DATA MINING AND DATA WAREHOUSING LAB

SEXISM DETECTION IN POPULAR ENGLISH LANGUAGE TV-SHOWS

TEAM NUMBER: 5

TEAM MEMBERS:
Vidushi Rai - 200911258
Shreeyanka Das- 200911270

ABSTRACT:

This paper presents a methodology to classify sexist content in TV shows using sentiment analysis of their scripts. The dataset used was trained on a benchmark dataset and includes dialogues annotated as sexist or neutral. The aim is to quantify the amount of sexism in popular TV shows and assess their impact on viewers. The study uses supervised learning algorithms (SVM, Decision Tree, AdaBoost, Naive Bayes, and LR) to detect and quantify sexism in TV shows by labeling the dialogs of a TV show as sexist or neutral. This approach can help raise awareness about negative stereotypes and tropes perpetuated in TV shows.

INTRODUCTION :

TV shows make up a significant portion of the cultural fabric of a time period. The most popular TV shows of a time often reflect the kind of tropes and stereotypes that are popular during the time period specific to the run of a particular TV show. Many of these popular TV shows spanning over decades subtly and slowly influence the psyche of the viewers based on the kind of content and themes portrayed in them. TV shows, especially sitcoms often are based around themes that aren't explicitly sexist, however, the underlying mixture of negative tropes, stereotypes, and portrayal of certain characters under a misogynistic lens often leads to the messaging to the viewership perpetuating negative stereotypes like misogyny over time.

The proliferation of streaming services and other English language media all over the world have made these English language TV shows an important part of sentiment analysis to identify sexist and misogynist troupes. A lot of the viewers of such TV shows tend to be adolescents and young adults who on exposure to such kind of media might pick up on the subtle negative messaging. Many popular TV shows such as *The Office*, *How I Met Your Mother*, and *Two and a Half Men* have influenced many young adults to emulate the characteristics of their characters. Many of these popular characters have many sexist troupes that are a running gag throughout the series. The aim of this project is to classify the sexist content in such TV shows using the TV scripts and the dialogues to quantify the amount of misogyny in a certain TV show given an example script.

Sentiment analysis of the scripts of TV shows helps in identifying these negative troupes and making the viewers aware of them. The impact of quantifying sexism in TV shows

using their scripts and classifying the dialogues into having sexist or neutral connotations also helps in assessing the amount of sexism in various TV shows over a period of time.

The dataset [27] used here has been trained on the benchmark dataset [3] which takes the script of two sets of popular TV shows. One set of TV shows is older TV shows that were popular around two decades back and the other set includes a much newer set of shows where the dialogues have been annotated as sexist or neutral. The original dataset contains the dialogue and the labeled connotation and the character who spoke them in that TV show. Since the aim is to quantify the overall sexism in the TV shows the character labels are removed from the dataset. This dataset tries to encompass various forms of sexism both explicit and implicit. Implicit sexism is much harder to detect and hence there is a benefit from the fact that the dataset is already trained on the benchmarked dataset. Given the script of any TV Show this methodology is able to detect and quantify the amount of sexism in a given TV show using a supervised learning approach. The five algorithms for this task are SVM, Decision Tree, LSTM, Naive Bayes, and LR.

LITERATURE REVIEW:

[1] This paper uses a semi supervised learning approach to detect sexism in a dataset trained against a benchmark dataset trained on a dataset of tweets that identified sexism and various other degrees of hate speech. This paper discovered that Bi-LSTM produces the best results with an improvement of 4.67% over the other algorithms. However, the results are only marginally better than those of other approaches such as SVM, RF, NB, and so on.

[2] This paper tries to quantify and evaluate implicit sexism in tweets. Implicit sexism also known as benevolent sexism is harder to classify. The authors use SVMs and other methods such as sequence to sequence models and FastText classifier. This paper gives the metrics to distinguish between implicit and explicit sexism.The study found that SVM has a slightly better f1 scores for the benevolent class and while the Seq2Seq classifier performs better for the hostile kind of sexism. FastText classifier has the best results with a f1 score 0.87- the highest.

[3] This is one of the benchmark dataset papers that presents a hate-speech annotated data set (includes many examples of sexist and misogynistic slurs). This paper describes

metrics for detecting sexist and misogynistic slurs, as well as other hostile and extreme forms of sexism, in a dataset.The study results show that the most effective approach is to use character n-grams of lengths up to 4 and add gender as an extra feature. The use of location or length is found to decrease the scores. The inclusion of gender information is observed to enhance the F1-score, while other features and their combinations have a negative impact on the system's performance. Statistical analysis reveals that adding location along with gender is significant with a p-value of 0.0355, whereas gender alone does not show statistical significance.

[4]Annotating and classifying data is an important aspect of NLP papers' dataset creation. This paper describes a method for reducing the need for expert annotators' annotation burden for sentiment analysis datasets. This paper proposes an agreement metric for integrating the annotation of different levels of annotators, both skilled and less skilled. This dataset and its annotations are tested by identifying the best performing features across multiple models and finalizing features based on f1 scores. The study finds that hate speech datasets should take into account the instances of hate speech being rare in real life into datasets as well so as to not make an unfair dataset.

[5]This paper utilizes deep learning techniques, GloVe, random embeddings and attention model based LSTMs are utilized to make better models that perform better on sifted datasets for sexism recognition.The paper demonstrates that the GloVe+BiLSTM+Attn model achieves an F1 score of 0.88, indicating that it is possible to achieve a level of performance in sexist detection that is comparable to previous research, using slightly different deep learning methods. Furthermore, it is important to note that this performance is achieved while constraining all data to be in a workplace context, which tends to involve more subtle forms of sexism such as "benevolent" sexism.

[6] This paper tries to tackle the problem statement through a two pronged approach. This paper finds an approach to classify not only the various kinds of sexism (i.e implicit, explicit or benevolent) but also try to find out the shades of sexism and a character analysis of each of the character’s sexism through a pre-trained BERT model.The addition of annotated dataset of sexist tweets as distant supervision helps improve the overall F1 score by 1.5% in identifying sexism in scripts. However, adding random data drops the performance by 14.7%. This indicates that randomly adding scripts can harm performance by injecting noise. The study tries to prove that distant supervision can improve performance while allowing analysis of different dimensions of sexism.

[7] This paper uses subtitles data set from movies over a century and tries to find fairness in the portrayal of different genders and bias in them. This paper shows the drawbacks of using a pre-trained BERT model and the kind of precautions we need to take while applying algorithms to analyze data that has heavy negative biases embedded in them.The study shows that pre-training BERT on film dialogue can introduce biases and social themes, with the strength and types of biases varying depending on the era of the film. Recent decades exhibit more explicit racial stereotypes while gendered associations are stronger in earlier decades. Underrepresentation of minority groups can also contribute to the lack of evidence for biases in datasets, which is a concern for downstream applications.

[8] This paper finds the relation between negative sexist stereotypes and the troupes they fall into. The authors evaluate the correlation between the popularity of a certain troupe and how it relates to a gendered view of whether that troupe is used because of it. The approaches used are NLTK and LDA to find that a classifier trained on Goodreads authors' books predicts author gender with 71% accuracy based on a binary feature vector of tropes. Female-authored books contain more female-leaning tropes, while male authors use more stereotypical tropes. Selection bias exists in the data, but the study only analyzes the tropes most predictive of author gender and does not attempt to identify an individual's gender.

[9] The paper shows how sexism is expressed in social media, specifically on Twitter. The authors develop an ML-based system to detect sexist expressions using BERT, which outperforms other algorithms with an accuracy of 74%. They also introduce a spanish language sexist tweets dataset. The authors acknowledge the limitations of their system due to linguistic phenomena and dataset size.

[10] The paper compares the effectiveness of traditional TF-IDF features and word embedding methods in sentiment analysis. The results show that word embedding outperforms TF-IDF in all metrics, particularly when used with RNN. The analysis also indicates that dataset size has a minimal influence on the results for word embedding. The study was extended to other datasets, where word embedding continued to show better performance. The findings suggest that word embedding is a more robust and efficient technique for sentiment analysis.

[11] This paper shows the drawbacks of how something is translated as offensive and obscene content in the subtitles in children's YouTube content when it's actually not so in reality.The authors use high performance language models such as BERT, XLM, XLNet, DistilBERT, and Megatron to find out where the error in subtitles and the subsequent transcription in content takes place.

[12] This paper provides a comparative analysis of gender bias and sexism in Bollywood and Hollywood movies over a seven decade long period using NLP sexism detection techniques on a corpus of multi-lingual data.

[13]This paper compares how men and women are portrayed in popular tv shows with the stereotypes and tropes attached to them. This paper provides the definition for gendered stereotypes and tropes.The study examined gender representation on Dutch television channels, finding that women were significantly underrepresented on men's channels and overrepresented in the fictional genre on those channels. Programs from the United States had a higher representation of women on men's channels, while Dutch and Other countries' programs had a significantly lower representation of women.

[14] This paper uses a TRAC-2 system to classify gendered aggression and misogyny in different contexts.The study found that misogyny was easier to detect than aggression in all languages due to its binary nature. The best performance was achieved in English sub-task B, where the system ranked 3rd out of 15 task systems (92% accuracy). The confusion matrices showed that covertly aggressive examples were more likely to be wrongly predicted than overtly aggressive, and Hindi had a higher overtly aggressive-non overtly aggressive confusion (73% accuracy) . Sub-task B confusion matrices showed a higher gendered-non- gendered confusion for Hindi, which was attributed to the significant difference in class distribution across the test data compared to the training and dev sets.

[15] This paper proposes a novel strategy for identifying sexism on the internet using annotated datasets and evaluating different ML models against each other.The study shows that adversarial examples can challenge the generalizability of models. BERT models perform better on such modified data with an f1 score of 0.80 and outperforming the other tested models.

[16] The study compared different approaches for classifying abusive language in tweets. The newly proposed HybridCNN outperformed the WordCNN in a one-step multi-class

classification, suggesting that the additional character input channel enhanced its performance. The CharCNN performed worse than the WordCNN. Two step approach performed better than one-step approach by more than 10 f1 points. Using HybridCNN on the first step and logistic regression on the second step performed better than using only HybridCNN.

[17] The study created two convolutional neural network (CNN) models for training and classifying tweets based on different input vector sets. The models were trained on word vectors generated using an unsupervised technique called word2vec and compared against a baseline of randomly generated vectors. Word2vec performed the best with an f1 score 78.3%. N-grams additionally increased precision.

[18] The study investigates the impact of using different types of counterfactual adversarial data (CAD) on the performance of hate speech and sexism detection models. The results show that models trained on adversarial data have higher f1 scores than non-counterfactual adversarial models.

[19] This study shows a pre-trained BERT model for Romanian language, achieved the highest Macro-F1 score of 0.75 for identifying sexist texts in the ROFEMPOL Corpus. However, all models performed better in identifying non-sexist texts than sexist ones. The machine learning models using BOW, TF-IDF, and BERT-based representations showed large performance differences between the two classes.The random forest model performed the best with an f1 score of 0.74 and 0.44 for non-sexist and sexist tweets respectively.

[20] The study explores the effectiveness of multitask learning in detecting sexist and gender stereotype content on Twitter. The results show that the inclusion of gender stereotype labels, as provided by an automatic classifier, outperforms two multitask baselines. Predicting the types of stereotypes proves to be more effective than identifying their presence, with an F-score of 0.796 compared to 0.776. However, the inclusion of gender stereotype information with sexist labels tends to decrease the results for all configurations except AngryBERT2 and AngryBERT4. AngryBERT4 was found to be better at predicting sexist content, with an F-score of 0.805 compared to 0.773 for the strong baseline BERT. The error analysis indicates that the majority of misclassified instances cannot benefit from the gender stereotype auxiliary task, confirming that sexist content does not necessarily involve stereotypes. In addition, humor, jokes, irony, and puns were found to be significant sources of classification errors. The results suggest that

accounting for these phenomena in hate speech detection is still an open problem. The best F-score obtained was 0.827, achieved by GS types, significantly outperforming BERT with a p-value of less than 0.05 using McNemar's Test statistic.

[21] The paper evaluates different machine learning models for the task of fine-grained multi-label classification to identify instances of sexism in online comments. The authors experiment with traditional machine learning and deep learning models, as well as propose new methods involving semi-supervised learning, multi-task learning, and objective functions that capture label correlations. They find that most proposed methods outperform all baselines across all metrics, with the best performing method involving auxiliary tasks and an L-cor loss function. The paper highlights the challenging nature of this task, but demonstrates that it is possible to identify instances of sexism with high accuracy using machine learning.

[22] This research analyzes 6.7 million case law documents for gender bias and finds that existing bias detection methods in NLP are inconsistent for testing bias in legal language. Two new approaches to building word lists for bias representation are suggested and tested on case law, finding both methods to be robust and consistent in identifying gender bias. The study also explores the intersectionality of gender and race bias in case law and the potential impacts on NLP systems' overall performance.

[23] The paper presents a study on the detection of hate and offensive speech in a novel movie subtitle dataset using various transfer learning techniques. The authors proposed a method to combine movie subtitle fragments and make social media text content more comparable for training purposes. They used domain adaptation and fine-tuning techniques for classification, evaluating three different ML models, namely Bag of Words, transformer-based, and BiLSTM-based models. The results showed that all three models performed well for the classification of the normal class, but BERT achieved a substantially higher F1-score in differentiating between offensive and hate classes.

[24] This study analyzes and finds novel methods to detect racism and evaluate the performance of hate speech classifiers against this metric. This study shows that such classifiers are more likely to negatively label data that is from a black aligned corpus.

[25] The best performing algorithm by far in this particular paper was RF-with sentiment. The goal was to differentiate and identify various forms of aggression in Twitter tweets. The proposed algorithm showed 97% recall and 56% precision.

METHODOLOGY:

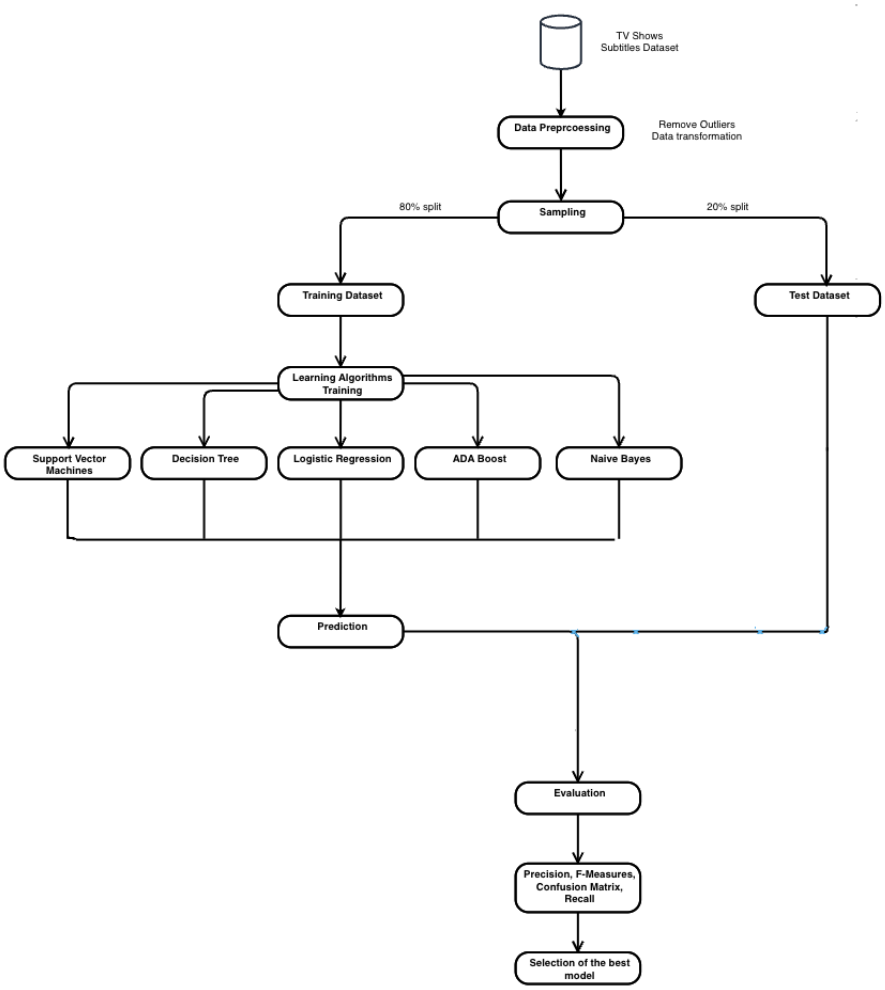

Figure-1 Data Flow diagram of the methodology used

Data Pre-processing:
The dataset [27] was initially processed by removing all rows with missing values and organizing the dialogues by their respective TV shows. To limit the scope of the current project, the attribute identifying the character speaking each dialogue was also removed. After the data cleaning, the dataset consisted of 33,570 annotated TV show dialogues. The data was then divided into a training set, containing newly annotated dialogues, and a test set.

Algorithms used:
Sexism is detected in the dataset using five different algorithms: SVM, Naive Bayes, Decision Tree, ADA Boost, and Logistic Regression. SVM, Naive Bayes, and Decision Tree are traditional machine learning algorithms, while AdaBoost is a statistical classification algorithm. All algorithms use a supervised learning approach.

1. **SVM**

Support Vector Machine (SVM) is frequently used in supervised machine learning for binary classification problems, but it may also be used for multiple-class classification and regression issues. The algorithm divides the feature space into two classes by drawing a hyperplane through it. The Maximal Margin Classifier, which maximizes the margin between the two classes, is the ideal hyperplane in SVM. The data points that are closest to the separating hyperplane are known as support vectors.

The goal of the hard margin SVM algorithm is to locate a hyperplane that precisely divides the classes.The goal of a soft-margin support vector machine is to maximize the margin width while minimizing the amount of slacks. Sequential Minimization Optimization can be used to resolve the resulting quadratic programming issue. The technique resolves this challenge by identifying the best hyperplane that maximizes the margin while properly classifying the most occurrences.

Hyper plane definition :

$$\boldsymbol{x} \cdot \tilde{\boldsymbol{w}} + \tilde{b} = 0$$

Hyperplane dividing the classes and maximizing M:

$$y_i \left(\boldsymbol{x_i} \cdot \tilde{\boldsymbol{w}} + \tilde{b} \right) \geq M \, .$$

$$y_i \left(\boldsymbol{x_i} \cdot w + b \right) \geq 1 \, .$$

$$\|\tilde{\boldsymbol{w}}\| = 1, \, \|\boldsymbol{w}\| = \tfrac{1}{M}$$

Support Vectors

$$\boldsymbol{x_i} * \boldsymbol{w} + b = 1 \text{ for positive class}$$

$$\boldsymbol{x_i} * \boldsymbol{w} + b = -1 \text{ for negative class}$$

Hard Margin SVM

$$\min_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{w}\|^2$$
$$\text{subject to } y_i(\boldsymbol{x_i} \cdot \boldsymbol{w} + b) \geq 1$$
$$\text{for } i = 1, \ldots, n$$

Soft Margin SVM and the parameter C

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\frac{1}{n}\sum_i \xi_i$$

$$\text{subject to } \begin{cases} y_i(\boldsymbol{x}\cdot\boldsymbol{w}+b) \geq (1-\xi_i) & \text{for } i = 1,\ldots,n \\ \xi_i \geq 0 & \text{for } i = 1,\ldots,n \end{cases}$$

Making it into an quadratic programming problem using dual Lagrangian Formulation

$$\min_{\boldsymbol{w},b,\xi} \max_{\alpha,\mu} \left[\frac{1}{2}\|\boldsymbol{w}\|^2 + C\frac{1}{n}\sum_i \xi_i - \sum_i \alpha_i \left[y_i(\boldsymbol{x_i}\cdot\boldsymbol{w}+b) - (1-\xi_i)\right] - \sum_i \mu_i\xi_i\right]$$

$$\max_{\alpha} \left[\sum_i \alpha_i - \frac{1}{2}\sum_{i,i'} \alpha_i\alpha_{i'}y_iy_{i'}\boldsymbol{x_i}\cdot\boldsymbol{x_{i'}}\right]$$

$$\text{subject to } \begin{cases} 0 = \sum_i \alpha_i y_i \\ 0 \leq \alpha_i \leq C & \text{for } i = 1,\ldots,n \end{cases} .$$

$$\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x_i} .$$

2. Naive Bayes

A Naive Bayes classifier is a probabilistic machine learning model that's used for this specific classification task. The crux of the classifier is based on the Bayes theorem.This is also a supervised learning algorithm. Multinomial, Bernoulli and Gaussian are various kinds of Naive Bayes Classifiers.

Bayes Theorem:

$$P(y|x_1,...,x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1,...,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Conditional Probability formula:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

3. Decision Tree

Decision Trees is a supervised learning algorithm that is used for both regression and classification tasks.As a result, decision trees are adaptable models that, if constructed properly, don't have their number of parameters increase as we add more features.They can also output either a categorical predictor or a numerical prediction .
Nodes and branches are the two types of elements used in their construction. One of the data's properties is assessed at each node in order to divide the observations during training or to direct a particular data point along a certain path during prediction.

Information and Gain:

$$\text{Info}(D) = -\sum_{i=1}^{m} pi \log_2 pi$$

$$\text{Info}_A(D) = \sum_{j=1}^{V} \frac{|Dj|}{|D|} \text{ X Info}(D_j)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

Gain Ratio:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

Gini Index:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

4. **Logistic Regression**

By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable. These binary results in logistic regression enable simple choice between two options. It is a supervised learning algorithm.

Sigmoid Function:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

$$\log \frac{p(x)}{1 - p(x)} = \alpha_0 + \alpha . x$$

$$p(x) = \frac{e^{\alpha_0 + \alpha}}{e^{\alpha_0 + \alpha} + 1}$$

Likelihood:

$$l(\alpha_0, \alpha) = \sum_{i=0}^{n} -\log 1 + e^{\alpha_0 + \alpha} + \sum_{i=0}^{n} y_i(\alpha_0 + \alpha . x_i)$$

Maximum Likelihood Estimation:

$$\frac{\partial l}{\partial \alpha_j} = \sum_{i=0}^{n} \big(y_i - p(x_i; \alpha_0, \alpha)\big) x_{ij}$$

5. **AdaBoost**

In adaptive boosting, weights are redistributed among instances, with examples that were mistakenly categorized receiving higher weights. Boosting increases learners sequentially, reducing bias and variation in supervised learning. With the exception of the first student, each succeeding learner is constructed from earlier-grown learners. Learning weaknesses are turned into strengths.

Sample weight

$$sample\ weight = \frac{1}{\#\ of\ samples}$$

Gini Impurity

$$Gini\ Impurity = 1 - (the\ probability\ of\ True)^2 - (the\ probability\ of\ False)^2$$

Amount of say

$$Amount\ of\ say = \frac{1}{2}\log\left(\frac{1 - total\ error}{total\ error}\right)$$

New Sample weights

$$New\ Sample\ Weight\ For\ \textbf{Incorrect}\ Samples = Sample\ weight * e^{amount\ of\ say}$$

$$New\ Sample\ Weight\ For\ \textbf{Correct}\ Samples = Sample\ weight * e^{-amount\ of\ say}$$

RESULTS:

SVM performs the best on the given dataset for classifying sexist content with a f1 score of 0.88 for classifying the sexist text. The second best performing algorithm for classification is the AdaBoost algorithm for classification of sexism with a f1 score of 0.87 for classifying sexist text.The third best algorithm for classification is the decision tree algorithm for classification of sexism with a f1 score of 0.85 for classifying text that is sexist. Both Logistic Regression and Naive Bayes perform rather poorly with classifying sexist subtitles of TV shows. Though LR and Naive Bayes have an weighted average f1 score of 0.96 that is due to it being able to classify neutral texts correctly. A large part of the corpus is text with neutral insinuations and hence the f1 score of classifying sexist texts correctly is the ideal metric for comparison.SVM performs marginally better than AdaBoost, however SVM's performance is better than the Decision Tree by almost 3.4% and AdaBoost beats SVM by having a f1 score 2.3% better. So in this experiment the algorithm that gives the most correct classification of sexism in a dialog corpus, in the decreasing order is SVM ,AdaBoost,Decision Tree ,Naive Bayes,LR.

Table-1 Sexist content in popular English TV sitcoms as classified by the respective algorithms.

Model		**Precision**	**Recall**	**f1-score**	**Support**
	Neutral	0.97	1	0.98	6393
	Sexist	0.99	0.28	0.44	321
LR	Accuracy			0.97	6714
	Macro Avg	0.98	0.64	0.71	6714
	Weighted Avg	0.97	0.97	0.96	6714
	Neutral	0.99	1	.0.99	6393
	Sexist	0.9	0.86	0.88	321
SVM	Accuracy			0.99	6714
	Macro Avg	0.95	0.93	0.94	6714
	Weighted Avg	0.99	0.99	0.99	6714
	Neutral	0.97	1	0.98	6393
	Sexist	0.99	0.28	0.44	321
Naive Bayes	Accuracy			0.97	6714
	Macro Avg	0.98	0.64	0.71	6714
	Weighted Avg	0.97	0.97	0.96	6714
	Neutral	0.99	0.99	0.99	6393
	Sexist	0.86	0.83	0.85	321
Decision Tree	Accuracy			0.99	6714
	Macro Avg	0.92	0.91	0.92	6714
	Weighted Avg	0.99	0.99	0.99	6714
	Neutral	0.99	1	0.99	6393
	Sexist	0.91	0.83	0.87	321
AdaBoost	Accuracy			0.99	6714
	Macro Avg	0.95	0.92	0.93	6714
	Weighted Avg	0.99	0.99	0.99	6714

Confusion Matrix of Various Algorithms Used:

1. **Logistic Regression**

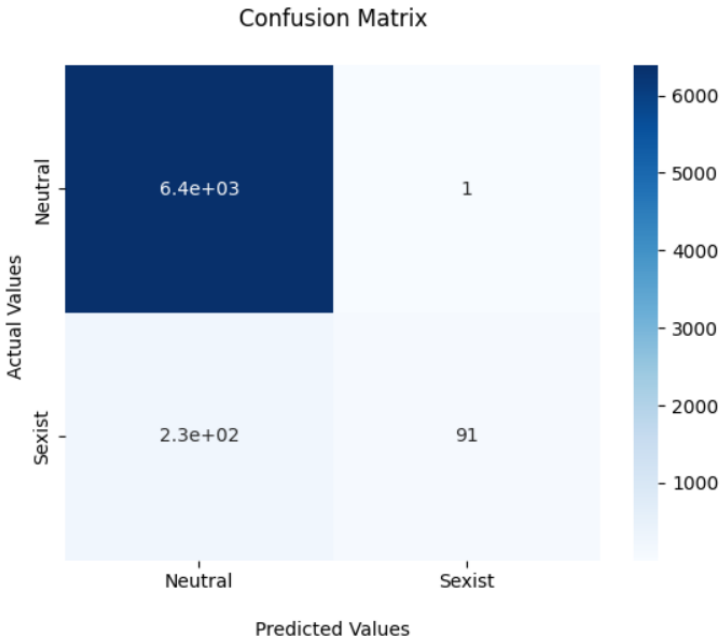

Figure-2 Confusion Matrix for Logistic Regression

2. **SVM**

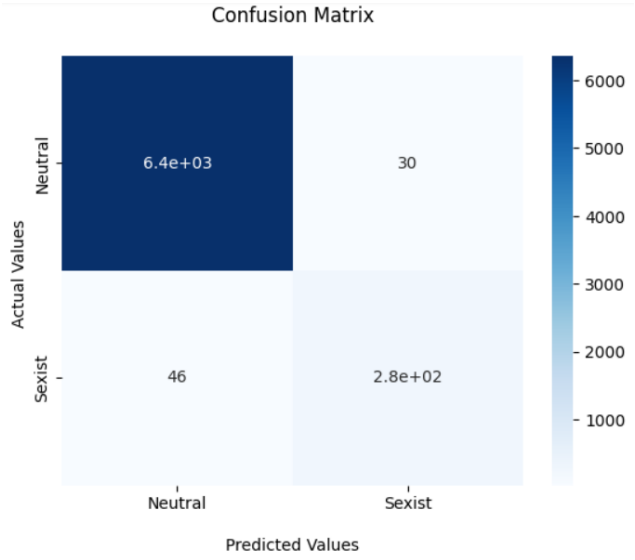

Figure-3 Confusion Matrix for SVM (Support Vector Machine)

3. Naive Bayes

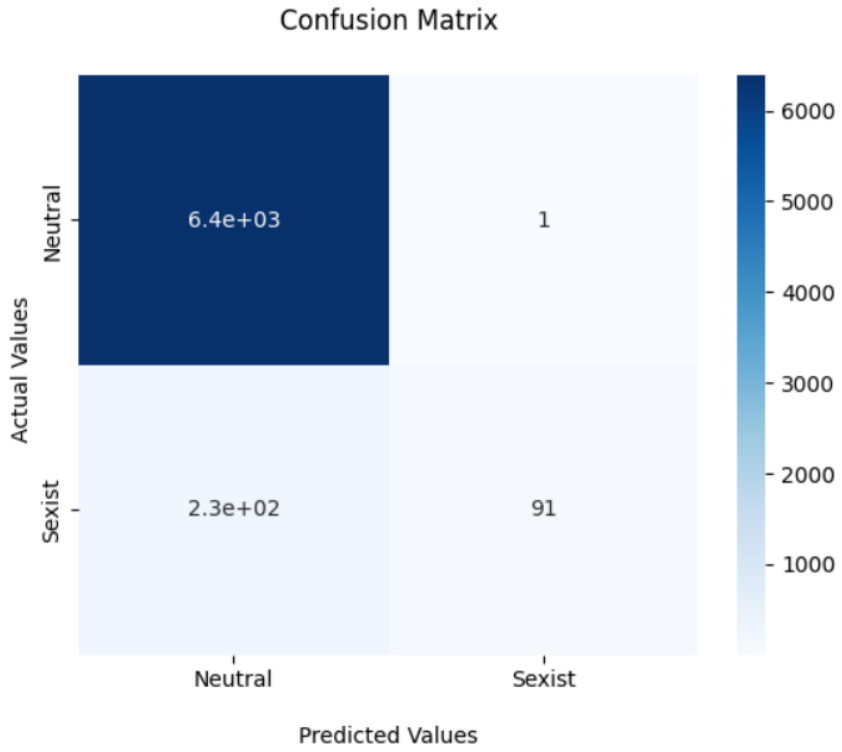

Figure-4 Confusion Matrix for Naive Bayes

4. Decision Tree

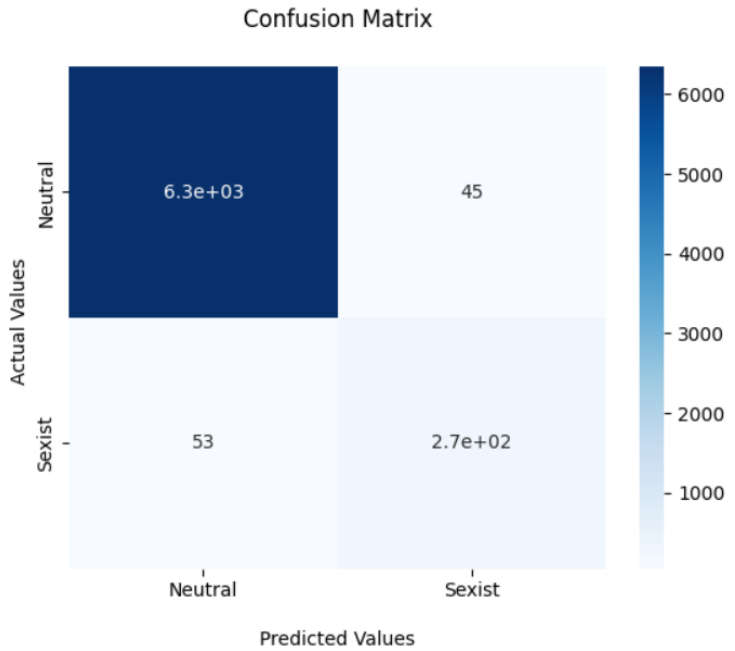

Figure-5 Confusion Matrix for Decision Tree

5. AdaBoost

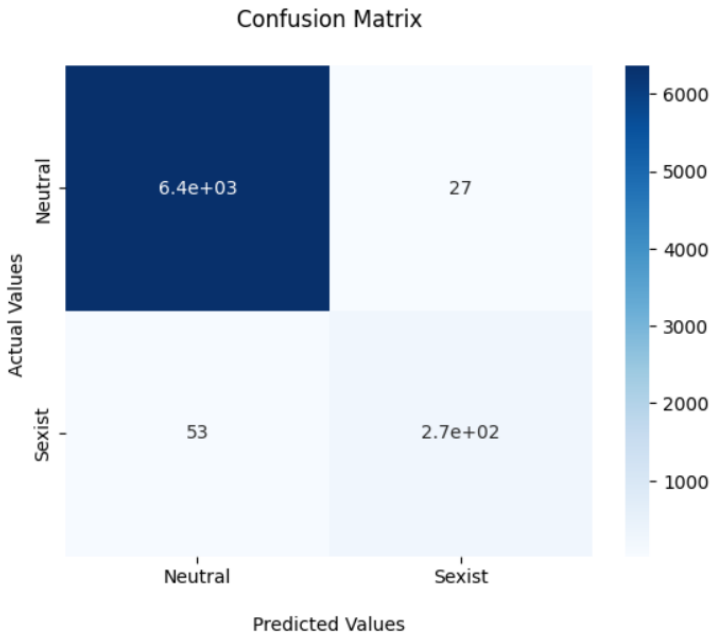

Figure-6 Confusion Matrix for AdaBoost

CONCLUSION:

Our project involves generating a labeled dataset of dialogues from TV show scripts to enable the quantification of sexism present in certain TV shows. This dataset can further be used to raise the awareness of viewers regarding the kind of content they are consuming. Our approach demonstrates the feasibility of using simple and computationally inexpensive algorithms for classifying and detecting sexism in TV shows. A similar approach can also be used for sexism detection in popular media of other languages on the availability of multilingual dataset and annotators. The reduction of sexism and the promotion of awareness of sexism in popular cultural media can greatly contribute to mitigating the problem of sexism.

CODE:

SVM
Naive Bayes
Decision Tree
Logistic Regression
AdaBoost

REFERENCES:

[1] Smriti Singh & Anand, Tanvi & Chowdhury, Arijit & Waseem, Zeerak,"Hold on honey, men at work": A semi-supervised approach to detecting sexism in sitcoms," in *Conference: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 2021.

[2] Akshita Jha, Akshita & Mamidi, Radhika,"When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data," in *Conference: Second Workshop on Natural Language Processing and Computational Social Science (NLP+CSS) at ACL, at Vancouver, Canada,* 2017.

[3] Waseem, Zeerak & Hovy, Dirk,"Hateful Symbols or Hateful People?," in *Predictive Features for Hate Speech Detection on Twitter. Conference: Proceedings of the NAACL Student Research Workshop,* 2016.

[4] Waseem, Zeerak,"Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter,"in *Conference: Proceedings of the First Workshop on NLP and Computational Social Science.,*2016.

[5] Dylan Grosz, Patricia Conde-Cespedes,"Automatic Detection of Sexist Statements Commonly Used at the Workplace ," in *Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD) Wokshop (Learning Data Representation for Clustering) LDRC*, 2020

[6] Lee Nayeon, Bang, Yejin ,Shin Jamin, Fung Pascale, "Understanding the Shades of Sexism in Popular TV Series," in *WiNLP*, 2019.

[7] Amanda Bertsch, Ashley Oh, Sanika Natu, Swetha Gangu, Alan W. Black, and Emma Strubell,"Evaluating Gender Bias Transfer from Film Data," in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 235–243, Seattle, Washington. Association for Computational Linguistics,* 2022.

[8] Gala, Dhruvil & Khursheed, Mohammad & Lerner, Hannah & O'Connor, Brendan & Iyyer, Mohit. "Analyzing Gender Bias within Narrative Tropes," *In Proceedings of the*

Fourth Workshop on Natural Language Processing and Computational Social Science, pages 212–217, Online. Association for Computational Linguistics, 2020.

[9] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz and L. Plaza,“Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data,” in *IEEE Access, vol. 8, pp. 219563-219576, 2020*

[10] Nhan Cach Dang , María N. Moreno-García and Fernando De la Prieta D. Soam and S. Thakur, "Sentiment Analysis Using Deep Learning: A Comparative Study," *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, *Gunupur, India, pp. 1-6, 2022*

[11] Ramesh, K., KhudaBukhsh, A. R., & Kumar, S. “Beach’ to ‘Bitch’: Inadvertent Unsafe Transcription of Kids,” Content on YouTube. *Proceedings of the AAAI Conference on Artificial Intelligence,* 2022.

[12] Kunal Khadilkar, Ashiqur R. KhudaBukhsh, Tom M. Mitchell, Gender bias, social bias, and representation in Bollywood and Hollywood, in *Patterns,Volume 3, Issue 2,* 2022.

[13] Daalmans, S., Kleemans, M., & Sadza, A. Gender Representation on Gender-Targeted Television Channels: A Comparison of Female- and Male-Targeted TV Channels in the Netherlands. *Sex Roles, 77*, 366 - 378, 2017

[14] Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, Thamar Solorio,”Aggression and Misogyny Detection using BERT: A Multi-Task Approach,” in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 126–131, Marseille, France. European Language Resources Association (ELRA),* 2020.

[15] Mattia Samory, Indira Sen , Julian Kohne , Fabian Flöck, Claudia Wagner,"“Call me sexist, but...”: Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples"," in *the Proceedings of the Fifteenth International Conference on Web and Social Media,* 2021.

[16]Ji Ho Park and Pascale Fung. “One-step and Two-step Classification for Abusive Language Detection on Twitter,”in *Proceedings of the First Workshop on Abusive Language Online, pages 41–45, Vancouver, BC, Canada. Association for Computational Linguistics* ,2017.

[17] Björn Gambäck and Utpal Kumar Sikdar. “Using Convolutional Neural Networks to Classify Hate-Speech,”in *Proceedings of the First Workshop on Abusive Language Online, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics* , 2017.

[18] Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. “Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection.” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4716–4726, Seattle, United States. Association for Computational Linguistics,* 2022.

[19] Andreea Moldovan, Karla Csürös, Ana-maria Bucur, and Loredana Bercuci.“Users Hate Blondes: Detecting Sexism in User Comments on Online Romanian News,” *In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 230–230, Seattle, Washington (Hybrid). Association for Computational Linguistics*, 2022.

[20] Patricia Chiril, Farah Benamara, and Véronique Moriceau. “Be nice to your wife! The restaurants are closed”: Can Gender Stereotype Detection Improve Sexism Classification?, ” in *Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2833–2844, Punta Cana, Dominican Republic. Association for Computational Linguistics,* 2021.

[21] Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. “Semi-supervised Multi-task Learning for Multi-label Fine-grained Sexism Classification,” in *Proceedings of the 28th International Conference on Computational Linguistics, pages 5810–5820, Barcelona, Spain (Online). International Committee on Computational Linguistics,* 2020.

[22] Baker Gillis,"Sexism in the Judiciary: The Importance of Bias Definition in NLP and In Our Courts," in *GeBNLP*, 2021

[23] Niklas von Boguszewski, Sana Moin, Anirban Bhowmick, Seid Muhie Yimam, and Chris Biemann," "How Hateful are Movies?" :A Study and Prediction on Movie Subtitles ," in *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021), pages 37–48, Düsseldorf, Germany,* 2021.

[24] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. "Racial Bias in Hate Speech and Abusive Language Detection Datasets," *In Proceedings of the Third Workshop on Abusive Language Online, pages 25–35, Florence, Italy. Association for Computational Linguistics,* 2019.

[25] Constantin Orăsan. "Aggressive Language Identification Using Word Embeddings and Sentiment Features," *In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pages 113–119, Santa Fe, New Mexico, USA. Association for Computational Linguistics,* 2018.

[26] Smriti Singh & Anand, Tanvi & Chowdhury, Arijit & Waseem,2021, "HHMW Dataset," "Hold on honey, men at work": A semi-supervised approach to detecting sexism in sitcoms," in *Conference: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*,.
[Online]. Available :https://github.com/smritisingh26/HHMWdataset