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. Link to implementation: `https://github.com/AlokDebnath/Sexist_Stereotype_Classification`
Link to website: `https://alokdebnath.github.io/Sexist_Stereotype_Classification/`
Link to video: `https://www.youtube.com/watch?v=okd5UwopDJE`

1 Introduction

"Guys are just stronger than girls", "Girls drive badly, make me a sandwich", these comments are considered offensive, harmful and toxic 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 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.

Given the subjectivity of the data being collected, contemporary methods use two-fold data creation mechanism on social media which allows for filtering of the relevant content, such as mentions of sexism, sexist stereotypes, abuse and so on. After this, synthetic data can be constructed and validated based on the data collection criteria [2]. Here, by synthetic data creation, we also include scraping and adding more content associated either from different platforms, or collecting more data and removing the negative samples to boost the number of positive samples in the final dataset. Most studies focus on the idea of ambivalent sexism which pertains to the sexism in everyday discourse [10]. These can be considered multiple criteria of accounting for sexist notions 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.

In this project, we perform sexist stereotype classification for instagram captions and comments. We identify all those captions and comments which pertain to sexism, sexist stereotypes, mansplaining and stories about abuse, which include those captions and comments which are about sexism as well

as sexist themselves. We use the data to develop a multi-layer classifier, based on which we classify the sexist comments and captions.

2 Related Work

In this section we consider and mention the related work on stereotype, abuse and other discrimination on social media.

[8] considers the role of gender in media and its use in various fields of technology. not pertaining to social media directly. but rather to technology and society.

[9] is a more relevant study, which assessed the effects of a sexist hashtag on the social media site Twitter. Twitter account that was anonymous or had personally identifying details. They were asked to share (i.e., retweet) or write posts incorporating a sexist hashtag. After exposure, participants completed two purportedly unrelated tasks, a survey and a job hiring simulation in which they evaluated male and female candidates' resumés. Higher interactivity led to more negative evaluations of female job candidates.

[10] is a relevant study on the use of compliments in the work place or on social media and when a comment may be categorized as sexist. The paper identifies benelovent sexism, which is a sexist comment disguised as a critique, compliment or passing off the behaviour as palatable. The paper captures the essence of the boundary by using a classification mechanism which is then used to define when a comment is be deemed offensive. This is quite a useful step in attribute and role stereotype detection as well.

[11] provides the first classification mechanism for sexism and abuse based speech which considers that the classes in which the data is distributed need not be mutually exclusive. While this was a milestone movement in classifying sexist comments, it must be known that previous work on this topic was actually based on mutually exclusive classes not for lack of recognition that classes were not overlapped, but because it makes the construction of classifiers much simpler, which is a useful characteristic for developments in this field. Nevertheless, the advent of multiclass classification with overlap is a useful discovery.

Finally, [12] moves away from the contemporary literature on sexism classification, which is different from sexism detection, because it has certain limitations in terms of the categories of sexism used and/or whether they can cooccur. To the best of our knowledge, theirs is the first work on the multi-label classification of sexism of any kind(s), and they contribute the largest dataset for sexism categorization. The work is an inspiration to people in this field as it provides not only a large dataset but also a thorough analysis of multiple classes of sexism, which can be used to automate the development of larger datasets and improve upon the classification and detection of future use in social media analysis and research.

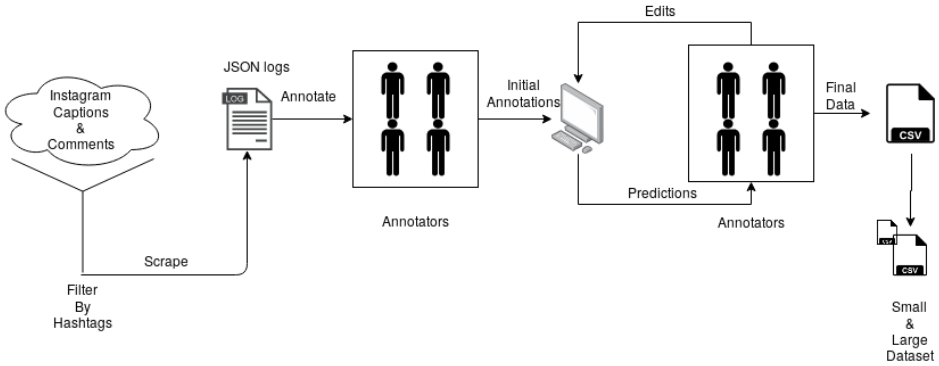

Figure 1: Data collection and annotation pipeline. We see here that the annotation and editing was an iterative process.

3 Dataset Creation

In this section, we explain the data collection mechanism for Instagram captions and comments. We use a semi-automated annotation mechanism using active learning models for quick and efficient annotation of large amounts of data. Figure 1 shows the data collection and annotation pipeline, which is described here.

3.1 Collection and Preprocessing

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.

3.2 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.

Using the annotations, we then further classified the sexist instagram captions and comments based on role based and attribute based sexism. Role

based sexism, also known as role stereotyping, refers to the generalization of false notions based on the idea that certain roles, occupations, professions and jobs are restricted only to, or suitable only for, a particular gender instead of another. On the other hand, attribute based sexism or attribute stereotyping, refers to misconceptions about the physiological, psychological or behavioural characteristics of people based on their gender.

It is also possible that the sexism may be a combination of both. However, due to the data skew on the number of sexist captions and comments, a combination of both would be difficult to identify and isolate, due to which it was not considered.

3.3 Automated Data Annotation

We use a small pool of labeled data, approximately 200 captions and comments, which are manually annotated. On this, we apply an active learning mechanism, using margin based measures of SVMs measures based on class probability for classification. We first compute the **1-Entropy** as:

$$M_{i,\text{1-Entropy}} = 1 - H(\hat{p}(\cdot|x_i)) \tag{1}$$

$$= 1 + \sum_j \hat{p}(c_j|x_i) \log \hat{p}(c_j|x_i) \tag{2}$$

where $\hat{p}(c_j|x_i)$ is the current estimate of the probability of class c_j given the example x_i. We use 1-Entropy instead of entropy, so all three measures will have lower values for less certain instances. 1-Entropy favors examples where the classifier assigns similar probabilities to all classes.

Using the 1-Entropy value, we determine the entropy of a caption or comment belonging to a particular class. In the event there are two most likely classes, we define the concept of a **margin** as follows.

$$M_{I,\text{Margin}} = |\hat{p}(c|x_i) - \hat{p}(c'|x_i)| \tag{3}$$

if the classes are c and c' [5]. Margin picks examples where the distinction be-tween two likely classes is hard. Finally, we calculate **MinMax** as a measure of the classifier's decision on a particular example. The other two measures also take into account the classifier's assessment of classes that were not chosen for the unlabeled example.

Performance estimation of this model is done by using the F measure, which is based on the number of true positives (TP), false positives (FP) and false negatives (FN). Based on the basic formula of F measure:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{4}$$

$$\implies F1 = \frac{2PR}{P + R} \tag{5}$$

$$\implies F1 = \frac{2TP}{2TP + FP + FN} \tag{6}$$

Therefore, we calculate these values as estimates for each iteration as:

$$\hat{TP} = \sum_i \sum_j \hat{p}(c_j|x_i) d_{i.j} \tag{7}$$

$$\hat{FP} = \sum_i \sum_j (1 - \hat{p}(c_j|x_i)) d_{i.j} \tag{8}$$

$$\hat{FN} = \sum_i \sum_j \hat{p}(c_j|x_i)(1 - d_{i.j}) \tag{9}$$

where $d_{i,j}$ is a flag such that $d_{i,j} = 1$ if $j = \arg\max_j \hat{p}(c_j|x_i)$ else $d_{i,j} = 0$ [6].

The active learning procedure gave the highest $F1$ score of approximately 61% which can be attributed to the high skew in the dataset, primarily due to Instagram's strict community guidelines which reduces the number of sexist or toxic comments and captions. More importantly, a large amount of the sexist and sexism data in the dataset consists of stories of sexism, which are usually much longer than the usual caption or comment, skewing the data considerably.

4 Classification Methodology

In this section, we describe the classifier design based on which the classification of the captions and comments is done. First, we have a binary classification for whether the given comment is sexist or not sexist. Then we show the design of multi-class classifier for attribute stereotyping, role stereotyping and non-stereotyped data. An overview of the model is shown in figure 2. We use GloVe embeddings [7] of 300 dimensions for both the classification tasks.

4.1 Binary Classification

For binary classification, we used a combination of recurrent networks followed by a softmax classifier. The classifier design was not considered more intricate due to the constraints on the data as mentioned above.

A recurrent network can be either an RNN (recurrent neural network) or an LSTM (long short term memory). Recurrent models are used in this project as they capture features of the previous cell as well as the current input, weighted on a non-linearity, usually a tanh function. Here we use simple recurrent model of size 100 dimensions. The difference in performance for RNNs and LSTMs comes from the fact that LSTMs have three gates which determine what information should be retained from the previous hidden states and what information should be discarded. LSTMs are preferred over RNNs in order to solve the vanishing gradient problem.

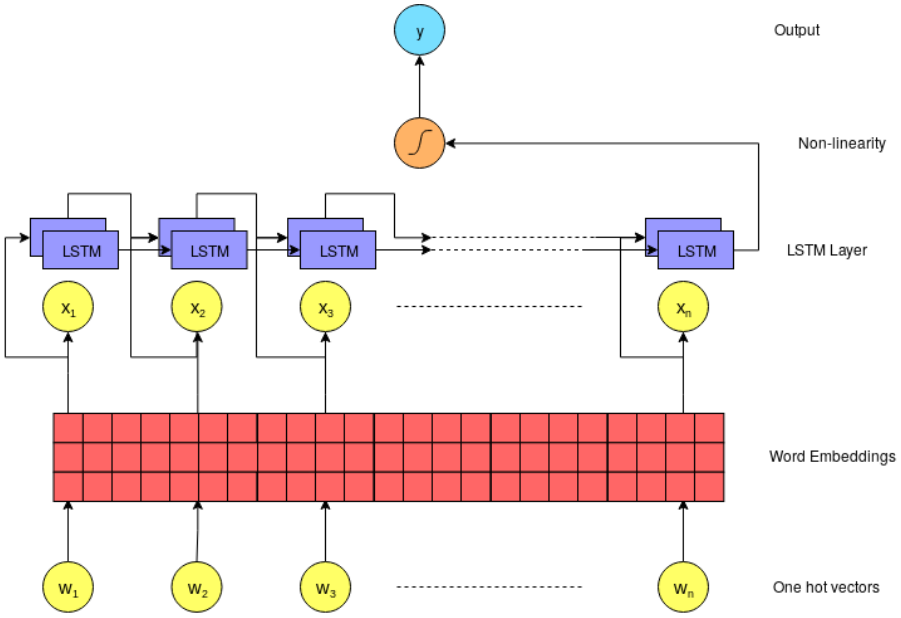

Figure 2: Model of the multiclass classifier. It is a stacked LSTM sequence followed by a non linearity

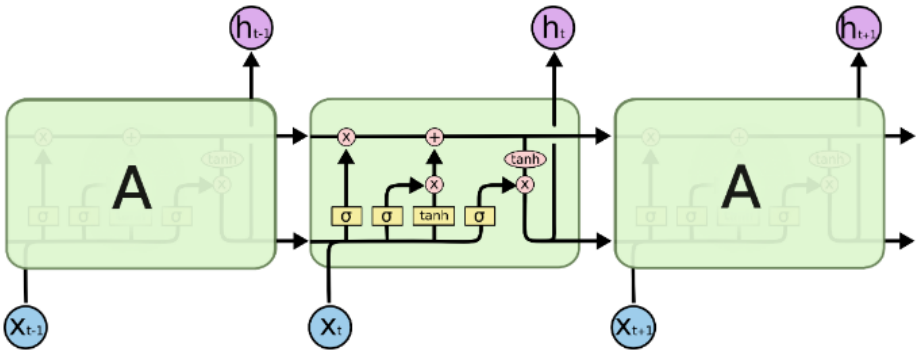

Figure 3: The LSTM internal representation

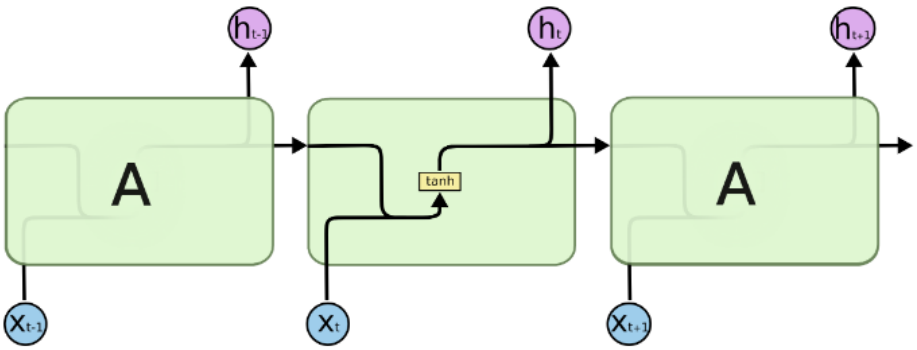

Figure 4: The RNN internal representation

Figures 3 and 4 show the internal representation of a single RNN cell and single LSTM cell.

A combined set of equations for recurrent networks are provided below. Here, f_t, i_t and o_t are the equations of the forget, inner and outer gates. The values are based on the sum of the individual weights, the hidden state value and the individual bias terms. The values are then concatenated, based on a concatenation function which identifies the weights of the previous cell and the gates mentioned above.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{10}$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{11}$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{12}$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{13}$$
$$h_t = o_t \cdot \sigma_h(c_t) \tag{14}$$

We run four experiments, a single LSTM of 50 and 100 dimensions, an two layer RNN of 100 dimensions and a two layer LSTM of 100 dimensions. The results of the experiments are given in the section below.

4.2 Multiclass Classification Model

For multi-class classification, we use a slightly more complicated model of a stacked LSTM. A stacked LSTM has multiple sequences of LSTMs in a stack, such that for the second layer onwards, the input is not the embedding, but the hidden state of the previous layer. The diagram 2 shows this configuration of stacked LSTMs.

We ran three experiments in multiclass classification between attribute stereotype, role stereotype and non-stereotype captions and comments, one with a single LSTM layer of 50 dimensions, one with LSTM layer of 100 dimensions and a stacked LSTM, each layer of 100 dimensions.

Model	Recall	Precision	F1 -Score	Accuracy
Single Layer LSTM (50 dimensions)	0.59	0.43	0.49	0.70
Single Layer LSTM (100 dimensions)	0.50	0.39	0.44	0.62
Two Layer LSTM (100 dimensions)	0.45	0.36	0.41	0.58
Single Layer RNN (100 dimensions)	0.43	0.37	0.39	0.57

Table 1: Experiment 1: Binary Classification Results

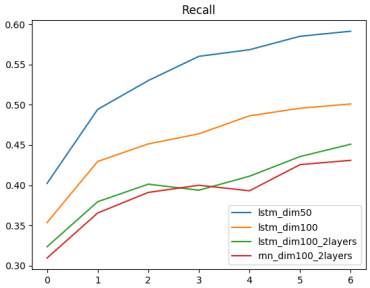

Figure 5: A graph showing recall values for binary classification

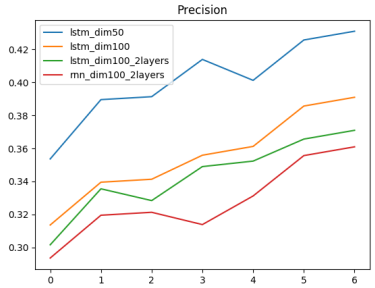

Figure 6: A graph showing precision values for binary classification

5 Results and Analysis

We can see in Table 1 the precision, recall and F1 values of the binary classification experiment are provided. We see two important observations here. First, we see that the lower dimension single layer LSTM performs the best despite being the simplest model. This is for two main reasons, which are as follows:

1. The number of data points on training are quite few, and the ratio of positive to negative samples are quite skewed. This causes larger models to overfit, and because of that the larger the model in dimension size, the worse it performs.

2. The data is skewed in more than one way. The comments which are sexist stereotypes tend to be much longer than those which are not sexist, specifically because the instagram scraping methodology only allows for scraping based on hashtags.

We show the graphs of precision, recall, accuracy and F1-score of the binary classification experiment in figures 10, 9 11 and 12. The effect of data overfitting is seen almost immediately. Further, note that sparsity and skew in the dataset requires better training data. Higher accuracies may be achieved by working with the better data.

We show the graphs of precision, recall, accuracy and F1-score of the binary classification experiment in figures 10, 9 11 and 12. The effect of

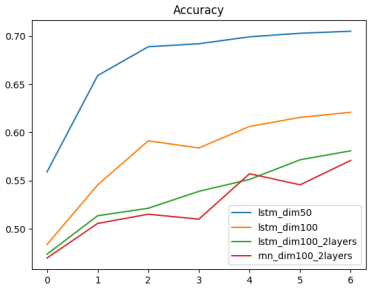

Figure 7: A graph showing accuracy values for binary classification

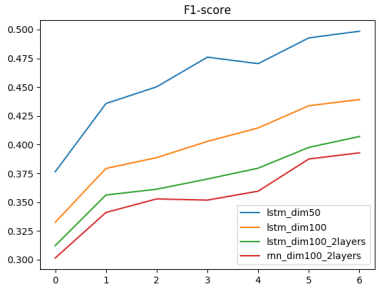

Figure 8: A graph showing F1 values for binary classification

Model	Recall	Precision	F1 -Score	Accuracy
Single Layer LSTM (50 dimensions)	0.55	0.49	0.518	0.605
Single Layer LSTM (100 dimensions)	0.51	0.44	0.472	0.535
Two Layer LSTM (100 dimensions)	0.484	0.443	0.462	0.513

Table 2: Experiment 2: Multi-class Classification Results

data overfitting is seen almost immediately. Further, note that sparsity and skew in the dataset requires better training data. Higher accuracy may be achieved by working with the better data.

Table 2 shows the results of the multiclass classification experiment. Again here we see that the simplest model performs the best. We also see that using a stacked LSTM shows a slight increase in performance, but the model runs the risk of overfitting.

We also show the loss values of each of the models, binary and multi-class classification. We see that while the loss falls most quickly for the model that stabilizes quickest, based on which the local minima is achieved. While the local minima is not the best performing, the model learns certain charcteristics of the data, such as the use of certain terms, length of caption or comment and so on.

6 Conclusion

In this project, we performed a study into the classification of Instagram captions and comments. We first annotated the data using a set of well formed guidelines. The deprecating API provided by Instagram inhibits the process the scraping the data off the site and they delete any comments or posts that are reported within a short period. This lead to a small number of sexist posts in our dataset to start with.

With this dataset, we started off with a manual annotation of small

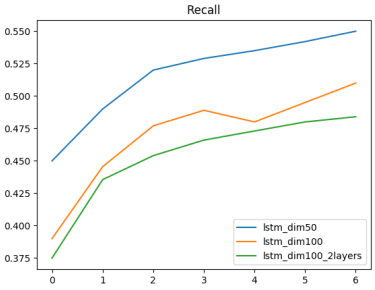

Figure 9: A graph showing recall values for binary classification

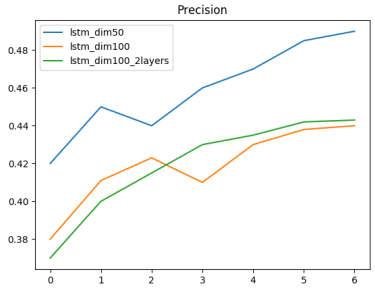

Figure 10: A graph showing precision values for binary classification

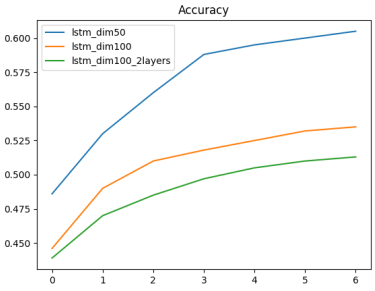

Figure 11: A graph showing accuracy values for binary classification

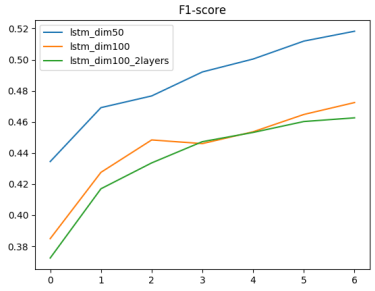

Figure 12: A graph showing F1 values for binary classification

number of posts, and using this seed data, we then used an active learning classifier in order to classify a large number of captions and comments. We cross verified the tags to see if the tags were right.

We then experimented with different classifiers, where an LSTM classifier with only 50 hidden layer dimension performed the best compared to higher dimension, or multi-layer classifier, even though we made sure that the training set had equal distribution between the 2 classes, for both binary and hierarchical classifier. This can be attributed to the skewed dataset that we have for this task. Further work over here would be to expand the dataset to include more sexist posts/captions.

The further work in this would be first to better the dataset by including more sexist captions. We can also identify more classes in the sexist types in the dataset, such as slut shaming, mansplaining, etc. The next step would be to experiment with other classifiers like Bi-LSTM, CNN, CNN-biLSTM-Attention, Hierarchical-biLSTM-Attention, and BERT, and with GloVe Twitter embedding along with GloVe Wikipedia.

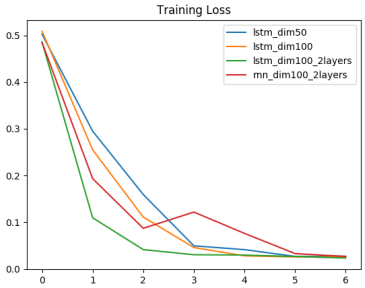

Figure 13: A graph showing decrease in training loss for binary classification

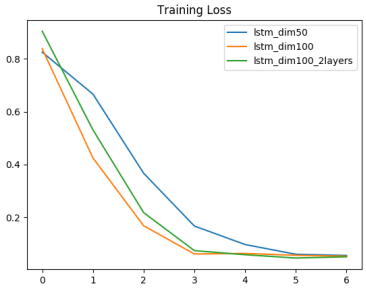

Figure 14: A graph showing decrease in training loss for multiclass classification

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

[5] Laws, F., Schätze, H. (2008, August). Stopping criteria for active learning of named entity recognition. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 465-472). Association for Computational Linguistics.

[6] Lewis, D. D. (1995, July). Evaluating and optimizing autonomous text classification systems. In SIGIR (Vol. 95, pp. 246-254).

[7] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 con-

ference on empirical methods in natural language processing (EMNLP). 2014.

[8] Carstensen, Tanja. "Gender and social media: sexism, empowerment, or the irrelevance of gender?." The Routledge companion to media & gender. Routledge, 2013. 501-510.

[9] Fox, Jesse, Carlos Cruz, and Ji Young Lee. "Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media." Computers in Human Behavior 52 (2015): 436-442.

[10] Jha, Akshita, and Radhika Mamidi. "When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data." Proceedings of the second workshop on NLP and computational social science. 2017.

[11] Karlekar, Sweta, and Mohit Bansal. "SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories." arXiv preprint arXiv:1809.04739 (2018).

[12] Parikh, Pulkit, et al. "Multi-label Categorization of Accounts of Sexism using a Neural Framework." arXiv preprint arXiv:1910.04602 (2019).