Studying and Predicting Misogyny and Sexism on Instagram

Aniruddha Prashant Deshpande, Mahek Mishra, Kristen Pereira

1. Introduction

No app is more integral to teens' social lives than Instagram. According to a recent study by the Pew Research Center, 72 % of teens use the platform, which now has more than 1 billion monthly users. Still, according to a recent Pew survey, 59 % 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 allows 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.

The escalation of online violence against women since 2016, particularly due to Internet anonymity, has been a concerning trend globally [1]. This rise can be attributed to the ease of creating anonymous profiles on social media platforms like Twitter and Instagram. These platforms provide a veil of anonymity, enabling perpetrators to engage in abusive behavior without fear of accountability. The cross-national nature of online violence against women is evident, with research indicating its exacerbation during events like the COVID-19 pandemic. Factors such as increased online activity, social isolation, and heightened tensions during crises contribute to this escalation.

1.1 Background

Content moderation on social media platforms is inherently challenging due to the vast amount of user-generated content. However, integrating automated detection and prediction systems can significantly aid human moderators in efficiently identifying and addressing misogynistic and sexist comments. This collaborative approach enhances the overall effectiveness of content moderation efforts. In the fight against online harassment, this project focuses on automatically detecting and flagging such comments to reduce their prevalence. By doing so, it strives to reduce the prevalence of online harassment and create a more positive and respectful online community. Gender-based discrimination often thrives in online spaces, perpetuating harmful stereotypes and biases. This project takes proactive steps to combat misogyny and sexism on Instagram by actively identifying and addressing such language. In doing so, it contributes to the promotion of gender equality, fostering an inclusive online environment where all users feel safe and respected.

1.2 Significance

The initial phase of content moderation involves employing a multi-modal correlation approach to analyze features across images and comments. Visual features selected are determined by considering aspects of women's physical appearance that are commonly targeted in misogynistic remarks, such as race and weight. This method holds promise in mitigating misogynistic environments surrounding posts, benefiting both the user and their followers while also combating gender-based discrimination. Insights gained from this study could potentially be adapted and expanded to promote positive interactions in posts related to the poster's sexual orientation (LGBTQIA+) and gender identity, thereby fostering a more inclusive online environment.

2. Related Work

Miranda et al. [2] performed an ethnographic investigation to scrutinize misogynistic hate speech on social media platforms that build upon prior research on hate speech against women, which suggests that In recent years, hate speech against women on social platforms has increased significantly, with studies showing higher rates of harassment compared to men, particularly evident in the Italian Hate Map Project and research by UNESCO [3]. The study analyzed 74 profiles that regularly shared hate speech content and focused on content that was misogynistic by manual curation. The result of the study showed a prevalence of indirect hate speech forms, like irony and ridicule, often conveyed through images, memes, and cartoons. The analysis showed the expressions frequently target feminist women, diverging from conventional beauty standards and reinforcing conservative gender roles. Fersinin et al. [4] discuss the Automatic Misogyny Identification (AMI) task introduced at IberEval 2018, which aims to identify and categorize misogynistic content in tweets in Spanish and English. It includes two subtasks: Subtask A identifies misogyny, while Subtask B categorizes misogynistic behaviors and classifies tweet targets. The paper details data collection methods and evaluation metrics, summarising the performance of participating teams. The majority of teams used Support Vector Machines (SVM) and Ensemble of Classifiers (EoC) for both Subtask A and Subtask B. Some teams also tried out deep learning methods. The feature sets usually included n-grams and embeddings. Teams using SVM mainly relied on n-gram-based approaches, while those using deep learning methods used word embeddings. Parikh et al. [5] introduce a new dataset obtained from crawling from the EverdaySexism project website. The dataset has accounts of sexism from the website and is categorized into 23 categories. Their work addresses limitations in existing work by presenting the first multi-label classification framework for various types of sexism; they also propose a neural framework for multi-label sexism classification. In [6], Pamungkas et al. compare different traditional ML methods and deep learning-based approaches to approach the Automatic Misogyny Identification (AMI) task, aiming to classify and categorize misogynistic content and behaviors. The study highlights the efficacy of classical machine learning models, particularly ensembles, for misogyny identification on Twitter. Sing et al.[7] took a novel approach to detect misogyny by integrating visual and textual cues, recognizing the importance of memes in shaping online discourse. Leveraging multimodal models like BERT+ViT, they perform the classification and categorization of misogynistic memes using the MAMI Dataset.

Our project doesn't aim to reproduce or replicate any prior work but instead seeks to build on the successes of previous papers, drawing inspiration from similar studies in the domain. The prior work primarily focuses on detecting hate speech in textual comments on social media, neglecting the analysis of image content. Furthermore, no work aims to predict what level of sexist/misogynistic comments a post can get based on the content. Our project aims to fill this gap by extending the analysis to image data and identifying features within images that might provoke hate comments. Our analysis might prove helpful to content moderators in prioritizing posts that might be susceptible to such comments, thus allowing them to spend their resources judiciously and modify their algorithms based on the correlations observed.

2. Objectives, Goals and Outcome

The objective of our project, which focused on Toxicity Detection on Instagram, was to examine and facilitate the identification of cyberbullying incidents within online photo-sharing platforms. We aimed to detect and forecast instances of Misogyny and Sexism on Instagram by correlating the features of individuals in a given Instagram post with types of Sexist/Misogynistic comments on that post. We planned to achieve this by training a neural network to analyze image features and predict the kind of sexist comments the post may attract, correlating features such as BMI, Age, Race, Emotion, and Clothing with various types of Sexist Classes like Body Shaming, Slut Shaming, and Sexual Harassment. We successfully accomplished the major objectives of our project by implementing Image Feature Extraction, Comment Label Classification, and Correlation between the two. We utilized Deepface for demographic analysis and a pre-trained ResNet model to calculate BMI. FastText classifiers trained on sexism categories were used for comment classification. The classifiers were tested using various models, including Random Forest, Decision Tree, KNN, Linear Regression, MLP Regression, and Gradient Boosting, achieving satisfactory performance with MAE ranging from 0.121 to 0.123. Due to time constraints, however, we were unable to explore multimodal models combining image and text data, investigate a wider range of features, or analyze comments on a per-user basis. Nevertheless, our work provides valuable insights into the relationship between demographic features and the likelihood of encountering sexism on Instagram, aligning with the expected outcomes. Future research could focus on integrating textual information, exploring additional features, and conducting a more granular analysis of user behavior to further enhance the effectiveness of our toxicity detection model on Instagram.

3. Data

The first task was to crawl Instagram and create 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 of Misogynistic Hate speech. Several papers were also looked into in order to categorize the type of Sexism/Misogyny which can be observed in the comments. Out of these, two papers [4] and [5] stood out the most and involved annotation schemas that made use of in-depth categories to classify Sexism. [4] also confirmed the fact that Public feminist influencers were heavily prone to Misogynistic Hate Speech.

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 criteria 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 that lets users upload multiple photos per post, and this feature was taken into consideration while downloading the images. Therefore, the set of comments on that post could be applied to every single image present in that given post.

Fig 1. - Examples of Categories Influencers Chosen for our Dataset: Body Positive Influencers - @plumptopretty, Influencers who preach Women's Rights/Sex Positivity - @Sex Positivity, Actresses and Models - @sydney_sweeney, Political Figures - @aoc, Fashion Bloggers - @komalpandeyofficial, Miscellaneous Influencers - @trintrin

Based on the above claim from [8], we curated a list of 50 Instagram influencers whom we thought were prone to such hate speech. A lot of them also tend to be the type of influencers who share their feministic thoughts frequently and, in fact, have become famous as influencers for being open when it comes to discussing issues involving feminism. Figure 1 summarises the categories of the influencers we thought would be prone to Sexist Hate speech:

Figures who preach Body Positivity: These types of influencers 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 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: **@plumptopretty**

Influencers who discuss Women's Rights / Sex Positivity: These influencers discuss various aspects of feminism, calling out misogynistic ideals of society. Some of them also discuss Sex positivity, which involves topics supporting LGBTQIA+, 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 stereotypes. Example: **@salonichopraofficial**, an actress turned Instagram influencer who heavily preaches women's rights equality involving topics like "Free the Nipple Movement."

Actresses and Models: Actresses and Models have been known to be victims of criticism and hate speech towards maintaining their bodies and, hence, are prone to Body Shaming. Since accessing and modeling go 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 over time. Examples: **@sydneysweeney, @malaikaaroraofficial**.

Political Figures: This category of influencers involves female political figures. We decided to add these influencers as they may be prone to misogynistic comments focused around Mansplaining, Hostile Work Environment, age-oriented sexist comments, Threats (especially due to differences in political beliefs), etc. Examples: **@aoc, @kamalaharris**

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

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**

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

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

4. Approach

This section is divided into three subsections, namely Image Feature Extraction, Comment Label Classification, and the Correlation of the above two. We combined the first two subsections to obtain the correlation between Image Features and the Comments Labels.

4.1 Image Feature Extraction

We plan to make use of the following image features of the person(s) in a post: **Age, Ethnicity, Facial Emotion, Clothing Features, and Body Mass Index (BMI).** Our decision to choose these features stems from the real-world intuition of what aspects of a person are usually targeted by a certain Sexist comment. Racist discrimination is linked to ethnicity, sexism in clothing, body shaming in weight, and age to societal roles. These correlations are further elaborated in Section 5.

We utilized DeepFace [9], a Python framework, to analyze demographic features like age, emotion, and race in images. Developed by Facebook Research, it's trained on 4 million facial images and can predict gender (though our dataset only includes feminine genders). There were situations wherein the face of the person in an image was 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.

An open-source pretrained model based on the medium article was used to calculate the BMI. It was trained on a range of 4000 images, each of a different individual, taken from the front of the subject. The model was created using the Keras ResNet50 class. 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.

Clothing feature extraction required a lot of training and effort, as no Pre-trained model was available. Thus, in order to extract the clothing features of the person in a post, we trained a model on a subset of the Deep Fashion Dataset [10] and trained a classifier over it. The Dataset used here had 290000 images and 46 classes, and we used Transfer Learning with Resnet50. We used the top 2 clothing label predictions for the final vector creation.

4.2 Comment Label Classification

For the comment label classification, we had to come up with certain classes of Sexism/Misogyny. We then used FastText to train a classifier on this dataset to classify all comments of each user post. FastText is a library for learning word embeddings and text classification created by Facebook's AI Research (FAIR) lab. The model allows for the creation of unsupervised/supervised learning algorithms for obtaining vector representations of words for any custom dataset, and the same can be understood from [11] and [12]. Facebook makes pre-trained models available in 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 and Classification

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]

The final output vector is illustrated as follows:

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.

y[misogyny_class] = no_of_comments(misogyny_class) / total_no_sexist_comments

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 final classifier in six regression models: random forest, decision tree, KNN, linear regression, MLP regression, and gradient boosting. This classifier then predicts the type of Sexist comments a given image is prone to. The final architecture of the classifier model is illustrated in Figure 3.

5. Results

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 the 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 expected behavior, as no sexist comment is usually based on the emotion of the person in the photo.

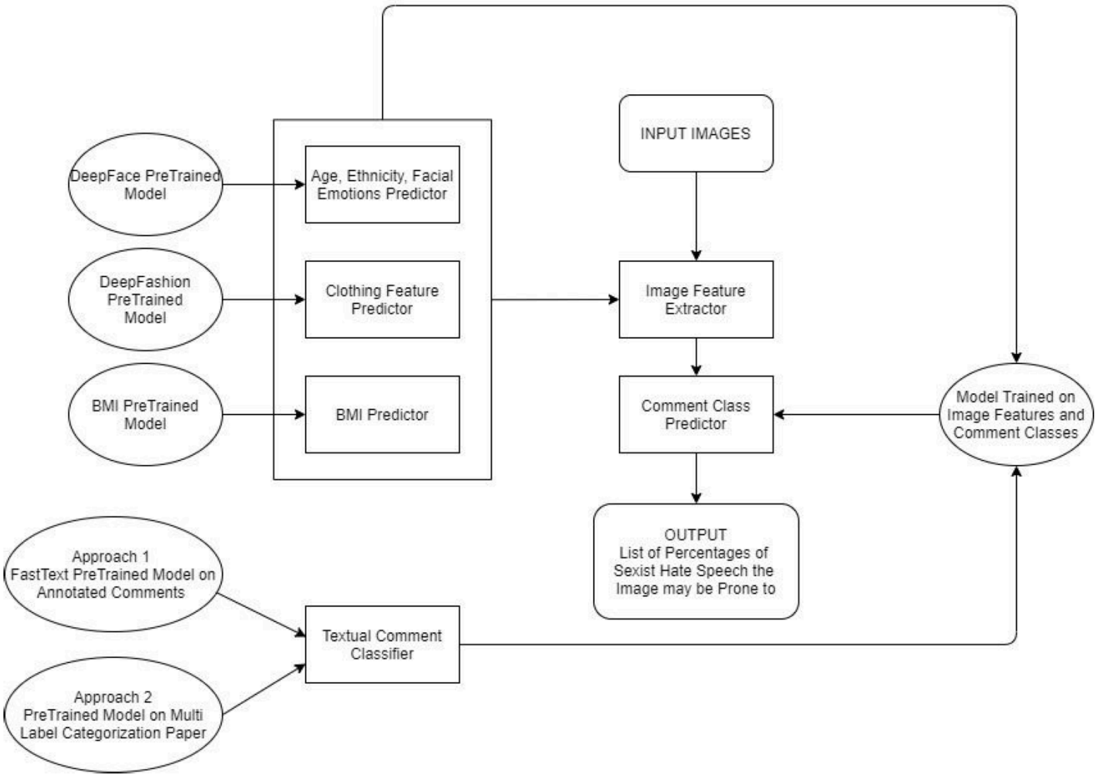

INPUT IMAGES
DeepFace PreTrained Model
Age, Ethnicity, Facial Emotions Predictor
DeepFashion PreTrained Model
Clothing Feature Predictor
Image Feature Extractor
BMI PreTrained Model
BMI Predictor
Comment Class Predictor
Model Trained on Image Features and Comment Classes
OUTPUT List of Percentages of Sexist Hate Speech the Image may be Prone to
Approach 1 FastText PreTrained Model on Annotated Comments
Textual Comment Classifier
Approach 2 PreTrained Model on Multi Label Categorization Paper

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

We can see from Figure 4 that most of the people were classified in the age group 30 -35, which was expected as the maximum influencers that were chosen belonged to this group. The Race prediction model showed bias where 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. The clothing classifier classified a full-size image fairly well into a category close enough to the original outfit. However, since the model was trained on a dataset that had full-length images, it should be noted that it misclassifies when the images are not full-length. As a result, we can see a lot of classifications in the jumpsuit category.

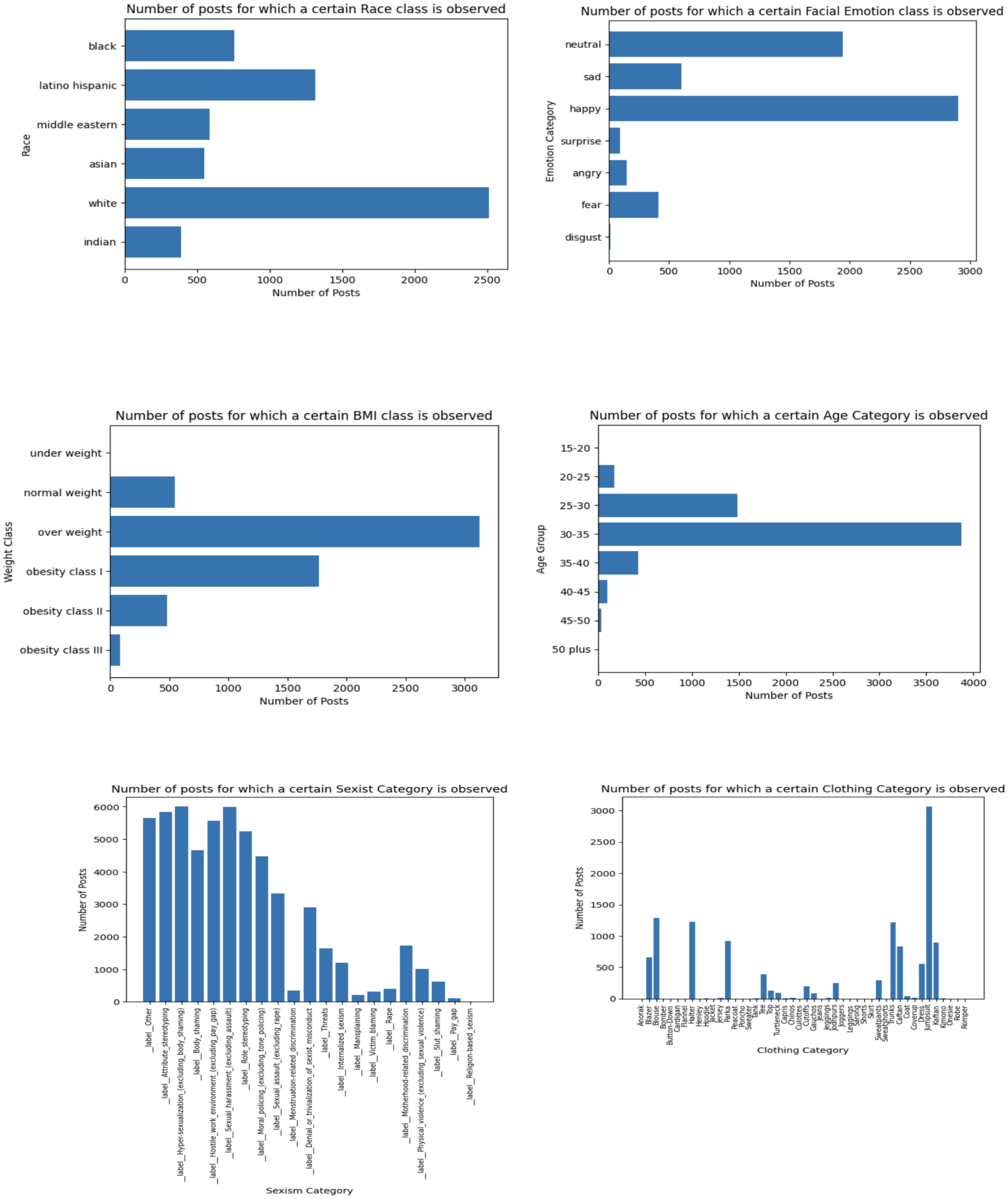
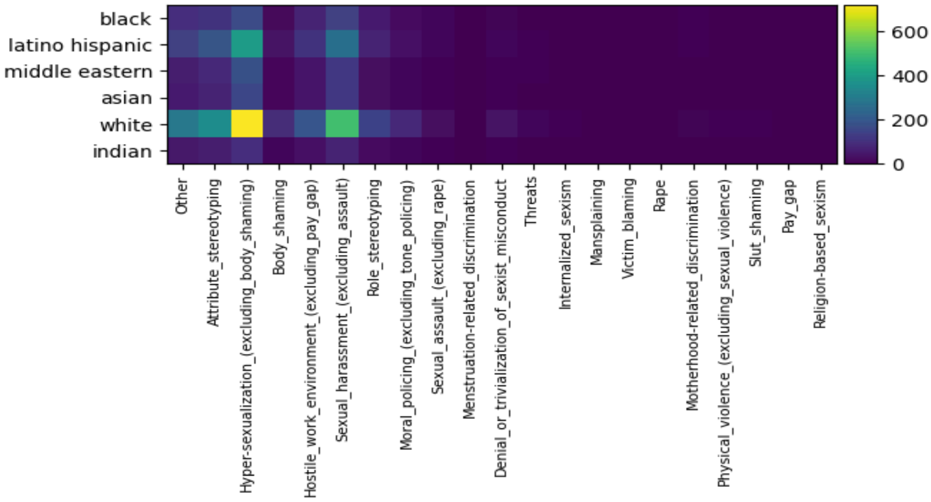


Figure 4 - Statistical Information of Visual Features on Posts based on Race, Emotion, BMI, Age, Sexism Category, Clothing Item

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 are omitted. When it comes to racial-based correlation studies, positive correlations are observed for Hypersexualisation and Sexual harassment types of comments for White and Latino Hispanic Races, but not as much for other races. We observed a higher positive correlation between Hostile Work Environment and Motherhood Related Discrimination Comment Classes with images in the age groups 30-35. This is corroborated by the statistics of average motherhood ages and working-class women’s age (with the intuition that women in the 20-25 age group are still in the education stages of their lives). This observation aligned with real-world intuition. An excessive positive correlation was observed between the Body-Shaming comment category and the Overweight class, confirming our intuition. The misclassification of clothes in the Jumpsuit and other categories was most observed in the Clothing Feature-based classification. This led to a heavy zero correlation between some clothing categories as they were heavily underclassified 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 wearing clothing perceived as 'revealing' according to societal norms." A similar observation was also made in our results. 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 like a Blouse, a pair of Trunks, or a Halter rather than other heavy outfits like a Blazer etc.



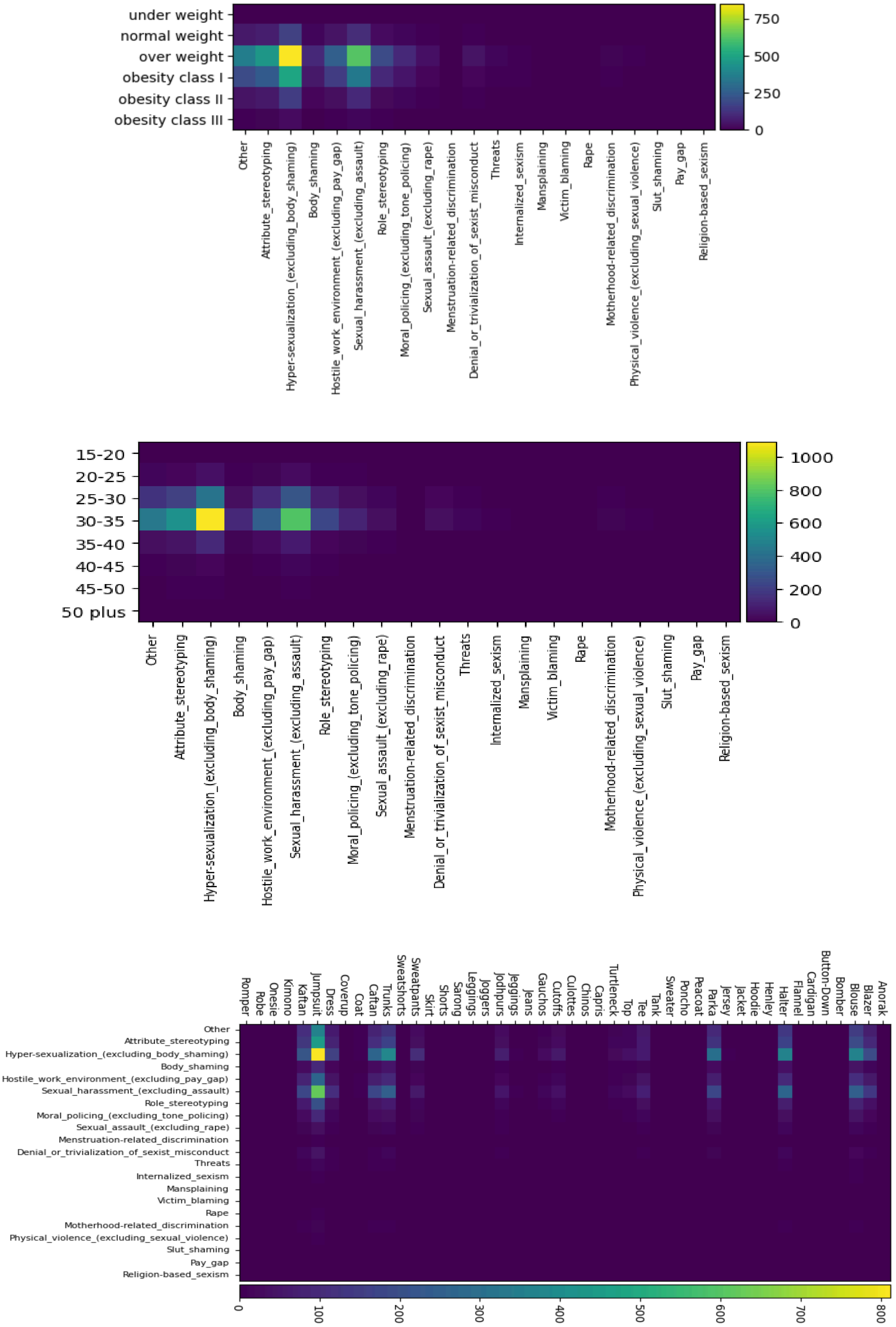


Figure 6 - Correlation Heatmaps for Sexist Comment Categories vs. Race, BMI, Age, & Clothing respectively

For the final classifier, the following models were Random Forest, Decision Tree, KNN, Linear Regression, MLP Regression, and Gradient Boosting. We chose the MAE metric for evaluation because our model predicts probability, and MAE is a suitable metric for such regression models.

Among these models, Gradient Boosting performed the best with the lowest MAE, followed closely by Random Forest and Decision Tree. KNN had the highest error, indicating less effective performance in this context.

Model	**MAE**
Random Forest	0.121941
Decision Tree	0.121839
KNN	0.123660
Linear Regression	0.122276
MLP Regression	0.123053
Gradient Boosting	0.121259

Table 1 - Mean Average Error for different classifier models

6. Discussion of Outcomes, Implications, and Conclusion

From this study, we understood how important and engaging the data collection and processing task is and how crucial it is to ensure data quality. Furthermore, we also observed how biases in the dataset seep into model predictions, as seen in the race classifier and age classifier. Our findings, as discussed above, mostly confirmed what most previous research suggested and what our intuition was saying. For example, a lot of overweight people were receiving hate comments regarding body shaming, whereas women in the age group 30-35 were getting Hostile Work Environment-related comments. On manual observation of some of the comments on posts, we noticed that most sexist comments were from accounts that were either meme pages or anonymous throwaway accounts with no posts. Very few people used their personal accounts to comment on such things. This might be related to the trolling effect we had talked about in class and something we can further research. Finally, our study shows noticeable correlations between visual features and comments. We can definitely use this multimodal information to create better moderation systems and understand people's reactions to certain kinds of posts. A new approach to understanding people's reactions by looking at image features was explored through this study. We believe that our short study sets the foundation for 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. This project was limited in time and resource constraints but still led us to find some correlations. We believe if more data could be scraped in the future, more features could be tested out to look for correlations and improve predictions. This could help algorithm curators on social media platforms execute better content moderation. Our current model architecture doesn't consider caption content. Bringing in embeddings of the caption may further lead to better results as well when it comes to better classification. Another possible improvement that could be incorporated in the future would be a translation of all comments into a single language and handling nonword tokens like emojis, as emojis often help relay the tone of a comment and would further help classify the comments correctly. There is much research going on in vision-language models that deal with multimodal embedding space to represent features. It would be interesting to study how these models can be leveraged to generate better features to predict sexist comments.

References

[1] Poland, B. (2016). Haters: Harassment, abuse, and violence online. U of Nebraska Press.

[2] Miranda, S. (2023). Analyzing Hate Speech Against Women on Instagram. Open Information Science, 7(1), 20220161. https://doi.org/10.1515/opis-2022-0161

[3] Posetti, J., Aboulez, N., Bontcheva, K., Harrison, J., & Waisbord, S. (2020). Violencia en línea contra las mujeres periodistas.

[4] Fersini, E., Rosso, P., & Anzovino, M. E. (2018). Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In Proceedings of IberEval@SEPLN. Retrieved from https://api.semanticscholar.org/CorpusID:51942244

[5] Parikh, P., et al. (2019). Multi-label Categorization of Accounts of Sexism using a Neural Framework. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 1642-1652). Hong Kong, China. doi: 10.18653/v1/D19-1174

[6] Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. Information Processing & Management, 57, 102360. Retrieved from https://api.semanticscholar.org/CorpusID:224976064

[7] Singh, S., Haridasan, A., & Mooney, R. (2023). Female Astronaut: Because sandwiches won't make themselves up there": Towards Multimodal misogyny detection in memes. In The 7th Workshop on Online Abuse and Harms (WOAH) (pp. 150-159). Toronto, Canada. doi: 10.18653/v1/2023.woah-1.15

[8] Council of Europe. (2016). Gender Equality Strategy: Combating Sexist Hate Speech. Retrieved from https://rm.coe.int/1680651592

[9] Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1701-1708). Columbus, OH, USA. doi: 10.1109/CVPR.2014.220

[10] Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1096-1104). Las Vegas, NV, USA. doi: 10.1109/CVPR.2016.124

[11] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

[12] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.