ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep393

Detection of Adult Content in Arabic Tweets Using Machine Learning Models

Aram Ibrahim Al-Anazi[1]

[1]Information Management Specialist

Publication Date: 2025/09/20

Abstract: This study evaluates the effectiveness of various machine learning and deep learning models in detecting adult content in Arabic tweets, addressing unique linguistic and cultural challenges. Using a dataset of 33,691 Arabic tweets, we implemented and compared Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), and AraBERT. The data underwent thorough preprocessing, including cleaning, tokenization, and segmentation into training, validation, and test sets. Performance metrics such as accuracy, F1 score, and confusion matrices were used to assess model efficacy. AraBERT achieved the highest accuracy (100%), demonstrating superior capability in capturing spatial patterns for content classification. CNN and RNN also performed well, with accuracies of 94.27% and 94.22%, respectively, while LSTM achieved an accuracy of 88.37%. These findings highlight AraBERT's potential for effective content moderation in Arabic digital spaces, contributing to safer online environments.

Keywords: *Arabic Tweets; AraBERT; Convolutional Neural Networks (CNN); Long Short-Term Memory (LSTM); Natural Language Processing (NLP); Recurrent Neural Networks (RNN); Text Classification.*

How to Cite: Aram Ibrahim Al-Anazi (2025) Detection of Adult Content in Arabic Tweets Using Machine Learning Models. *International Journal of Innovative Science and Research Technology*, 10(9), 1060-1065. https://doi.org/10.38124/ijisrt/25sep393

I. INTRODUCTION

Adult content detection has become an increasingly critical issue for online platforms due to the rapid expansion of data exchanged across digital spaces. Social media platforms, websites, and other online environments handle vast amounts of information every second, resulting in a dynamic yet demanding content moderation ecosystem. The availability of explicit information online, often without adequate age controls or permission procedures, raises substantial concerns, particularly for vulnerable users such as children and teenagers. The risks of exposure to inappropriate content extend beyond immediate harm, potentially affecting psychological development, behavior, and societal norms [1]. Traditional approaches for identifying adult content have evolved significantly, from manual moderation to automated systems powered by machine learning (ML) and deep learning (DL) technologies. Early methods, such as keyword filtering and image-based detection, were limited in their ability to handle the complexities of modern content. Advances in ML and DL have led to the development of sophisticated algorithms capable of analyzing large datasets, recognizing subtle patterns, and providing real-time results [2]. Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and transformer-based architectures like BERT and its variants have shown promising results in text and image classification tasks, including the detection of explicit content [3], [4].

Despite these advancements, there remains a significant gap in research focused on detecting adult content in the Arabic digital realm compared to other languages. Arabic presents unique linguistic challenges, including its rich morphology, diverse dialects, and frequent use of colloquialisms, which complicate automated content moderation [5]. Additionally, the inclusion of emoticons, URLs, and hashtags in Arabic tweets further complicates text processing, necessitating more advanced and adaptive algorithms [6].

This study aims to address these challenges by evaluating the effectiveness of various ML and DL models in detecting adult content in Arabic tweets. The study investigates models including CNN, RNN, LSTM, and AraBERT to determine the most effective methods for achieving this aim. The results reveal the capabilities and limitations of current systems and highlight the need for continued innovation in content moderation. Finally, our study contributes to developing safer digital environments, especially for Arabic- speaking users, by improving our understanding of how to efficiently and reliably filter explicit information in text content.

II. RELATED WORK

The dataset is organized into several categories to capture the diverse linguistic and cultural aspects of Arabic social media content. It includes tweets in various Arabic

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep393

dialects, such as Gulf, Levantine, and Egyptian, showcasing the linguistic diversity of the Arabic-speaking world. The dataset also features informal text styles common on social media, including emojis, hashtags, and abbreviations. Additionally, it contains sensitive language, including vulgar words and satirical and metaphorical references to adult entertainment, reflecting the nuanced ways in which adult content is expressed in Arabic. Contextual information, such as hashtags, mentions, and links, is retained to enhance the accuracy of natural language processing models in identifying semantic and syntactic patterns. Several studies have focused on the automatic identification of adult content, whether in image-based, video-based, or text-based formats.

Appati et al. (2021) provided a comprehensive review of image analysis techniques for adult content detection, focusing on both traditional and deep learning methods [1]. They categorized the approaches to Region of Interest (ROI) techniques and deep learning methods. ROI techniques, such as skin pixel ratio and explicit content weighting, utilize algorithms like Support Vector Machines (SVM) and K Nearest Neighbors (KNN) for classification. In contrast, deep learning methods, particularly Convolutional Neural Networks (CNNs), have shown superior performance in detecting explicit content due to their ability to learn complex patterns from large datasets.

Gajula et al. (2020) proposed a model based on supervised learning using the Support Vector Machine (SVM) algorithm to detect and blur explicit content in images. The model was trained on a dataset containing 7,000 pornographic images and 7,000 normal images, achieving an accuracy of 97.8% during testing. The classification process was based on the amount of skin percentage exposed in the images. If an image was classified as pornographic, it was then processed to blur the explicit content using image processing techniques. This approach ensured that end-users were not exposed to inappropriate material, making it a robust solution for protecting children from adult content on the internet [7].

Dubettier et al. (2023) conducted a comparative study to evaluate the efficacy of various methods for detecting sexual content in images, aiming to protect children and enhance digital forensic investigations. The study assessed five tools: the nsfw model, NudeNet, NuDetective, SkinDetection, and DeepPornDetection, each employing distinct techniques such as skin detection, deep learning, or transfer learning. Using three datasets with varying degrees of explicit content and complexity, the researchers found that the nsfw model and NudeNet achieved the highest accuracy across datasets, while DeepPornDetection performed best on the dataset it was trained on, indicating a training bias. The study highlighted several challenges, including the insufficiency of skin detection alone, as high skin exposure does not always indicate sexual content, and some explicit images have low skin exposure. Additionally, the findings underscored the subjectivity in distinguishing between acceptable and harmful content due to cultural and environmental factors. While the nsfw model and NudeNet showed potential, the authors concluded that further enhancements are needed. They recommended future research to explore adaptive criteria based on cultural variables for more precise screening [4]. Ochoa et al. (2012) explored machine learning-based strategies for recognizing pornographic video material, emphasizing the integration of spatial and temporal data [2]. Their study demonstrated the effectiveness of combining spatial features, such as skin detection and color histograms, with temporal features like shot duration and camera motion. The use of SVM classifiers achieved an accuracy rate of up to 94.44%, highlighting the potential of deep learning techniques in this domain.

Barrientos et al. (2020) conducted a comprehensive study on the use of machine learning techniques to detect inappropriate erotic content in text, with a focus on protecting children from exposure to such material. The study highlighted the growing need for automated moderation tools due to the impracticality of manual moderation in the face of increasing user-generated content. The researchers employed twelve different models, including three text encoders (Bag of Words, TF-IDF, and Word2Vec) and four classifiers (SVM, Logistic Regression, k- Nearest Neighbors, and Random Forest), to identify unsuitable material. Using a dataset of over 110,000-word samples from Reddit, categorized as sexual or neutral, they found that the combination of TF-IDF and SVM (linear kernel) achieved the highest performance, with an accuracy of 97% and an F-score of 0.96. This study underscored the effectiveness of machine learning approaches in real-time content filtering, particularly for social networks. The authors also suggested future research directions, including the exploration of deep learning models and feature reduction techniques to further enhance content detection systems. The findings demonstrated the practicality and potential of automated tools in maintaining a safer online environment for children [3].

Hamdy et al. (2021) conducted a study focused on identifying explicit content in Arabic tweets. The researchers created a dataset comprising 50,000 Twitter accounts, with 6,000 identified as adult content accounts. This dataset was meticulously annotated using Arabic-related keywords and hashtags. The analysis revealed that adult tweets are generally shorter, use fewer words, and contain more URLs and emojis compared to non-adult tweets. Significant patterns were identified in the use of words, emojis, and hashtags. The study evaluated several machine learning models, including Support Vector Machines (SVM), FastText, multilingual BERT, and AraBERT. Among these, AraBERT outperformed the others, achieving an F1 score of 96.8% when combining user data with tweet content. The researchers concluded that even basic information, such as usernames and brief descriptions, can be effective in identifying sexual content accounts. They suggested that future research could explore multimodal content analysis to further enhance detection accuracy [8]. These studies provided a foundation for our research, highlighting the strengths and limitations of existing approaches. Our study aims to build on this foundation by evaluating the performance of various machine learning and deep learning models in detecting adult content in Arabic tweets, addressing the identified gaps in the literature (Table 1).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep393

Table 1 Summary of the Key Findings from the Related Work, Comparing the Performance of Different Models and Techniques in Detecting Adult Content

Study	Approach	Dataset	Accuracy	Key Findings
Appati et al.	ROI Technique s, CNNs	Image Data	94.44%	CNNs out perform Traditional methods
Ochoa et al.	SVM, Spatial and Temporal Features	Video Data	94.44%	Integration of spatial and temporal data improves accuracy
Gajula et al.	SVM, Skin Percentage	Image Data	97.8%	High accuracy in detecting and blurring explicit content
Barrientos et al.	TF-IDF, SVM	Text Data	97%	Effective real-time content filtering for social networks
Dubettier et al.	nsfw model, NudeNet	Image Data	High accuracy	Challenge s in skin detection and cultural subjectivity
Hamdy et al.	AraBERT	Arabic Tweets	96.8%	Effective in identifying explicit content in Arabic tweets

III. DATASET

In this study, we used the dataset from the previous study by Hamdy et al. (2021) [8], we evaluated adult content on Arabic Twitter. The dataset consists of 33,691 samples with two main columns, named "text2" and "categories." The "text2" column contains the full tweet text or main textual content, which includes various linguistic structures ranging from Fusha (formal Arabic) to Ammiya (colloquial/spoken Arabic) and code- switching between Arabic and English. The "categories" column indicates whether the tweet contains adult content or not, with two classes: ADULT (1) and NOT_ADULT (0).

The dataset was meticulously collected and annotated by expert linguists and content reviewers to ensure high quality and reliability. They manually evaluated and annotated tweets using Arabic-related keywords and hashtags to accurately identify and classify adult content.

The dataset encompasses a wide range of linguistic, artistic, and expressive features, including tweets in Gulf, Levantine, and Egyptian Arabic dialects, reflecting the linguistic diversity of the Arabic-speaking world. It also includes informal text styles common on social media, such as emojis, hashtags, and abbreviations. Additionally, the dataset contains vulgar words and satirical and metaphorical references to adult entertainment, capturing the nuanced ways in which adult content is expressed in Arabic.

Contextual information typically found in social media content, such as hashtags, mentions, and links, is retained in the dataset. This contextual data is crucial for advanced natural language processing (NLP) models, as it helps identify semantic and syntactic patterns that improve the accuracy of adult content classification.

➢ *Data Preprocessing*

To prepare the dataset for model training, several preprocessing steps were undertaken (Figure 1):

- Data Cleaning: Duplicate tweets and language flaws that could skew the study findings were removed.
- Tokenization: The text input was tokenized using a tokenizer to convert the tweets into a format suitable for model training.
- Label Encoding: The labels were converted to numerical values using LabelEncoder to facilitate model training and evaluation.
- Data Splitting: The dataset was split into training, validation, and test sets using a 70/30 ratio, ensuring a robust evaluation of the models.
- Dataset Size and Diversity The dataset's large size and diversity enable robust machine-learning training for the automatic identification and classification of adult content on Arabic social media sites. This wealth of cultural and expressive dimensions helps in crafting strong, generalizable models that are well-suited to modeling the complexity and nuances of Arabic text.
- Generalizability With such diverse and realistic data, the models trained on this dataset can generalize well across various scenarios, making this dataset an essential part of the solution for building tools to monitor and filter out adult content in Arabic language social media websites. The inclusion of contextual information and the representation of different dialects and informal text styles further enhance the dataset's utility for developing effective content moderation systems.

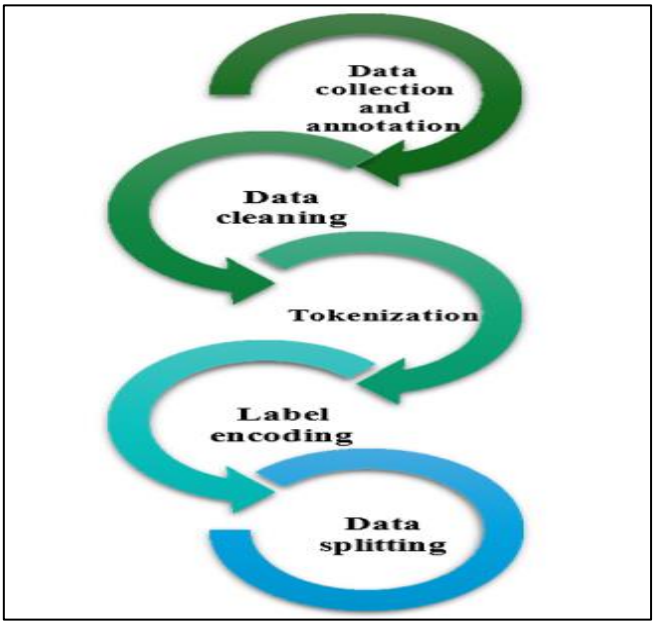

Fig 1 Data Processing Steps

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep393

IV. METHODOLOGY

In this study, we utilized a dataset of 33,691 tweets classified as adult or non-adult material. The methodology involved several key processes to ensure the accuracy and reliability of the results:

➢ *Data Preprocessing*

- Data Cleaning: The dataset was meticulously cleaned to remove duplicate tweets and correct language flaws that could skew the study findings. This step ensured that the data was of high quality and free from inconsistencies.
- Tokenization: The text input was tokenized using a tokenizer, which converted the tweets into a format suitable for model training. This process involved breaking down the text into individual tokens or words, which could then be analyzed by the models.
- Label Encoding: Labels were converted to numerical values using LabelEncoder. This step facilitated the training and evaluation of the models by providing a standardized format for the classification labels.
- Data Splitting: The dataset was split into training, validation, and test sets using a 70/30 ratio. This approach ensured that the models were trained on a substantial portion of the data while retaining enough data for validation and testing to assess model performance accurately.

➢ *Models used*

We employed several machine learning and deep learning models to evaluate their effectiveness in detecting adult content in Arabic tweets:

- Convolutional Neural Network (CNN): The CNN model consisted of embedding layers, convolutional layers (Conv1D), and pooling layers. This architecture enabled the capture of spatial patterns in the text, making it suitable for text classification tasks.
- Long Short-Term Memory Network (LSTM): The LSTM model was designed to handle sequential dependencies in the text. It included layers for long-term memory, allowing it to capture contextual meaning across long sequences, which is particularly beneficial for lengthier sentences.
- Recurrent Neural Network (RNN): A simple RNN was used to analyze text sequentially using tanh activation functions. This model was appropriate for short to medium-length texts. For binary classification, we employed a dense output layer with sigmoid activation.
- AraBERT Model: Specifically designed for Arabic, the AraBERT model was trained on a large corpus of Arabic texts and optimized to handle the unique features of the language, including its morphology and syntax. Natural language processing tasks were performed using the pretrained aubmindlab/bert-base- arabertv02 model, set up with Trainer to have a low learning rate and a small batch size. Few-shot learning was also investigated by evaluating performance with a small collection of examples.

➢ *Model Training and Evaluation*

- Training: Each model was trained on the training set, with hyperparameters tuned to optimize performance. The training process involved adjusting the model parameters to minimize the loss function and improve classification accuracy.
- Validation: The validation set was used to monitor the models' performance during training and prevent overfitting. Techniques such as dropout and batch normalization were employed to enhance model generalization.
- Testing: The final evaluation of the models was conducted on the test set. Performance metrics such as accuracy, F1 score, and confusion matrices were used to assess how successfully each model identified explicit content from safe material.

➢ *Evaluation Metrics*

- Accuracy: The proportion of correctly classified tweets out of the total number of tweets.
- F1 Score: The harmonic means of precision and recall provide a balanced measure of model performance.
- Confusion Matrices: Graphical representations of the true positives, false positives, true negatives, and false negatives, allow for a detailed comparison of model performance across categories.

Table 2 A Structured Overview of the Methodology Steps, Model used, and Evaluation Metrics

Step	Description
Data Preprocessing	
Data Cleaning	Removing duplicate tweets and correcting language flaws.
Tokenization	Converting text input into tokens suitable for model training.
Label Encoding	Converting labels to numerical values.
Data Splitting	Splitting the dataset into training (70%), validation, and test sets (30%).
Models Used	
Convolutional Neural Network (CNN)	Embedding layers, Convolutional layers (Conv1D), and Pooling layers to capture spatial patterns.
Long Short-Term Memory Network (LSTM)	Handling sequential dependencies with layers for long-term memory.
Recurrent Neural Network (RNN)	Analyzing text sequentially using tanh activation functions.
AraBERT	Pretrained on a large corpus of Arabic texts, optimized for Arabic morphology and syntax.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep393

Step	Description
Model Training and Evaluation	
Training	Adjusting model parameters to minimize loss and improve accuracy.
Validation	Monitoring performance during training to prevent overfitting.
Testing	Final evaluation using accuracy, F1 score, and confusion matrices.
Evaluation Metrics	
Accuracy	The proportion of correctly classified tweets.
F1 Score	Harmonic means of precision and recall.
Confusion Matrices	Graphical representation of true positives, false positives, true negatives, and false negatives.

V. RESULTS

The findings of this study demonstrated the varying efficacy of different machine learning and deep learning models in classifying adult content in Arabic tweets. The models evaluated include AraBERT, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory networks (LSTM). Each model's performance was assessed using metrics such as accuracy and F1 score, and confusion matrices were constructed to provide a detailed comparison.

Results revealed that AraBERT emerged as the most successful model, achieving an impressive accuracy of 100%. This model's ability to efficiently capture spatial patterns in the text makes it highly effective for content classification. The high accuracy indicated that AraBERT can reliably distinguish between adult and non-adult content in Arabic tweets, making it a valuable tool for content moderation.

In addition, the CNN model also performed well, with an accuracy of 94.27%. This model's architecture, which includes embedding layers, convolutional layers (Conv1D), and pooling layers, enables it to capture spatial patterns in the text effectively. The high accuracy of the CNN model suggests that it was well-suited for handling short text sequences and can be a robust option for detecting adult content in Arabic tweets.

Moreover, the RNN model achieved an accuracy of 94.22%, compared to the CNN model. The RNN's ability to analyze text sequentially using tanh activation functions makes it suitable for short to medium-length texts. The model's performance indicates that it can effectively handle brief text sequences and classify them accurately.

The LSTM model, designed to handle sequential dependencies in the text, achieved an accuracy of 88.37%. While this model is good at capturing contextual meaning across long sequences, its performance was somewhat lower than that of the CNN and RNN models. The LSTM's ability to manage temporal dependencies makes it beneficial for lengthier sentences, but it may be less effective for shorter text sequences compared to the other models.

The results highlighted the strengths and limitations of each model in classifying adult content in Arabic tweets. AraBERT's superior performance can be attributed to its design, which was specifically optimized for the Arabic language, including its morphology and syntax. The CNN and RNN models also demonstrated strong performance, indicating their suitability for text classification tasks involving short to medium-length texts. The LSTM model, while effective in handling longer sequences, showed lower accuracy, suggesting that it may be more suitable for tasks requiring more context or longer text analysis.

Confusion matrices were constructed for each model to provide a detailed comparison of their performance across categories. These matrices illustrate the true positives, false positives, true negatives, and false negatives, offering insights into each model's classification capabilities. The high accuracy and low error rates of the AraBERT, CNN, and RNN models indicate their reliability in distinguishing between adult and non-adult content.

VI. CONCLUSION

In conclusion, the study demonstrates that AraBERT is the most effective model for detecting adult content in Arabic tweets, with the CNN model being a close second. The RNN model also performs well, while the LSTM model, although effective in handling longer sequences, shows slightly lower performance levels. These findings underscore the importance of selecting the appropriate model based on the specific characteristics of the text data and the classification task at hand. By providing a detailed analysis of each model's performance, this study contributes to the development of more accurate and reliable content moderation systems for Arabic social media platforms. The results highlight the need for continued innovation and improvement in machine learning and deep learning models to enhance their effectiveness in detecting explicit content (Table 3) (Figure 2).

Table 3 Comparison of the Performance of Each Model in Terms of Accuracy and F1 Score

Model	Accuracy (%)	F1 Score
AraBERT	100.00	1.00
CNN	94.27	0.94
RNN	94.22	0.94
LSTM	88.37	0.88

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep393

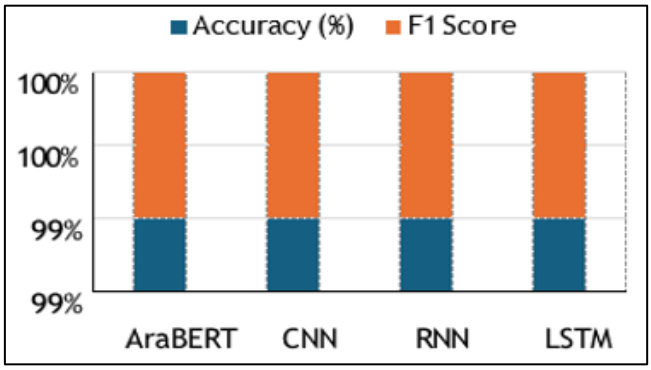

Fig 2 The Performance of Each Model in Terms of Accuracy and F1 Score

Tweets, contributing to the development of safer and more accurate content moderation systems.

FUTURE WORK

Future research in this area could focus on several key aspects to enhance model performance and improve the accuracy of detecting adult content in Arabic tweets. Advanced preprocessing techniques for Arabic text, such as handling diacritics and addressing colloquial idioms and spelling variations, can significantly improve model accuracy. Exploring hybrid and ensemble models by combining different models to leverage their strengths can lead to improved outcomes. Fine-tuning pretrained models specifically for Arabic or domain-specific datasets can increase their ability to understand nuanced language characteristics. Incorporating multimodal data, such as images or videos, along with text can provide additional context and improve classification accuracy. Increasing the dataset size and diversity by expanding the dataset and incorporating more diverse examples can improve the models' generalizability. Developing models optimized for real-time performance and ensuring scalability is crucial for practical content moderation. Investigating adaptive criteria based on cultural variables and addressing ethical implications, including privacy concerns and potential biases, is essential for developing responsible and fair systems. By focusing on these areas, future research can significantly enhance the accuracy and effectiveness of models for detecting adult content in Arabic.

REFERENCES

[1]. J. K. Appati, K. Y. Lodonu, and R. Chris-Koka, "A Review of Image Analysis Techniques for Adult Content Detection: Child Protection," *https://services.igi global.com/resolvedoi/resolve.aspx?doi=10.4018/IJSI.2021040106*, vol. 9, no. 2, pp. 102–121, Jan. 1AD, doi: 10.4018/IJSI.2021040106.

[2]. V. M. T. Ochoa, S. Y. Yayilgan, and F. A. Cheikh, "Adult video content detection using machine learning techniques," *8th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2012r*, pp. 967–974, 2012, doi: 10.1109/SITIS.2012.143.

[3]. G. M. Barrientos, R. Alaiz-Rodríguez, V. González-Castro, and A. C. Parnell, "Machine learning techniques for the detection of inappropriate erotic content in text," *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 591–603, Jun. 2020, doi: 10.2991/IJCIS.D.200519.003/METRI CS.

[4]. A. Dubettier, T. Gernot, E. Giguet, and C. Rosenberger, "A Comparative Study of Tools for Explicit Content Detection in Images," *Proceedings - 2023 International Conference on Cyberworlds, CW 2023*, pp. 464–471, 2023, doi: 10.1109/CW58918.2023.00077.

[5]. H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, "Arabic Offensive Language on Twitter: Analysis and Experiments," *WANLP 2021 - 6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop*, pp. 126–135, Apr. 2020, Accessed: Jan. 11, 2025. [Online]. Available: https://arxiv.org/abs/2004.02192v3

[6]. Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, p. 102048, Apr. 2024, doi: 10.1016/J.JKSUCI.2024.102048.

[7]. G. Gajula, A. Hundiwale, S. Mujumdar, and L. R. Saritha, "A machine learning based adult content detection using support vector machine," *Proceedings of the 7th International Conference on Computing for Sustainable Global Development, INDIACom 2020*, pp. 181–185, Mar. 2020, doi: 10.23919/INDIACOM49435.2020.90 83700.

[8]. H. Mubarak, S. Hassan, and A. Abdelali, "Adult Content Detection on Arabic Twitter: Analysis and Experiments," 2021. Accessed: Jan. 11, 2025. [Online]. Available: https://aclanthology.org/2021.wanlp- 1.14/