A Semi-Supervised Approach for Sexism Detection in K-Pop Posts

Britney G. Beligan
Computer Science Department
New Era University
britney.beligan@neu.edu.ph

Aubrey Min M. Lasala
Computer Science Department
New Era University
aubreymin.lasala@neu.edu.ph

Abstract *- Hallyu, the global rise of Korean pop culture, has notably influenced the entertainment industry. Social media, especially for K-POP, serves as a participatory space where fans support idols, share content, and drive trends. However, it also hosts verbal abuse, including sexism, manifested through gender-based prejudice and hate comments. This study detects sexism in K-POP-related posts using semi-supervised learning. It uses two datasets: labeled (8,160 posts) and unlabeled (10,782 posts). A baseline hybrid model (Stacked Embeddings + CNN with Attention) trained on the labeled data achieved 85.83% test accuracy. Self-training generated pseudo labels for the unlabeled dataset, which were added to retrain the baseline model. Over five iterations, performance improved by 6%, with final results of 91.84% accuracy, 90.10% precision, 94.79% recall, 92.38% F1-score, and 96.96% ROC-AUC. Error analysis on misclassified tweets was conducted. The model is deployed in a K-pop Sexism Detector Google Extension prototype connected to the Reddit API. Future work includes expanding the annotated dataset, using transformer models and embeddings, adopting KPOP discourse embeddings, and exploring other platforms.*

Keywords - *Attention, Convolutional Neural Networks, Hybrid Deep Learning Model, Natural Language Processing, Stacked Word Embeddings, Semi-Supervised Learning, Sexism, Tweets, Twitter, Pseudo label, Self Learning*

I. INTRODUCTION

Korean pop music, widely known as K-pop, is a popular genre of music originating from South Korea and one part of the Korean Wave or Hallyu, a term referring to the popularity of Korean pop culture like tv shows, music, and movies across Asia and other parts of the world (Romano, 2018).

As a music genre and a thriving industry, K-pop has reached global influence, with artists consistently dominating international streaming platforms like YouTube and Spotify, with different music styles such as pop, electronic, hip-hop, R&B, etc. Its cultural impact extends globally due to its captivating performances, music, fashion, and entertainment trends. Social media has become an essential participatory space for multiple popular cultures, especially K-pop. Nowadays, fans have advanced with technology to actively support and engage with their idols, circulate information, and produce numerous gifts or products based on them, such as fan fiction, fan videos, fan art, etc (Saraswati & Nurbaity, 2020).

For example, BTS (also known as Bangtan Sonyeondan or Bangtan Boys) has garnered an “ARMY” of committed fans on every platform and established a close relationship with them. Saraswati and Nurbaity (2020) stated that fans feel connected with their songs and their activity on social media. Yet, this active participation is often bashed by numerous trends on Twitter (now X) that have echoes massively.

The industry is also facing challenges like traditional notions of masculinity and femininity, with artists often pushing boundaries in terms of beauty standards and gender expression. One of the biggest challenges faced by the idols is the instances of verbal abuse, cyberbullying, and physical violence involving sexism that affect their mental health (Vision D., 2024).

II. BACKGROUND

Sexism describes prejudice or discrimination based on sex or gender that affects every level of society (Villines, 2021). The European Institute for Gender Equality (2024) defines sexism as being rooted in beliefs about the fundamental nature of women and men and the societal roles they are expected to fulfill. These sexist assumptions manifest as gender stereotypes, which often rank one gender as superior to another. This hierarchical thinking may be conscious and hostile or unconscious in the form of bias. While sexism can affect everyone, it disproportionately impacts women. On social media, the depiction of sexism carries real-world implications. Repeated exposure to biased or harmful content can influence audience attitudes and behaviors, normalizing discriminatory thinking and perpetuating inequality. This highlights the urgent need for tools to detect and reduce sexism in media content to promote fairer gender representation.

Twitter (now X) is recognized as the most active platform for K-pop-related interactions. This engagement is visible through voting events, donations, trend hashtags, and fandom projects (Nugraha & Komsiah, 2023). Despite the implementation of time and tweet limits on aggressive accounts and Twitter's existing regulations, its systems remain insufficient in detecting domain-specific, sexism-related posts. This study responds to that gap by focusing on sexism detection in K-pop posts. Previous studies have shown that sexism within the K-pop industry and fan community is harmful, as idols are often subjected to physical standards imposed by companies and fans, overshadowing their talents and achievements. Research has explored sexism in news coverage and across various social media platforms, including Twitter. This study seeks to provide focused insight into K-pop's unique environment while advancing methods for automated sexism detection.

To address this issue effectively, the study adopts a hybrid model architecture combined with semi-supervised learning techniques. Hybrid learning approaches have improved classification accuracy, especially in low-volume and high-dimensional data scenarios. Given the limited availability of labeled data related to K-pop sexism, this study utilizes semi-supervised learning methods, specifically self-training and pseudo-labeling, to achieve its objectives. The broader goal of this research is to foster a more respectful online space while challenging the harmful biases perpetuated within the K-pop industry. By applying practical machine learning solutions to a significant social issue, this work aims to contribute to academic knowledge and advocate for a more inclusive online community.

Therefore, the study seeks to raise awareness within the online community, particularly among K-pop audiences, about the harmful impact of sexist hate comments by developing a model to detect sexism using semi-supervised learning on K-pop-related posts. The specific objectives of the study are as follows: (1) to apply preprocessing techniques to both a labeled sexism dataset and an unlabeled K-pop dataset, (2) to classify K-pop tweets as sexist or non-sexist through a hybrid deep learning model enhanced by semi-supervised learning, (3) to evaluate the model's performance using appropriate evaluation metrics, and (4) to implement the model in a web extension prototype designed to detect sexism in Reddit posts.

III. LITERATURE REVIEW

Sexism remains deeply embedded in the K-pop industry, particularly in the portrayal and treatment of female idols. Women are pressured to conform to rigid beauty standards—like "ant waists" and "11-abs"—that overshadow their talents (Joshi, 2022; Nickeled & Dimed, 2021). Unlike male idols, who have more freedom in image expression, female idols are often confined to roles emphasizing fragility and submission (IvyPanda, 2022; Lin & Rudolf, 2017). This double standard extends to media coverage and online discourse, where women face harsher criticism and sexualization compared to men (Ingrid, 2024; Rika et al., 2024). Such portrayals not only reinforce societal gender norms but also influence public perception, as the media serves both as a reflection and reinforcer of ideology (Dai & Xu, 2014).

To counter this, automated sexism detection has gained momentum. Early work by Waseem and Hovy (2016) and Jha and Mamidi (2017) laid the groundwork using annotated tweets and models like SVM and BiLSTM. Deep learning approaches, including CNNs and LSTMs with word embeddings, have since improved detection accuracy (Badjatiya, 2017). Competitions like EXIST and EDOS have furthered innovation through interpretable, transformer-based models (Rallabandi et al., 2023; Kirk et al., 2023).

Word embeddings are central to these advances. GloVe captures global word co-occurrence (Pennington et al., 2014), while fastText uses subword information for rare terms (Bojanowski et al., 2017). Their combination has proven effective in detecting hate speech (Mollas et al., 2022). Though BERT excels in contextual understanding, GloVe and fastText remain practical for resource-constrained settings (Agarwal et al., 2021; Badri et al., 2022).

Hybrid deep learning models enhance performance by integrating multiple techniques. CNNs extract local text patterns (Chavan, 2023), and combining them with LSTMs and attention improves the classification of complex content like sexism (Kalra & Zubiaga, 2021). Attention mechanisms highlight relevant input segments, boosting both accuracy and interpretability (Bahdanau et al., 2014; Berggmann & Stryker, 2024). Recent studies show hybrid models consistently outperform single-architecture systems (Jayapal et al., 2021; Kumar et al., 2024; Belbachir et al., 2024).

Given limited labeled data in niche domains like K-pop sexism, semi-supervised learning (SSL) is a viable solution. Self-training—where models iteratively label and learn from unlabeled data—has shown promise in offensive speech and sexism detection (Berton & Duarte, 2023; Alsafari & Ssadaoui, 2021). Advanced SSL techniques, like pseudo-labeling with confidence thresholds or domain-adapted models, enhance performance even further (Abburi et al., 2021; Li et al., 2021). These approaches demonstrate that hybrid architectures and SSL are practical tools for combating online sexism, particularly in underrepresented contexts.

IV. METHODOLOGY

The project follows four main phases: Data Preparation, Model Training, Semi-supervised Learning, and Evaluation and Testing. In Data Preparation, K-POP tweets and a labeled sexism dataset are collected, cleaned using social media text preprocessing, and split into training, validation, and test sets. The environment is also set up for model training. Model Training involves building and training a hybrid model over several epochs, using validation data to adjust parameters. The trained model is then saved for the next phase. In Semi-supervised Learning, self-training generates pseudo-labels for unlabeled data, which are added to the labeled set to retrain and improve the model. Finally, Evaluation and Testing measure the model's performance using accuracy, precision, recall, and F1-score, along with a prototype to show real-world use.

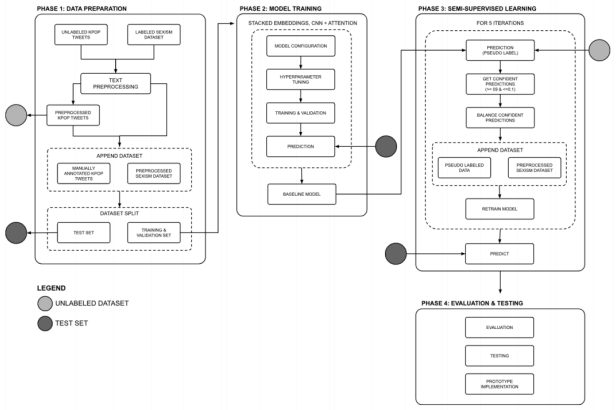

Figure 1. Project Design of the Study

Data Gathering

The unlabeled dataset consists of 11,089 K-pop-related tweets. Of these, 1,040 English tweets were sourced from Sainez and Wu (2022), focusing on the Kim Garam controversy. The remaining 10,049 tweets were manually scraped using keywords related to K-pop scandals such as Lisman, Jenslut, and Aesplastic. After removing duplicates, non-English rows, and empty rows, the two datasets were merged.

The EXIST 2021 dataset, containing 5,644 training and 2,208 test samples labeled as sexist or non-sexist, was used for labeled data. To capture K-pop-specific sexism nuances, 307 K-pop tweets were manually annotated and added to training and test sets, ensuring domain relevance (Anand & Singh, 2021). With this, the final unlabeled dataset consisted of 10,782 tweets.

Data Preprocessing

Tweets were cleaned using standard social media text techniques, including the removal of URLs, HTML tags, and punctuation; anonymizing mentions; expanding contractions; normalizing numbers, slang, and elongated characters; splitting CamelCase hashtags; removing non-ASCII characters; and converting to lowercase. These steps reduce noise and improve embedding quality. The pseudocode below summarizes the preprocessing techniques applied:

Algorithm:

Input: A raw text string

Output: A cleaned and standardized version of the text

1. *Remove all URL*
2. *Remove all HTML tags*
3. *Replace all user mentions (e.g., @username) with "[USER]"*
4. *Expand all contractions using Python library contraction*
5. *Remove all punctuation marks*
6. *Convert all numeric digits into words using Python library inflect*
7. *Split CamelCase words*
8. *Remove hashtag symbols ('#')*
9. *Normalize elongated characters (e.g., "soooo" → "so") using a repetition reduction rule*
10. *Replace slang and sexist abbreviations using a curated dictionary*
11. *Remove all non-ASCII characters*
12. *Convert entire text to lowercase*
13. *Return the cleaned text*

Data Splitting

The labeled dataset was stratified and randomly split into training (60%), validation (15%), and test (25%) sets. The training data consisted of 4,673 rows, validation data consisted of 1,169 rows, and the test set consisted of 2,279 rows. With this, he overall dataset used for training the baseline model consists of 8,121 rows.

Model Development

The model leveraged stacked word embeddings, combining pre-trained GloVe (non-trainable) and FastText (trainable) embeddings to capture general and domain-specific semantics. Figure 1 illustrates the overall model architecture. Input sequences of tokenized tweets are passed through parallel embedding layers, concatenated, and regularized with SpatialDropout1D (rate 0.4) to reduce overfitting. A 1D CNN with 64 filters (kernel size 3) extracted local features, followed by LayerNormalization for training stability. An attention mechanism highlighted important features by computing attention weights over CNN outputs. The attended features were pooled via GlobalAveragePooling1D and GlobalMaxPooling1D, concatenated, and fed into fully connected layers with dropout (rate 0.3). The output layer used sigmoid activation for binary classification. The model was compiled with the Adam optimizer (learning rate 5e-5) and binary cross-entropy loss.

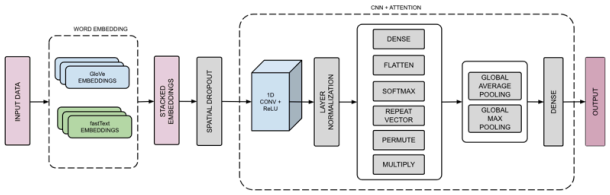

Figure 2. Model Architecture

Model Training

Training ran for up to 50 epochs with batch size 32, employing early stopping and learning rate reduction callbacks based on validation performance to avoid overfitting. Hyperparameters were tuned accordingly. The best-performing model was saved for generating pseudo-labels in the semi-supervised phase.

Self-Training

The saved model predicted labels for the unlabeled dataset. Two confidence thresholds were used to select pseudo-labeled samples: ≥ 0.9 for sexist and ≤ 0.1 for non-sexist predictions. An equal number of confident samples from each class was selected to maintain class balance. These pseudo-labeled samples were merged with the original labeled training data, shuffled, and split into new training and validation sets for five iterations.

The model was retrained on this expanded dataset with early stopping and learning rate reduction callbacks. This iterative process continued until the unlabeled data no longer yielded sufficiently confident predictions. Metrics such as accuracy, loss, and the number and distribution of pseudo-labeled samples were monitored and visualized to track learning progress.

The final model was evaluated on a held-out test set to assess generalization after semi-supervised learning.

Evaluation

This study used performance measurements such as accuracy, precision, recall, F1-score, Confusion Matrix, ROC curve, and AUC score.

Accuracy measures the overall correctness of the model by calculating the proportion of accurate

predictions (both positive and negative) to the total predictions.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \quad \textit{eq. 1. Accuracy}$$

Precision evaluates how many of the predicted sexist posts were actually sexist, thereby reflecting the model's reliability in identifying positive instances.

$$PC = \frac{TP}{TP + FP} \quad \textit{eq. 2. Precision}$$

Recall assesses the model's ability to identify all actual sexist content by calculating the ratio of true positives to the total number of actual positives.

$$RC = \frac{TP}{TP + FN} \quad \textit{eq. 3. Recall}$$

The F1-score, the mean of precision and recall, is a balanced metric when dealing with class imbalance, providing a more nuanced insight than accuracy alone.

$$F1 = \frac{2 * PC * RC}{PC + RC} \quad \textit{eq. 4. F1-score}$$

A confusion matrix was generated to visualize the distribution of true positives, true negatives, false positives, and false negatives, offering a granular view of the model's strengths and weaknesses.

The ROC curve was plotted to illustrate the trade-off between the true positive rate and the false positive rate at various classification thresholds.

The AUC score, which represents the area under the ROC curve, was used as a single scalar value to summarize the model's discriminative power across all thresholds. A higher AUC indicates a stronger ability of the model to distinguish between sexist and non-sexist content, thereby reinforcing its effectiveness in practical deployments.

Prototype Implementation

The model was deployed in a Google Chrome extension developed in Visual Studio Code. The backend used Flask to host the model, tokenizer, and preprocessing pipeline, exposing inference via HTTP. The extension accepts Reddit post URLs and uses the Reddit API (via PRAW) to dynamically fetch posts and top-level comments. Retrieved texts are preprocessed identically to the training data before classification, enabling live detection of sexism on Reddit content.

V. RESULTS

Preprocessing techniques were applied to the labeled and unlabeled datasets to ensure consistency and model readiness. Table 1 presents a randomly selected dataset text with data preprocessing results, where the removal of URLs and HTML, replacement of mentions, removal of punctuation, camel case splitting, and lowercasing were applied.

Table 1. Data Preprocessing Results

Process	Result
input	Absence of clean drinking water in Buikwe district responsible for increased cases of #domesticViolence the women say that they walk long journeys looking for clean get tired and fail to please their husbands in bed,the men spank them @username @username @username https://t.co/vKj6sSWdtA
Remove URL & HTML	Absence of clean drinking water in Buikwe district responsible for increased cases of #domesticViolence the women say that they walk long journeys looking for clean get tired and fail to please their husbands in bed,the men spank them @username @username
Replace Mentions	Absence of clean drinking water in Buikwe district responsible for increased cases of #domesticViolence the women say that they walk long journeys looking for clean get tired and fail to please their husbands in bed,the men spank them [USER]
Remove punctuation	Absence of clean drinking water in Buikwe district responsible for increased cases of domesticViolence the women say that they walk long journeys looking for clean get tired and fail to please their husbands in bed the men spank them [USER]
Camel case split	Absence of clean drinking water in Buikwe district responsible for increased cases of domestic Violence the women say that they walk long journeys looking for clean get tired and fail to please their husbands in bed the men spank them [USER]
Lowercase	absence of clean drinking water in buikwe district responsible for increased cases of domestic violence the women say that they walk long journeys looking for clean get tired and fail to please their husbands in bed the men spank them [user]

To classify the K-pop-related tweets as sexist and non-sexist, the study followed a multi-stage approach centered on a binary text classification task. The process involved manually annotating a subset of K-POP-related tweets to add to the labeled data. The annotation process was guided by criteria drawn from prior research on online sexism, as detailed in Table 2, ensuring consistency and reliability in labeling.

Table 2. Manual Annotation Criteria

Labels	Criteria
Sexist	• Use of derogatory sexist terms (e.g., Jenslut, Stiffsoo, and Aesplastic) • Objectification • Slutshaming • Insults and stereotyping • Body-shaming • Undermining their career based on their gender • Extreme gender-based comparison • Double standards
Non-sexist	• Contain no sexual, appearance-based, or gender-based attacks • Neutral or critical • Typical fan discourse

Afterwards, a baseline model (Stacked Embeddings, CNN + Attention) was trained. The model was compiled using the Adam optimizer with a learning rate of 5e-5. The loss function used was binary cross-entropy, which is appropriate for binary classification problems. The training underwent a total of 36 epochs, where

ReduceLROnPlateau was called. Figure 3 presents the model summary for the baseline model.

Model: "model_34"

Layer (type)	Output Shape	Param #	Connected to
input_36 (InputLayer)	[(None, 100)]	0	[]
embedding_70 (Embedding)	(None, 100, 200)	6000000	['input_36[0][0]']
embedding_71 (Embedding)	(None, 100, 300)	9000000	['input_36[0][0]']
concatenate_69 (Concatenate)	(None, 100, 500)	0	['embedding_70[0][0]', 'embedding_71[0][0]']
spatial_dropout1d_38 (SpatialDropout1D)	(None, 100, 500)	0	['concatenate_69[0][0]']
conv1d_35 (Conv1D)	(None, 100, 64)	96064	['spatial_dropout1d_38[0][0]']
layer_normalization_34 (LayerNormalization)	(None, 100, 64)	128	['conv1d_35[0][0]']
dense_103 (Dense)	(None, 100, 1)	65	['layer_normalization_34[0][0]']
flatten_35 (Flatten)	(None, 100)	0	['dense_103[0][0]']
activation_35 (Activation)	(None, 100)	0	['flatten_35[0][0]']
repeat_vector_35 (RepeatVector)	(None, 64, 100)	0	['activation_35[0][0]']
permute_35 (Permute)	(None, 100, 64)	0	['repeat_vector_35[0][0]']
multiply_35 (Multiply)	(None, 100, 64)	0	['layer_normalization_34[0][0]', 'permute_35[0][0]']
global_average_pooling1d_34 (GlobalAveragePooling1D)	(None, 64)	0	['multiply_35[0][0]']
global_max_pooling1d_34 (GlobalMaxPooling1D)	(None, 64)	0	['multiply_35[0][0]']
concatenate_70 (Concatenate)	(None, 128)	0	['global_average_pooling1d_34[0][0]', 'global_max_pooling1d_34[0][0]']
dropout_74 (Dropout)	(None, 128)	0	['concatenate_70[0][0]']
dense_104 (Dense)	(None, 64)	8256	['dropout_74[0][0]']
dropout_75 (Dropout)	(None, 64)	0	['dense_104[0][0]']
dense_105 (Dense)	(None, 1)	65	['dropout_75[0][0]']

Total params: 15,104,578
Trainable params: 9,104,578
Non-trainable params: 6,000,000

Figure 3. Model Summary

The probability of each input of sexism is acquired using the predict method. Table 3 presents the first ten texts related to K-pop topics in the test set. The table displays the text, the true label, the predicted label, and its probability score, which is calculated using a threshold of 0.5. Probabilities greater than 0.5 are indicated as 1 or sexist. On the other hand, all text below 0.5 will be considered as 0 or non-sexist.

Table 3. Probability score of K-pop related topics

Text	True Label	Predicted Label	Probability
jisoo is a hag with <OOV> <OOV> face above all that she is a talentless <OOV> but you are worried about rs focus on mass reporting her fancams because she is getting dragged for her stiff ass	1	1	0.742176
user lisa fashion <OOV> she dresses like a man and grandma jisoo is talentless you all blinks are so disgusting	1	1	0.787914
user ya everyone <OOV> they are talentless girls and <OOV> girls and plastic faces they are not worth to be queens when so many talented girls group are in korea also a childish like you does not know how to <OOV> between what is right and wrong that is why you are still like this	1	1	0.634255
the whole world knows jisoo is dozen talentless hag keep delusional because blonks are brainless like dozen plastic <OOV> <OOV>	1	1	0.772143
user dozen hag jisoo should thank her fans for setting her up every hour because jisoo knows she is the worst dozen to exist yet her fans hate her so badly jisoo dozen cannot do anything she is just talentless and hag	1	1	0.744415
drag her all you want it are not damaging her image nor taking away her jobs at the end of everything your faves will still be irrelevant and talentless jisoo is happy in her home sleeping while <OOV> her dog	0	1	0.747227
to the quotes now if we called jenma an ugly talentless whore <OOV> a stripper rose a ceo <OOV> and jisoo a filler	1	1	0.802888
user yg s princess <OOV> jisoo is still talentless no wonder why they used to give her less lines even then the lead rapper	1	1	0.676956
user jisoo is a talentless hoe	1	1	0.816363
user <OOV> slut shaming jennie for this but you forgot jisoo s entire career is known for being talentless dozen you are favorite look like old grandma the audacity to compare <OOV> jennie with that grandma and talentless jisoo	1	1	0.799508

After applying semi-supervised techniques, the test set is reevaluated on the final model. Table 4 displays the text, the true label, the predicted label, and its corresponding probability score, which demonstrates significant improvement.

Table 4. Probability score after applying semi-supervised techniques

Text	True Label	Predicted Label	Probability
jisoo is a hag with <OOV> <OOV> face above all that she is a talentless <OOV> but you are worried about rs focus on mass reporting her fancams because she is getting dragged for her stiff ass	1	1	0.960702
user lisa fashion <OOV> she dresses like a man and grandma jisoo is talentless you all blinks are so disgusting	1	1	0.956055
user ya everyone <OOV> they are talentless girls and <OOV> girls and plastic faces they are not worth to be queens when so many talented girls group are in korea also a childish like you does not know how to <OOV> between what is right and wrong that is why you are still like this	1	1	0.95529
the whole world knows jisoo is dozen talentless hag keep delusional because blonks are brainless like dozen plastic <OOV> <OOV>	1	1	0.975697
user dozen hag jisoo should thank her fans for setting her up every hour because jisoo knows she is the worst dozen to exist yet her fans hate her so badly jisoo dozen cannot do anything she is just talentless and hag	1	1	0.96535
drag her all you want it are not damaging her image nor taking away her jobs at the end of everything your faves will still be irrelevant and talentless jisoo is happy in her home sleeping while <OOV> her dog	0	1	0.967388
to the quotes now if we called jenma an ugly talentless whore <OOV> a stripper rose a ceo <OOV> and jisoo a filler	1	1	0.969776
user yg s princess <OOV> jisoo is still talentless no wonder why they used to give her less lines even then the lead rapper	1	1	0.965162
user jisoo is a talentless hoe	1	1	0.977107
user <OOV> slut shaming jennie for this but you forgot jisoo s entire career is known for being talentless dozen you are favorite look like old grandma the audacity to compare <OOV> jennie with that grandma and talentless jisoo	1	1	0.963685
directors who work with jisoo keep showering her with <OOV> as an actress but koreaboo stan twitter want us to believe she is talentless	0	1	0.979382
user aespa are bunch of plastic hoes by the way	1	1	0.645674
user aespa already ended those eight talentless plastic hoes by the way <OOV>	1	1	0.941537
user and i lo the plastic whores of aespa	1	1	0.969955
user aespa are untalented plastic no <OOV> cannot dance and do not deserve to exist as a group <OOV> clapback on their fave with <OOV> user why should <OOV> defend aespa they are misogynist and hate women	1	1	0.709635
fuck you casual aespa listener a casual listener would not drag them for plastic surgery like antis all do you are just another lying ass aespa anti who says they love the music but hate the fans when they hate aespa too	0	0	0.383514
anyway plastic <OOV> aespa will never go on <OOV> nominee in those plastic girls life	1	0	0.336773
that is why slut plastic aespa will never be successful aespa will not last long because <OOV> <OOV> will soon be thrown back to her country bringing the same vicioc does not make aespa look better <OOV> they are just a bunch of lip <OOV> bitches every time lip <OOV> oh so useless	1	1	0.967017
and you all wonder why those plastic melting face bitches called aespa always get dragged sm stans and their obsession with hybe idols is insane	1	1	0.838814
aespa <OOV> no blush nude lips too much <OOV> while the last pic was done by different makeup artist for her <OOV> which <OOV> her features non <OOV> should shut the fuck up	1	1	0.713315

To evaluate the final model's performance, The Confusion Matrix, ROC Curve, and ROC-AUC Score were generated using Python's sklearn library. Figure 4 displays the confusion matrices before (left) and after (right) semi-supervised learning, offering a clear comparison of the model's ability to classify sexist and non-sexist content.

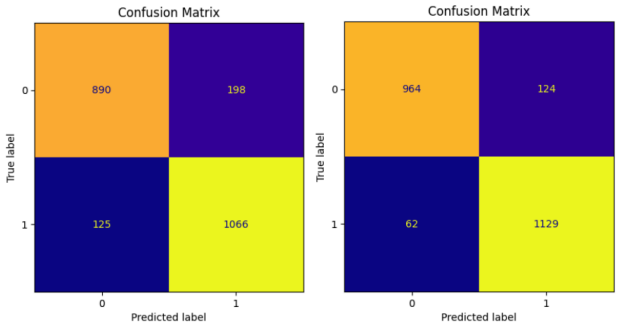

Figure 4. Side-by-Side Confusion Matrix of Baseline and Final Model

The baseline ROC curve (AUC = 92.41%) indicates strong initial performance but with room to reduce false positives. In contrast, the final model's ROC curve (AUC = 96.96%) is steeper and more convex, reflecting improved accuracy and robustness in classification. Figure 5 shows the side-by-side improvement of the ROC Curve and ROC-AUC Score.

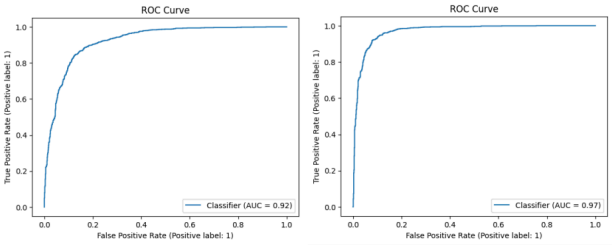

Figure 5. Side-by-Side ROC Curve of Baseline and Final Model

Using a test set of 2,279 samples, Table 5 shows that the baseline model achieved an accuracy of 85.83%, which increased to 91.84% after applying semi-supervised learning—a 6% improvement. Precision improved from 84.34% to 90.10%, indicating a decrease in false positives. Recall rose from 89.50% to 94.79%, showing better detection of actual sexist content. The F1 score also increased from 86.84% to 92.38%, confirming the final model's greater reliability and balance.

Table 5. Evaluation of the Baseline and Final Model

Evaluation Metric	Baseline Model	Final Model
Accuracy	85.83%	91.84%
Precision	84.34%	90.10%
Recall	89.50%	94.79%
F1 Score	86.84%	92.38%

The proposed model demonstrates highly effective performance compared to recent studies. As shown on Table 6, using data from the EXIST 2021 shared task, the top systems (AI-UPV and SINAI_TL) achieved around 76–77% accuracy and F1-score on the English dataset. In contrast, the baseline model in this study (using stacked embeddings with CNN + Attention) achieved 85.83% accuracy and 86.84% F1-score after augmenting the dataset with manually annotated K-pop tweets. This improvement is attributed to the inclusion of domain-specific data and tailored preprocessing.

This aligns with Singh and Anand (2021), who also improved their baseline Bi-LSTM model from 76.03% to 83.0% accuracy and from 82.8% to 82.8% F1-score using semi-supervised learning with sitcom dialogue data. While the final model in this study cannot be directly compared to EXIST 2021 systems due to data augmentation, it clearly shows the effectiveness of domain adaptation for sexism detection in K-pop-related social media content.

Table 6. Results comparison with existing studies

Model	Accuracy	F1-Score
EXIST 2021 *(task1_AI-UPV_1)*	76.68%	76.57%
EXIST 2021 *(task1_SINAI_TL_3)*	77.72%	77.47%
Singh & Anand (2021) *(Baseline Bi-LSTM)*	76.03%	-
Singh & Anand (2021) *(Final Bi-LSTM after SSL)*	83.03%	82.82%
Baseline Model *(with K-pop)*	85.83%	86.84%
Final Model *(after SSL)*	91.84%	92.38%

To implement the model, the researchers developed the K-pop Sexism Detector, a Google Extension that utilizes the final model to detect sexism and display its probability score in a Reddit thread.

The front-end interface, as shown in Figure 6 of the Google extension, was designed with usability in mind. It adopted a tab-based layout, allowing users to easily scroll through and view each Reddit post and comment. For every piece of text displayed, the interface included the original content, the predicted label, and a probability score indicating the model's confidence in its prediction. This enables researchers to evaluate the classification outcome and the confidence in the model's decision.

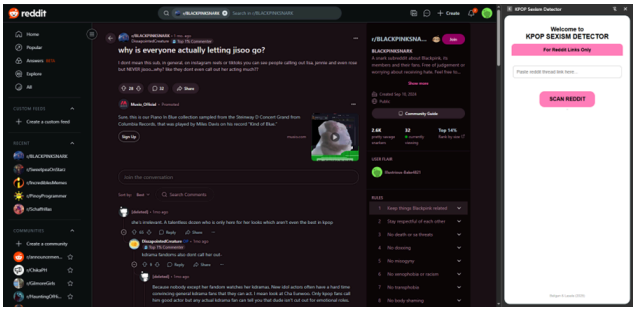

Figure 6. Front End of Protoype

To test the functionality of the Google Extension, the researchers began by searching for a suitable Reddit thread that could provide relevant data on sexism. Within Reddit, they chose a thread from the subreddit BLACKPINKSNARK, which is notorious for critical comments about the K-pop girl group BLACKPINK, making it a potentially rich source of sexist commentary. After finding a relevant thread, the researchers copied its URL and pasted it into a Google Chrome plugin built explicitly for natural language processing activities. This

automatically extracted the content of the Reddit post and comments, scanning each text section for sexist language.

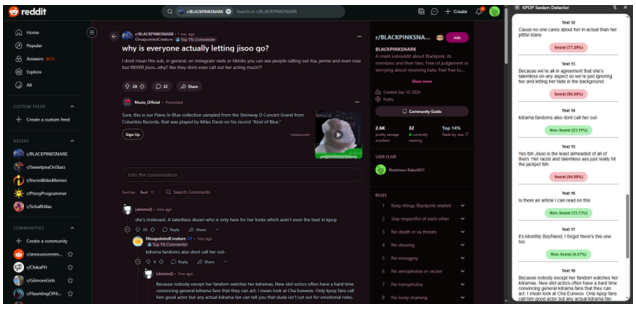

Figure 7. Retrieval and Detection of K-pop Sexism Detector

The K-pop Sexism Detector extension then displays each comment with a classification label (e.g., sexist or non-sexist) and its probability score, indicating the model's confidence in its evaluation, as shown in Figure 7. Those highlighted in red have sexist predictions, while those highlighted in green have non-sexist predictions. This technique enables the researchers to collect and evaluate user-generated content from the K-pop fan community in an automated and replicable manner.

Error Analysis

A key issue identified was the limited scope of the training data. Since the labeled and pseudo-labeled K-pop tweets were manually scraped, they only reflected a narrow range of sexist discourse. The model performed well on familiar controversies (e.g., BLACKPINK derogatory comments, the Garam bullying scandal, and Burning Sun) but struggled with new or subtle forms of sexism not seen during training, resulting in false negatives. This highlights the model's difficulty in generalizing to emerging topics on K-pop Twitter, particularly without continuous domain adaptation, which can cause performance degradation over time due to temporal bias.

A second limitation was the presence of out-of-vocabulary (OOV) terms. Expressions like “wah” and Korean words such as “aegyosal” and “nugu” were not well-represented in the embeddings. Although stacked embeddings improved coverage, they still lacked depth in capturing Korean slang and bilingual code-switching. Creating custom Korean slang embeddings could address this, but it was beyond the scope of the study.

Table 7. Misclassified K-pop Tweets

Text	True Label	Predicted Label	Probability
we should bully garam to teach her a lesson so you are admitting you are also a <OOV> bully	1	0	0.09
anyway plastic <OOV> aespa will never go on <OOV> nominee in those plastic girls life	1	0	0.34
user that is because the bribery made by his girlfriend lisman has no talent so there is nothing to be proud of from a that girl his girlfriend gives <OOV> so that people want to <OOV> with lisman who cannot sing	1	0	0.33
lisman s mom probably does since he does not have a big a forehead unlike her daughter whom she told to cover up she knew the disaster if she did not	1	0	0.32
blackpink is signed to the worst korean and western labels both of them hate women	0	1	0.85
her rich <OOV> boyfriend bought it	0	1	0.87
jennie willingly <OOV> in a sexist fantasy show and them four dozens constantly insult women in their shitty music a man writes for them be serious	0	1	0.94
directors who work with jisoo keep showering her with <OOV> as an actress but koreaboo stan twitter want us to believe she is talentless	0	1	0.98
soojin was literally bullied left and right her case was treated worse than rapists by <OOV> her situation still being used as a drag by haters and being made fun of by the way shut your stupid ass about my before anyone else	0	1	0.94
you all laughed about <OOV> suicide watch the misogyny and <OOV> she faced and the overall situation which left her deeply traumatized do not lie	0	1	0.95
drag her all you want it are not damaging her image nor taking away her jobs at the end of everything your faves will still be irrelevant and talentless jisoo is happy in her home sleeping while <OOV> her dog	0	1	0.97

A third issue involved the model’s reliance on surface-level lexical patterns rather than a deeper understanding of semantics. As shown in Table 7, the model occasionally provided high-confidence predictions for incorrect labels, indicating an overreliance on explicit or repetitive features. This stemmed from both the limited diversity of training data and the embeddings’ inability to capture semantic nuances. While the model performed well on explicit sexism, it struggled with implicit cases resembling toxic language.

VI. CONCLUSION

This study successfully explored automated sexism detection in K-pop-related tweets by addressing key challenges specific to this domain, including noisy text, limited labeled data, and evolving community discourse. This study contributed a domain-specific solution to the growing demand for automated sexism detection. The application of eleven tailored preprocessing techniques standardized and enhanced the quality of both labeled and unlabeled datasets, enabling more effective feature learning and improving the model's performance in handling informal, social media-style text.

The development of a hybrid deep learning model (stacked GloVe and fastText embeddings, CNN, and attention mechanisms), combined with a semi-supervised self-training approach, significantly improved the model’s capacity to detect sexist content. Integrating domain-specific, manually labeled K-pop data into the training process enabled the model to generalize better to niche, context-rich language. Although the model performed well across multiple evaluation metrics, error analysis revealed limitations in capturing implicit, culturally embedded, or slang-heavy expressions in K-pop. This underscores the need for more semantically aware architectures and the inclusion of richer, multilingual data in future work.

The deployment of the final model in a functional Chrome extension demonstrated its practical application,

enabling real-time, automated sexism detection in Reddit threads. This prototype validated the model's usability and relevance in real-world K-pop community contexts. By extracting and classifying Reddit comments in real-time, the extension provided users with an accessible and interpretable tool for identifying sexist content. This practical application validates the model's utility beyond academic settings and opens pathways for further development in content moderation tools.

VII. RECOMMENDATIONS

Extend Slang and Emoji Normalization. Continue update the slang vocabulary and add emoji normalization to reflect newly emerging K-pop fan slang and abbreviations. Extending the preprocessing techniques will enable future work to take advantage of sentiment or implicit meanings in social media posts.

Integrate Semantically Aware Models. Transformer-based embeddings and models are proven to be more advanced in text classification as it is trained on contextual features, which can improve domain adaptation and sexism detection. Due to hardware constraints, the researchers weren't able to experiment with Transformers.

Adopt word embeddings on K-pop discourse. To address the key limitation of this study, future research should incorporate fine-tuned embeddings based on Korean or K-pop-specific corpora. With this, it will minimize the multiple OOVs present during he study and support multilingual text and enhance model generalization and word embedding representations.

Explore other social media platforms. Use Facebook or Reddit as other sources of unlabeled data for K-pop sexism detection due to the heightened restrictions on Twitter. Likewise, 10,782 unlabeled data are insufficient for training deeper models like Bi-LSTM and Transformer-based models like BERT.

ACKNOWLEDGEMENT

From the bottom of their hearts, the researchers would like to express their sincere gratitude. The completion of this thesis has been a journey marked by both challenges and personal growth. It would not have been possible without the effort, dedication, and time of the researchers, as well as those who have helped make this study possible:

To the adviser, Dr. Marc P. Laureta, for his patience, valuable constructive criticism, and belief in their capabilities. His guidance challenged and brought clarity to the development of this work.

To their BSCS classmates, for the shared encouragement and support, which reminded them that they were not alone on this path.

To their beloved parents and families, for their unconditional love, sacrifices, and unwavering support, which served as the foundation that made this accomplishment possible, and for their presence and motivation during moments of doubt and exhaustion.

Most of all, to the Almighty God for His constant Grace that brought them peace through the most challenging days. He gave them strength and wisdom throughout this study and its successful completion.

REFERENCES

Abburi, H., Parikh, P., Chhaya, N., et al. (2021). Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. Data Science and Engineering, 6, 359-379. https://doi.org/10.1007/s41019-021-00168-y

Abney, S. (2007). Semisupervised Learning for Computational Linguistics (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420010800

Agarwal, S., Sonawane, A., & Chowdary, C. R. (2023). Accelerating automatic hate speech detection using parallelized ensemble learning models. Expert Systems with Applications, 230, 120564.

Ali, A. A., Latif, S., Ghauri, S. A., Song, O., Abbasi, A. A., & Malik, A. J. (2023). Linguistic features and Bi-LSTM for identification of fake news. Electronics, 12(13), 2942. https://doi.org/10.3390/electronics12132942

Al Bataineh, A., Reyes, V., Olukanni, T., Khalaf, M., Vibho, A., & Pedyuk, R. (2023). Advanced Misinformation Detection: A Bi-LSTM Model Optimized by Genetic Algorithms. Electronics, 12(15), 3250. https://doi.org/10.3390/electronics12153250

Alsafari, S., & Sadaoui, S. (2021). Semi-supervised self-training of hate and offensive speech from social media. Applied Artificial Intelligence, 35(15), 1621-1645.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. Proceedings of the 26th International Conference on World Wide Web Companion. DOI: 10.1007/s00530-017-0552-0

Badri, N., Kboubi, F., & Chaibi, A. H. (2022). Combining FastText and Glove word embedding for offensive and hate speech text detection. Procedia Computer Science, 207, 769–778.

Belbachir, F., Roustan, T., & Soukane, A. (2024). Detecting Online Sexism: Integrating Sentiment Analysis with Contextual Language Models. AI, 5(4), 2852-2863.

Bergmann, D., & Stryker, C. (2024, December 4). What is an attention mechanism? IBM. https://www.ibm.com/think/topics/attention-mechanism

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.

Bergmann, D. (2023). What is Semi-Supervised Learning? | IBM. https://www.ibm.com/topics/semi-supervised-learning

Dai, H. and Xu, X. (2014). Sexism in News: A Comparative Study on the Portray of Female and Male Politicians in The New York Times. Open Journal of Modern Linguistics, 4, 709-719. doi: 10.4236/ojml.2014.45061.

Duarte, J. M., & Berton, L. (2023). A review of semi-supervised learning for text classification. Artificial Intelligence Review, 56(2). https://doi.org/10.1007/s10462-023-10393-8

Effrosynidis, D., Symeonidis, S., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems With Applications, 110, 298–310. https://doi.org/10.1016/j.eswa.2018.06.022

GeeksforGeeks. (2023, June 8). Bidirectional LSTM in NLP. GeeksforGeeks. https://www.geeksforgeeks.org/bidirectional-lstm-in-nlp/

GeeksforGeeks. (2025, February 7). Understanding TFIDF (Term FrequencyInverse Document Frequency). GeeksforGeeks. https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

Georgios, K. P., Ramampiaro, H., & Langseth, H. (2018). Detecting offensive language in tweets using deep learning. arXiv preprint arXiv:1801.04433, 1-17.

Girl Groups and Gender Gaps: The Globalization of Gender Stereotypes through K-POP. (2021, April 11). NICKELED AND DIMED. https://nickledanddimed.com/2021/04/11/girl-groups-and-gender-gaps-the-globalization-of-gender-stereotypes-through-K-POP/

Gupta, A., Nayyar, A., Arora, S., Jain, R. (2021). Detection and Classification of Toxic Comments by Using LSTM and Bi-LSTM Approach. In: Luhach, A.K., Jat, D.S., Bin Ghazali, K.H., Gao, XZ., Lingras, P. (eds) Advanced Informatics for Computing Research. ICAICR 2020. Communications in Computer and Information Science, vol 1393. Springer, Singapore. https://doi.org/10.1007/978-981-16-3660-8_10

Hamborg, F. (2023). Revealing Media Bias in News Articles: NLP Techniques for Automated Frame Analysis (p. 238). Springer Nature. https://library.oapen.org/viewer/web/viewer.html?file=/bitstream/handle/20.500.12657/61876/978-3-031-17693-7.pdf?sequence=1&isAllowed=y

Hamborg, F., Donnay, K., & Gipp, B. (2019). Automated identification of media bias in news articles: an interdisciplinary literature review. International Journal on Digital Libraries, 20(4), 391–415. https://doi.org/10.1007/s00799-018-0261-y

Ibm. (2025, April 17). Semantic layer. What is semantic layer? Retrieved May 6, 2025, from https://www.ibm.com/think/topics/semantic-layer

Ingrid, C. (2024, March 11). Gender inequality and double standards in the Korean pop industry. Medium.

https://medium.com/@ingridientss/gender-inequality-and-double-standards-in-the-korean-pop-industry-def5dc2f4675

IvyPanda. (2022, June 21). Gender Bias in K-POP: Gender Bias in Korean Society. https://ivypanda.com/essays/gender-bias-in-K-POP-gender-bias-in-korean-society/

Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. https://www.semanticscholar.org/paper/When-does-a-compliment-become-sexist-Analysis-and-Jha-Mamidi/161dfc76ace91bedf2770b3de5c6aac882453749

Joshi, S. (2022, November 11). Sexism in K-POP. The High. https://thehighisl.com/1616/opinion/sexism-in-K-POP/

Kalra, A., & Zubiaga, A. (2021). Sexism Identification in Tweets and Gabs using Deep Neural Networks. Papers With Code. Retrieved from https://paperswithcode.com/author/amikul-kalra

Karisani, P., & Karisani, N. (2021). Semi-supervised text classification via self-pretraining. In Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21). https://doi.org/10.1145/3437963.3441814

Kim, S. W., Lee, Y. G., Tama, B. A., & Lee, S. (2020). Reliability-Enhanced camera lens module classification using Semi-Supervised Regression Method. Applied Sciences, 10(11), 3832. https://doi.org/10.3390/app10113832

Kirk, H. R., Yin, W., Vidgen, B., & Röttger, P. (2023). Semeval-2023 task 10: Explainable detection of online sexism. arXiv preprint arXiv:2303.04222.

Kumar, A., Kumar, S., Passi, K., & Mahanti, A. (2024). A Hybrid Deep BiLSTM-CNN for Hate Speech Detection in Multi-social media. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(8), 1-22.

Li, C., Li, X., & Ouyang, J. (2021). Semi-Supervised Text Classification with Balanced Deep Representation Distributions. Annual Meeting of the Association for Computational Linguistics.

Lin, X., & Rudolf, R. (2017). Does K-POP Reinforce Gender Inequalities? Empirical Evidence from a New Data Set. Asian Women/Asian Women, 33(4), 27–54. https://doi.org/10.14431/aw.2017.12.33.4.27

Mavaie, P., Holder, L., & Skinner, M. K. (2023). Hybrid deep learning approach to improve classification of low-volume high-dimensional data. BMC bioinformatics, 24(1), 419.

Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2022). ETHOS: a multi-label hate speech detection dataset. Complex & Intelligent Systems, 8(6), 4663-4678.

Morinière, P. (2016). No to sexism hate speech. Council of Europe. Retrieved from https://rm.coe.int/no-to-sexism-hate-speech/16805a315d

Nugraha, R., & Komsiah, S. (2023). Utilization Of New Media As Digital Fandom Among Korean Pop (K-POP) Fan Groups On The Social Media Platform Twitter. International Journal of Progressive Sciences and Technologies, 40(1), 200-207. doi:http://dx.doi.org/10.52155/ijpsat.v40.1.5584

Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., & Goodfellow, I. J. (2018). Realistic evaluation of deep semi-supervised learning algorithms. arXiv preprint arXiv:1804.09170. Retrieved from https://doi.org/10.48550/arXiv.1804.09170

Patil, S., Rodrigues, A., Telangi, R., & Chavan, V. (2022). A review on text classification based on cnn. International Journal of Scientific Research in Science and Technology, 622-624.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Rallabandi, S., Singhal, S., & Seth, P. (2023). SSS at SemEval-2023 Task 10: Explainable Detection of Online Sexism using Majority Voted Fine-Tuned Transformers. arXiv preprint arXiv:2304.03518.

Rao, P., & Taboada, M. (2021). Gender Bias in the News: A scalable topic modelling and visualization framework. Frontiers in Artificial Intelligence, 4. https://doi.org/10.3389/frai.2021.664737

Rika, A. (2024). DOUBLE STANDARDS IN HATE COMMENTS AGAINST K-POP ARTISTS: PRAGMATICS STUDY. ELP (Journal of English Language Pedagogy), 9(1), 1-15.

Rodrigo-Ginés, F. J., Carrillo-de-Albornoz, J., & Plaza, L. (2021). UNEDBiasTeam at IberLEF 2021's EXIST Task: Detecting Sexism Using Bias Techniques. In IberLEF@ SEPLN (pp. 522-532).

Rodríguez-Sánchez, F., Carrillo-De-Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., & Donoso, T. (2021, September 6). EXIST: sEXism Identification in Social neTworks. https://nlp.uned.es/exist2021/

Rodríguez-Sánchez, F., Carrillo-De-Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., & Donoso, T. (2021, September 6). Overview of EXIST 2021: SEXism Identification in Social NETworks. Rodríguez-Sánchez | Procesamiento Del Lenguaje Natural. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389

Romano, A. (2018, February 26). How K-POP became a global phenomenon. Vox. https://www.vox.com/culture/2018/2/16/16915672/what-is-K-pop-history-explained

Sainez, A., & Wu, J. J. (2022, May). K-pop Sentiment Analysis. GitHub. https://github.com/tsainez/K-pop-sentiment-analysis.git

Saraswati, L. A., & Nurbaity. (2020). BTS ARMY's #BTSLOVEYOURSELF: a worldwide K-POP fandom participatory culture on Twitter. KnE Social Sciences. https://doi.org/10.18502/kss.v4i14.7899

Sharifirad, S., & Matwin, S. (2019, February 27). When a Tweet is Actually Sexist. A more Comprehensive Classification of Different Online Harassment Categories and The Challenges in NLP. arXiv.org. https://arxiv.org/abs/1902.10584

Shulman, H., & Simo, H. (2021). Poster: WallGuard - A Deep Learning Approach for Avoiding Regrettable Posts in Social Media. In 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS) (pp. 1142-1143). Washington, DC, USA: IEEE. https://doi.org/10.1109/ICDCS51616.2021.00127

Singh, S., Anand, T., Chowdhury, A. G., & Waseem, Z. (2021). "Hold on honey, men at work": A semi-supervised approach to detecting sexism in sitcoms. In 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 (pp. 180-185). Online: Association for Computational Linguistics.

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine Learning, 109, 373-440. https://doi.org/10.1007/s10994-019-05855-6

Villines, Z. (2021, June 30). What is sexism? https://www.medicalnewstoday.com/articles/what-is-sexism

Vision, D. (2024, February 12). What is K-POP. Dance Vision. https://blog.dancevision.com/what-is-K-pop

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop (pp. 88-93). San Diego, California: Association for Computational Linguistics.

What is sexism? (2024, June 19). European Institute for Gender Equality. https://eige.europa.eu/publications-resources/toolkits-guides/sexism-at-work-handbook/part-1-understand/what-sexism?language_content_entity=en

Word embeddings. (n.d.). TensorFlow. https://www.tensorflow.org/text/guide/word_embeddings

Zahra, A. A. (2024). Misogynistic in the K-Pop Industry: Analyzing Gender Bias Towards Female Idols. Serat: Journal of Literature & Cultural Studies, 1(2), 49-56.