to further better the ImageFeatures we took into consideration top 2 clothing predictions instead of just the one.

5.1.3 BMI Features:

The main inaccuracies observed in this subsystem was the over classification (i.e. misclassification) of perfectly fit people into the overweight category as can be seen in Figure 4. This led to an overshot in the number of people in the overweight category instead of an expected higher number in the normal weight category.

5.1.4 Sexist Comments Classification:

Originally [3] didn't have a Non-Sexist category in their dataset. Their dataset was curated using stories from a website consisting of everyday sexism stories [1]. They only had stories involving Sexism/Misogyny. Another property about their dataset was that they didn't involve emojis. Instagram being a social platform heavily uses emojis in comments and hence can't be ignored. Although these emojis were converted into a textual form during the preprocessing step, they originally weren't a part of the training dataset. Therefore, inaccuracies may have arisen in the comment classification step due to presence of emojis. Another reason we saw heavy bias towards certain classes could be due to Non-English comments written in Latin Scripts.

5.2 Correlation of Image Features with Sexism Classes

The correlation between an Image Feature and the Sexism Comment Categories was done by calculating class wise summation of weighted Y feature vector values (from Section 4) for each class of each Image Feature Type. Since each of the feature vectors' weights add up to 1, uniformity is maintained. We saw zero correlation for some of the Sexism comment categories because of the inaccuracies brought in by the classifier model.

In Ideal cases we expected strong correlations between certain Image Features and certain Sexism Comment Categories. Although such correlations weren't always observed due to the above mentioned biases and inaccuracies, some of the positive correlations that we were expecting out of the study can be discerned. Figure 5 summarises the observed correlations of Sexism categories with respect to that of each Image Feature. Following is the description of the correlations that we did observe:

5.2.1 Correlation Between Demographics and Sexism Categories:

As mentioned above, since emotions didn't play a role in deciding on the sexism category classifications, no correlations between Emotion Categories and Sexism Categories were observed and hence is omitted. When it comes to a racial based correlation study, positive correlations are observed for Hypersexualisation and Sexual harassment type of comments for White and Latino Hispanic Races, but not as much for other races. The overtly positive correlation for the White race came due to the excessive misclassification of people of Indian ethnicities into the White Race class.

This suggested that influencers of Indian descent were a lot more prone to Hypersexualisation and Sexual harassment. This was an observation we were expecting and the dataset supported our expectations. We noticed that the dataset was more or less distributed in the 20-40s age group. This went hand in hand with the age group distribution of Instagram as a social media platform as well [2]. We observed a higher positive correlation for Hostile Work Environment and Motherhood Related Discrimination Comment Classes to that of images with age groups in the range 30-35. This corroborated with the statistics of average Motherhood ages [3] and a working class woman's age as well (with the intuition that women in the 20-25 age group are still in the education stages of their lives.) This observation corroborated with respect to real world intuition.

5.2.2 Correlation Between Weight Classes/BMI and Sexism Categories:

Weight Classes (using BMI values) by intuition should play a role in Body Shaming based comment categories. An excessive positive correlation was observed between Body Shaming comment category and Overweight class because of excessive misclassification of fit user images into Overweight classes. Even on regularizing this anomaly in distribution of classes, we still observe higher positive correlation between images present in Obese Class I/Overweight classes and Body Shaming comments, hence confirming our intuition.

5.2.3 Correlation Between Clothing Features and Sexism Categories:

The excessive misclassification of clothes in the Jumpsuit category and the under-classification of clothes in many

[1] everydaysexism.com

[2] https://www.statista.com/statistics/325587/instagram-global-age-group/

[3] https://www.statista.com/chart/15144/when-women-become-mothers/

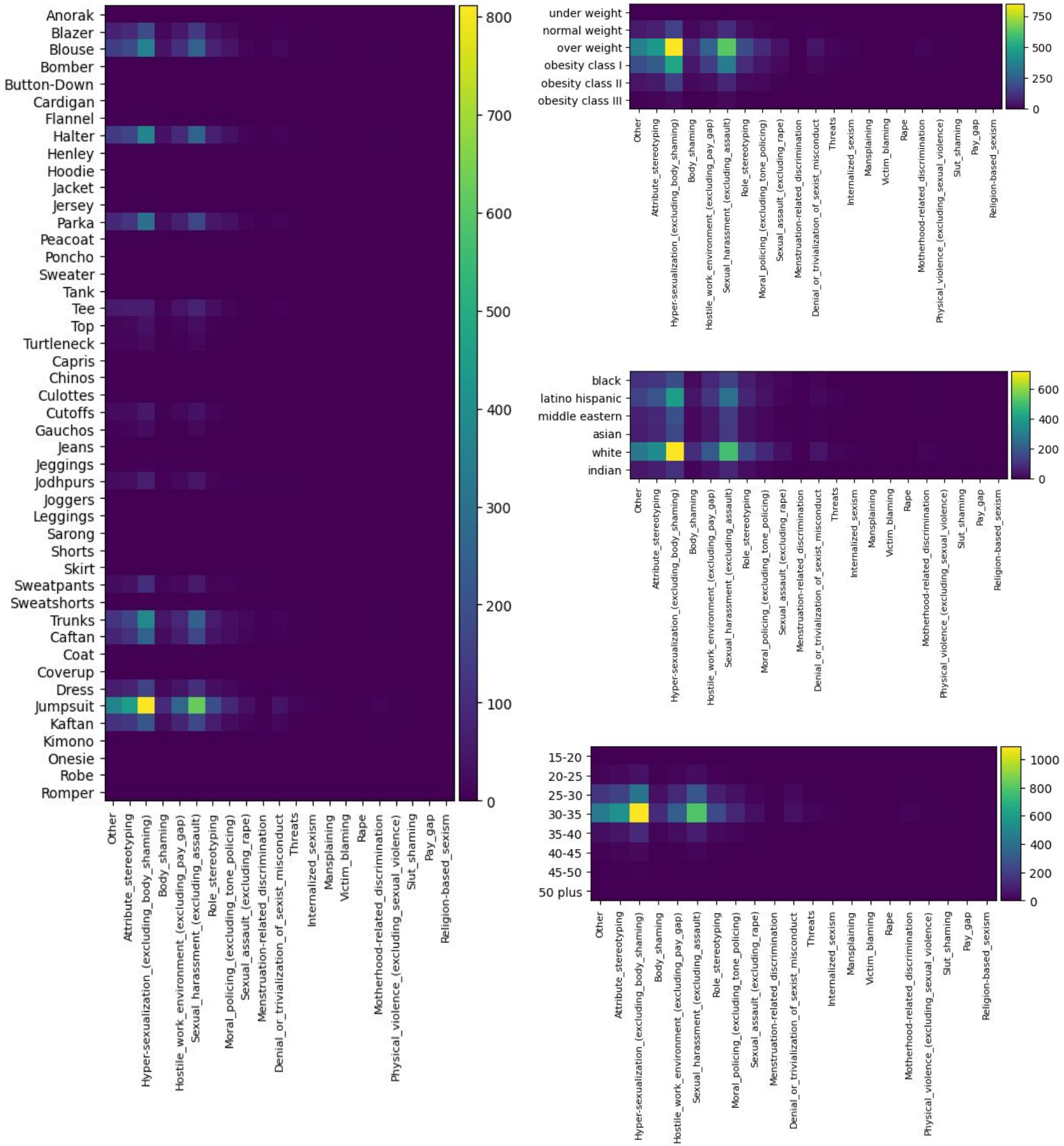

Figure 5 - Correlation Matrices between Sexist Comment Label type and Clothing Types, Weight Classes, Race types, Age Classes (in order left-to-right top-to-bottom) respectively.

other categories was most observed in the Clothing Feature based classification. This led to heavy zero correlation between some clothing categories as they were heavily under classified due to inaccuracies of the clothing feature classifier model. Irrespective of this issue we did observe some of the positive correlations that we were expecting based on our real world intuition. Intuitively we observed cases of Hypersexualisation, Sexual Assault, Sexual Harassment, and Body Shaming related sexist comments in situations when a person was said to have been wearing 'Revealing Clothes' as per our society. A similar observation was also made in our dataset. We saw a heavy positive correlation when it came to the above types of sexist comments and clothing types. Sexist comments of Hypersexualisation, Sexual Assault, Sexual Harassment, and Body Shaming types were observed more in outfits which showed a lot more skin like a Blouse, a pair of Trunks, or a Halter rather than outfits which cover the skin like a Blazer, a pair of Sweatpants, or a Tee.

6 Regression Networks for Predicting Sexist Comments

Based on the above created dataset, we trained 6 regression models which predicts the weighted Y vector based on the extracted Image features as explained in Section 4. These were trained with an 80-20% Train-Test split. Table 1 summarizes the networks' hyperparameters and their respective error loss values over the test set.

Regression Model	**Hyper-Parameters**	**Mean Absolute Error**	**Mean Squared Error**
Random Forest	Number of Trees: **50**	0.021941	0.002260
Decision Tree	Max Depth: **5**	0.021839	0.002197
KNN	Nearest Neighbours: **5**	0.023660	0.002572
Linear Regression	Fit Intercept: **True**, Normalize: **False**	0.022276	0.002283
MLP Regression	Max Iterations: **600**, Solver: **SGD**, Learning Rate: **Adaptive**, Batch Size: **200**	0.023053	0.002407
Gradient Boosting	Number of Boosting Stages: **100**, Loss: **least squares regression** Learning Rate: **0.1**, Criterion: **Friedman MSE**, Max Depth: **3**	0.021259	0.002085

Table 1: Regression Model Hyper-Parameters and Error Loss Values

These discussed regression models achieved very low error loss values as they considered the training dataset as their basis while training their network weights. They just considered the inaccuracies of the training dataset as the correct true values and trained accordingly and hence ran independently of the Image Feature Extraction Models. This was an expected behaviour.

To actually classify a certain test image, the model takes in the image and predicts the image features for that given image. It should be ensured that the face is visible clearly in the image. In the cases of images with faces covered with some object, the user may opt to manually add their image features. These image features are then sent as an input to all the 6 regression models to predict the weighted Y vector.

Higher the weight for a given sexist label type in the Y vector, higher is the chance that the image is prone to receive comments under that category. Based on this criteria, Top 3 predictions from each of the 6 regression models are chosen. These Top 3 Predictions from all 6 regression models are taken to calculate the frequency for each of the predicted labels. Finally, Majority Voting over all the regression models' predictions yields top 3 most frequently occuring labels and are then displayed as final predictions to the classifier model.

Example:
Random Forest Top 3 predictions: [Sexual Assault, Rape, Body Shaming]
Decision Tree Top 3 predictions: [Sexual Assault, Body Shaming, Other]
KNN Top 3 predictions: [Body Shaming, Threats, Rape]
Linear Regression Top 3 predictions: [Rape, Body Shaming, Sexual Assault]
MLP Regression Top 3 predictions: [Body Shaming, Sexual Assault, Other]
Gradient Boosting Top 3 predictions: [Rape, Sexual Assault, Threats]

Label Wise Frequencies: [Sexual Assault - 5, Rape - 4, Body Shaming - 4, Other - 2, Threats - 2]

Final Predictions - [Sexual Assault, Rape, Body Shaming]

Figure - 6 explains the final classifier architecture.

7 Future Work

The biggest challenges we face in this study are the inaccuracies and misclassifications that the individual feature extraction subsystems bring in while creating the dataset. The Ideal Machine Learning Solution to this is to manually annotate every Image Feature and Comment Type to get results as close to the real world truth when it comes to Sexism. This is a rather tedious and long task, but we expect much better results in this scenario. This is so because post annotating one can train a CNN which can learn all the Image Features at once, creating much better feature vectors that also take into consideration the relationships each type of Image Features share. Our

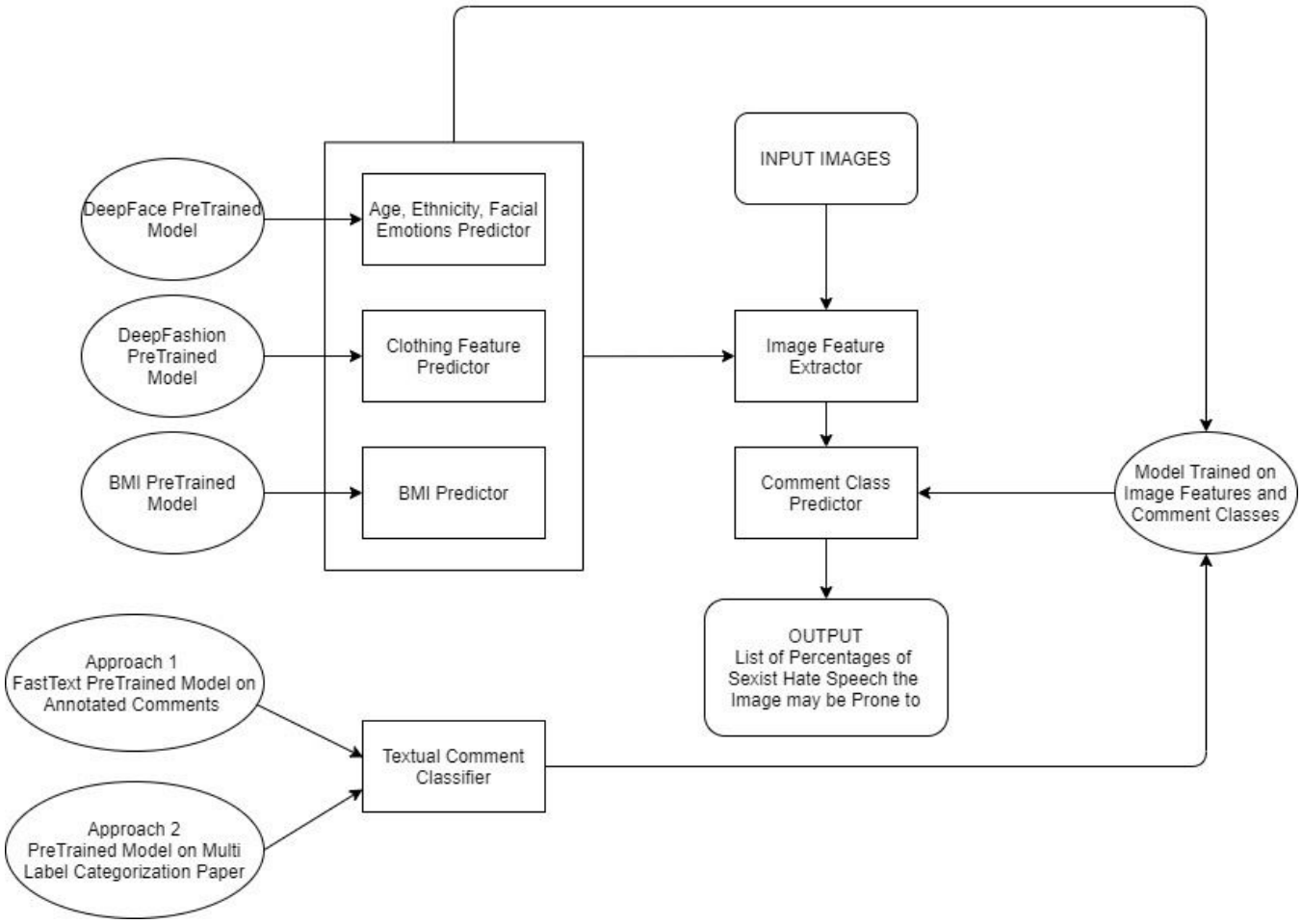

Figure 6: Final System Architecture that predicts Top 3 Sexist comment types that a given input Image may be prone to.

(**Note**: Instead of List of percentages as the output, we decided to change it to that of the top 3 labels based on these percentage values, as it is a more user friendly output.)

architecture fails here as each of the Image Features are calculated individually and that too based on different training datasets. Manually annotated comments can also take into consideration emojis, which our comment feature extraction model doesn't.

Using a purely annotated dataset will also help in creating stronger better trained relationships between the comments and the images itself. Manual annotation also maintains uniformity in distribution of features of the dataset, where our model fails to a certain extent due to its dependence on the individual feature extraction models. Clothing Features can be better annotated using manual annotation as well. That way some clothing features like collar types, sleevelengths can be explicitly part of the Image features.

Our architecture doesn't take into consideration the contents of the caption. Bringing in embeddings of the caption may further lead to better results as well when it comes to better classification. This will be useful in situations of correlating the image and caption features to the sexist comments' weight vector as if manually annotated, better attention models can be created that can possibly also predict whether or not the person in the image is the target towards sexism as it provides an added context to the comments.

During the creation of a dataset if one restricts to comments and captions belonging to a single language. This would also require semi-manual annotation to remove comments in other languages and even in other scripts, thus making the comment space uniform.

8 Conclusion

We believe that our short study sets the foundation towards studying Sexism on Instagram and establishing stronger correlations between what exactly in an image is getting targeted by a certain comment. Better Computer Vision and NLP models would definitely play an essential role in this task and could bridge the gap between the predictions and the ground truth. These concepts could play a crucial role in eliminating not just Sexism, but perhaps even Hate Speech on Social Media as a whole.

9 References

[1] Council of Europe - Combating Sexist Hate Speech
[2] E. Fersini , P. Rosso , and M. Anzovino - Overview of the Task on Automatic Misogyny Identification at IberEval 2018
[3] Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, et. al. - Multi-label Categorization of Accounts of Sexism using a Neural Framework
[4] Yaniv Taigman, Ming Yang Marc, Aurelio Ranzato, Lior Wolf - DeepFace
[5] Leo Simmons - Estimating Body Mass Index from Face Images Using Keras and Transfer Learning
[6] Furkan Kinli - [Deep Learning Lab] Episode-4: Deep Fashion
[7] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang1, Xiaoou Tang - DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annot
[8] SRK - Getting started with Text Preprocessing
[9] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov - Enriching Word Vectors with Subword Information
[10] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov - Bag of Tricks for Efficient Text Classification

10 Appendix

Github Link: https://github.com/aniruddhapdeshpande99/IG-Misogyny-Classification

Please refer to the following tables that define the numeric indices for emotion, race, clothing and sexist comment labels.

Sexist Comment Label Category	**Numeric Value (Index)**
Other	0
Attribute stereotyping	1
Hyper-sexualization (excluding body shaming)	2
Body shaming	3
Hostile work environment (excluding pay gap)	4
Sexual harassment (excluding assault)	5
Role stereotyping	6
Moral policing (excluding tone policing)	7
Sexual assault (excluding rape)	8
Menstruation-related discrimination	9
Denial or trivialization of sexist misconduct	10
Threats	11
Internalized sexism	12
Mansplaining	13
Victim blaming	14
Rape	15
Motherhood-related discrimination	16
Physical violence (excluding sexual violence)	17
Slut shaming	18
Pay gap	19
Religion-based sexism	20

Table 2: Sexist Comment Label Category Index

Clothing Type	Numeric Value (Index)
Anorak	0
Blazer	1
Blouse	2
Bomber	3
Button-Down	4
Cardigan	5
Flannel	6
Halter	7
Henley	8
Hoodie	9
Jacket	10
Jersey	11
Parco	12
Peacoat	13
Poncho	14
Sweater	15
Tank	16
Tee	17
Top	18
Turtleneck	19
Capris	20
Chinos	21
Culottes	22
Cutoffs	23
Gauchos	24
Jeans	25
Jeggings	26
Jodhpurs	27
Joggers	28
Leggings	29
Sarongs	30
Shorts	31
Skirts	32
Sweatpants	33
Sweatshorts	34
Trunks	35
Caftan	36
Coat	37
Coverup	38
Dress	39
Jumpsuit	40
Kaftan	41
Kimono	42
Onesie	43
Robe	44
Romper	45

Table 3: Clothing Categories Index

Emotion	Numeric Value (Index)
Neutral	0
Sad	1
Happy	2
Surprise	3
Angry	4
Fear	5
Disgust	6

Table 4: Emotion Class Index

Race	Numeric Value (Index)
Black	0
Latino Hispanic	1
Middle Eastern	2
Asian	3
White	4
Indian	5

Table 5: Race Class Index