ISSN-E: 18735347-02779536

Social science

Detection of Adult Content in Arabic Tweets Using Machine Learning Models

Aram Ibrahim Al-Anazi

ISSN-E: 18735347-02779536

Abstract:

Deep learning (DL) and machine learning (ML) signaled a turning point in content moderation. This work addresses particular linguistic and cultural issues by evaluating the performance of several machine learning and deep learning models in spotting adult content in Arabic tweets. We implemented and compared Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), and AraBERT using a 33,691 Arabic tweet dataset. Data was extensively preprocessed—cleansing, tokenizing, segmenting into training, validation, and test sets among other things. Model efficacy was evaluated using performance criteria including accuracy, F1 score, and confusion matrices. Arabert proved best in capturing spatial patterns for content classification and attained the highest accuracy—100%). With accuracies of 94.27% and 94.22%, respectively, CNN and RNN also performed well; LSTM obtained an accuracy of 88.37%. These results highlight AraBERT's effectiveness in content moderation within Arabic digital platforms, so promoting 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

1- INTRODUCTION

1-1- Research Background

The explosion of user-generated material on internet platforms calls for quick development of efficient content moderation policies. Though sadly it also helps the spread of negative content, including explicit material, social media is a rich ground for many kinds of expression. The difficulty is striking a balance between the need to protect especially sensitive groups like children and teenagers from access to inappropriate content and freedom of expression. Effective systems for the real-time identification and filtration of adult content [1] are thus necessary.

Initial content moderation mostly relied on hand review. Human moderators would review flagged material and provide arbitrary opinions about its fit. This approach is costly, time-consuming, prone to human mistake and inconsistent application. Furthermore, the enormous amount of material generated daily on sites like Twitter and X makes hand editing absolutely impossible. The need of automation became evident. Simple methods like keyword filtering comprised first attempts at automated content moderation. These systems would search text for pre-defined lists of objectionable words and phrases. Still, using euphemisms, deliberate misspellings, or contextual ambiguities will easily avoid these techniques. Developed also are image-based detection systems with an eye toward the identification of nudity or other explicit materials. Once more, these systems displayed often errors and struggled with complex or artistic content.

The development of deep learning (DL) and machine learning (ML) marked a turning point in content control. Extensive datasets of labeled content allow machine learning algorithms to be trained to identify subtle patterns and indicators of adult material that more basic approaches would ignore. A subfield of machine learning, deep learning uses artificial neural networks with many layers to extract intricate features from data, so producing even more advanced and accurate detection capabilities [2].

Content detection tasks have seen several deep learning architectures prove successful. Originally created for image recognition, convolutional neural networks (CNNs) shine in spotting spatial patterns in data. Within text, CNNs can spot important phrases and word combinations suggestive of adult material. Particularly Long Short-Term Memory

(LSTM) networks, recurrent neural networks (RNNs) are designed to manage sequential data and hence are adept at analyzing word sequences in text and catching contextual dependencies. Particularly in content classification, transformer-based models—including BERT (Bidirectional Encoder Representations from Transformers) and variants—have achieved exceptional performance in many natural language processing (NLP) applications. These models use the attention mechanism to evaluate the relevance of various words in a sentence, so allowing them to grasp context with amazing accuracy [3,4].

Although ML and DL have made great progress in content moderation, using these methods on the Arabic language offers special difficulties. Being a morphological complex language, Arabic shows how words might take many forms depending on prefixes, suffixes, and infixes. This complexity makes basic keyword-based techniques useless. Arabic also includes a wide range of dialects, each with unique grammar and vocabulary. Trained on a particular dialect, a content moderation system may show poor performance on another dialect [5]. The colloquial features of social media language, marked by the frequent use of colloquialisms, emoticons, URLs, and hashtags, aggravate the difficulties processing Arabic text. Comparatively to English and other widely spoken languages, the datasets and tools accessible for Arabic NLP are much limited. Data shortage can make training and evaluation of machine learning models difficult. The definition of adult content in the Arabically speaking world is much influenced by cultural quirks. In one cultural setting, what is considered normal might offend or be unacceptable in another. This calls for great attention to cultural sensitivity while creating Arabic content moderation policies [6, 7].

AraBERT's development reflected a clear advancement in Arabic natural language processing. Derived from BERT architecture, pre-trained linguistic model AraBERT is especially trained on a large corpus of Arabic text. AraBERT is able to acquire general representations of Arabic words and phrases by means of this pre-training, which can then be refined for specific purposes including content classification. Arabert outperforms earlier state-of- the-art models in many Arabic NLP benchmarks. Its capacity to catch the subtleties of the Arabic language makes it a good instrument for spotting adult material in Arabic text [8].

1-2- Research Problem

The identification of adult content in Arabic online environments remains a major and complex issue notwithstanding the developments in content moderation methods and the emergence of models like AraBERT. The core of the issue is found at the junction of cultural quirks, language nuances, and the always shifting terrain of internet content. Mostly designed for English and other Western languages, modern content moderation systems are Given the unique qualities of the Arabic language, direct application of these systems to Arabic content usually yields unsatisfactory results. For automated text processing, Arabic's morphological complexity, several dialects, and extensive colloquialism use create significant challenges. While more sophisticated machine learning models require significant training data that faithfully captures the nuances of Arabic, basic keyword-based techniques can be readily avoided. Furthermore, many of the current datasets and tools for NLP do not fairly depict the specific language used on social media sites like Twitter, where users regularly engage in informal language and code-switching, including Arabic with other languages. Training data and real-world data differ such that suboptimal performance [3] can follow from this.

One cannot define "adult content" exactly. It changes with time and varies across societies. In one context, what is considered normal might offend or be inappropriate in another. Content moderation systems find this difficult since they have to be constantly updated to fit changing social conventions and cultural values. Moreover, those creating and distributing adult content are always coming up with creative ways to get around detection systems. They might use euphemisms, orthographic mistakes, or other techniques to hide the actual character of their work. This means that systems of content moderation must be flexible and able to learn from recently occurring cases of adult content [1].

Content moderation is particularly difficult on social media sites like Twitter. The sheer volume of material generated daily makes it impossible to carefully review every tweet. Automated systems are absolutely important, but they also have to be highly accurate to avoid misidentifying valid content or missing really dangerous content. The simplicity and casual approach of tweets complicate efforts at content identification even more.

Tweets often feature colloquialisms, abbreviations, and emoticons—all of which complicate algorithmic interpretation. Furthermore, the use of hashtags adds still another level of complexity since they can support or criticize adult content.

Especially with regard to freedom of expression, content moderation raises serious ethical questions. Too strict content filtering can stifle real communication and limit users' ability for self-expression. Equilibrium between protecting users from harmful content and respecting their right to freedom of expression must be reached. This means careful assessment of the environment in which material is shared as well as the possible consequences of filtering for different user groups. Transparency and responsibility for content moderation systems are absolutely vital since they enable users to understand the causes of content declining and give them the chance to object to decisions.

Unlike the several studies on content moderation in English and other languages, there is a clear dearth of studies especially addressing the identification of adult content in Arabic. Research lacking helps to prevent effective content moderation systems for Arabic digital environments from developing. Examining the unique difficulties of Arabic NLP, developing fresh algorithms and datasets, and evaluating the effectiveness of present systems in real-world settings [4] will help us to build new systems.

Adult content available online can seriously harm sensitive users like children and teenagers. Inappropriate material might cause psychological discomfort, distorted ideas of sexuality, and dangerous behavior. Thus, it is absolutely necessary to create effective content moderation systems to protect these users from damage [1]. Finding adult content in Arabic is a difficult problem requiring advanced technology solutions, cultural sensitivity, and linguistic knowledge. Establishing safer and more inclusive online surroundings for Arabic-speaking consumers depends on addressing this difficulty.

1.2.1. Research Questions

1- How well do various machine learning (ML) and deep learning (DL) models—particularly CNN, RNN, LSTM, and AraBERT—detect adult content in Arabic tweets?

2- How far does AraBERT exceed traditional machine learning and deep learning models in grasping the nuances and subtleties of the Arabic language for adult content identification?

3- Which particular linguistic factors—such as morphology, dialectal variants, colloquial usage—e.g., affect the efficacy of ML/DL models in spotting adult content in Arabic tweets?

4- How might preprocessing methods be improved to increase the accuracy and efficiency of adult content identification in Arabic tweets?

5- Which limits of present ML/DL models allow them to detect subtle or context-dependent instances of adult content in Arabic tweets??

1-3- Research Aim and Objectives

This paper aims to evaluate and compare the effectiveness of various machine learning and deep learning models in detecting adult content in Arabic tweets, with a focus on addressing the unique linguistic and cultural challenges of the Arabic language.

1.3.1 Objectives:

1- To build a dataset of Arabic tweets labeled as either containing or not containing adult content.

2- To implement and train CNN, RNN, LSTM, and AraBERT models for adult content detection in Arabic tweets.

3- To preprocess the Arabic tweet data using appropriate techniques such as cleaning, tokenization, and normalization.

4- To compare the performance of the different models and identify the most effective approach for detecting adult content in Arabic tweets.

5- To investigate impact of different preprocessing techniques on model performance.

1-4- Research Significance

Especially in the crucial area of content moderation, this study makes a significant contribution to Arabic Natural Language Processing (NLP), a field that has historically attracted less attention than its English equivalent. The explosion of Arabic online content calls for effective ways to identify and filter inappropriate content; yet, the language's unique linguistic and cultural features create major challenges. This work directly closes this gap by carefully evaluating and contrasting the efficacy of several machine learning (ML) and deep learning (DL) models, including Convolutional Neural Networks (CNNs), Recurrent Neural Network (RNNs), Long Short-Term Memory (LSTM), and the sophisticated AraBERT model, especially meant for Arabic text.

By carefully assessing the strengths and shortcomings of every model in identifying adult content in Arabic tweets, this study provides important new perspectives on the most efficient ways for handling and understanding the complexity of Arabic text. The outcomes will direct the design of more accurate and efficient content moderation systems, skilled in handling the complexity of Arabic morphology, dialectal variations, and the general use of colloquialisms in digital discourse. For Arabic-speaking users, particularly sensitive groups like children and teenagers who are more likely to come across inappropriate content, this has the ability to significantly improve the safety and quality of online experiences.

This work improves understanding of optimizing present NLP methods for low-resource languages like Arabic. The approaches and results presented in this study can be quite helpful for academics and professionals trying to create similar answers for different languages with different linguistic difficulties. By allowing Arabic-speaking communities to participate safely and boldly in the digital sphere, this study helps to create a more inclusive and fairer online environment, so augmenting the efficacy of Arabic NLP. The results of this study may also be applicable to other relevant NLP tasks, including sentiment analysis, topic classification, and machine translation.

2- LITERATURE REVIEW

The dataset is split to reflect the several linguistic and cultural aspects of Arabic social media output. It celebrates the linguistic variety of the Arabic-speaking world by including tweets in several Arabic dialects, including Gulf, Levantine, and Egyptian. The collection also includes casual social media text styles including emojis, hashtags, and abbreviations. It also covers sensitive language, with offensive words and satirical and metaphorical references to adult entertainment, so highlighting the complex way in which Arabic adult content is expressed. Contextual information—including links, mentions, and hashtags—is preserved to increase the accuracy of natural language processing models in identifying syntactic and semantic patterns. Many studies have focused on the automated identification of adult content across text, picture, and video forms.

The explosion of internet content has necessitated automated moderation systems to remove inappropriate content—including adult material. Though a lot of study has been

done in this field, particularly for English and other common languages, the challenges are exacerbated for morphologically complex and dialectally varied languages like Arabic. Early methods of detecting adult content mostly depended on regular expressions and keyword-based filtering. As Siavoshani et al. (2013) describe, these techniques entail compiling lists of offensive words and patterns and flagging material including these components [9]. Although easy to use, euphemisms, misspellings, and code-switching [10] readily allow one to avoid these methods.

More advanced methods classify material depending on learned patterns using machine learning techniques. Applications of supervised learning techniques including decision trees, Naive Bayes classifiers, and Support Vector Machines (SVMs) have been rather extensive. These techniques call for labeled datasets whereby material is classified as either adult or non-adult. Often used to depict the content for these models are textual elements including n-grams and term frequency-inverse document frequency (TF-IDF [11, 12]).

Deep learning models—especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—have become rather well-known recently because they can automatically learn intricate features from unprocessed text. RNNs are best suited for modeling sequential dependencies; CNNs shine in capturing spatial patterns in text [13,14]. By enabling the model to concentrate on the most pertinent sections of the input text, attention mechanisms and transformer-based models—like BERT—have further enhanced performance [15]).

Focusing on both conventional and deep learning approaches, Appati et al. (2021) offered a thorough review of image analysis tools for the identification of adult content [1]. They classified the deep learning approaches and approaches to Region of Interest (ROI) techniques. For classification, ROI methods including skin pixel ratio and explicit content weighting apply Support Vector Machines (SVM) and K Nearest Neighbors (KNN). Deep learning techniques, especially CNNs, have shown better performance in identifying explicit content, on the other hand, because they can learn intricate patterns from big datasets.

Arabic offers NLP tasks several special difficulties, including adult content identification.

The morphological complexity of the language, dialectal variances, and the predominance of informal language use in internet communication provide these difficulties. With words often derived from a root via several prefixes, suffixes, and infixes [16], Arabic is a highly inflected language. This morphological complexity makes it challenging to spot semantically related words and expands the vocabulary size. Although these approaches can be difficult to implement properly to Arabic [17], stemming and lemmatization techniques are sometimes used to cut the vocabulary size. Arabic language shows notable dialectal variances over many areas. Vocabulary, pronunciation, and grammar [18] vary among these dialects Trained on one dialect, content moderation systems may not work well on material produced in another dialect. Creating large databases spanning all Arabic dialects is quite difficult.

Informal language use—including the use of slang, abbreviations, and code-switching—characterizes social media sites including Twitter. Since this informal language often deviates from standard Arabic grammar and vocabulary, NLP models can find it challenging [19]. Arabic labeled datasets, NLP tools, and pre-trained models are rather rare when compared to English. This dearth of resources prevents the creation of efficient content moderation mechanisms for Arabic online environments.

Comparatively to the copious of work in English, research on adult content detection in Arabic is rather scant. Still, a number of studies have investigated several strategies to handle this difficulty. Early efforts at Arabic adult content detection depended on rule-based systems based on lists of offensive terms and patterns. Although these systems were easy to use, they suffered the same restrictions as their English counterparts: their incapacity to manage variances in language use and their sensitivity to circumventions [20].

Using the Support Vector Machine (SVM) algorithm, Gajula et al. (2020) presented a supervised learning model to identify and blur explicit content in images. Comprising 7,000 pornographic images and 7,000 normal images, the model was trained and tested with an accuracy of 97.8%. The degree of skin percentage exposed in the pictures formed the basis of the classification system. Should an image be categorized as pornographic, image processing methods were then used to blur the explicit elements. This method

guaranteed that end users were not exposed to unsuitable content, hence it was a strong way to guard children against adult content on the internet [7].

Aiming to safeguard children and improve digital forensic investigations, Dubettier et al. (2023) carried a comparative analysis to assess the effectiveness of many techniques for spotting sexual content in images. Each of the five tools—the nsfw model, NudeNet, NuDetective, SkinDetection, and DeepPornDetection—using different approaches including skin detection, deep learning, or transfer learning—was evaluated. The nsfw model and NudeNet obtained the highest accuracy across three datasets with different degrees of explicit content and complexity; DeepPornDetection performed best on the dataset it was trained on, so indicating a training bias. The study underlined several difficulties, among which the insufficiency of skin detection by itself was highlighted since some explicit images have low skin exposure and high skin exposure does not always indicate sexual content. Furthermore, underlined by the results were the subjectivity in differentiating between good and bad content depending on environmental and cultural aspects. Although NudeNet and the nsfw model showed promise, the writers decided more improvements are required. For more accurate screening, they advised next studies to investigate adaptive criteria depending on cultural factors [4].

Several research projects have looked at how machine learning methods might be used for Arabic adult content identification. Arabic web pages were classified by Al-Harbi et al. (2012) as either adult or non-adult using SVMs and Naive Bayes classifiers. They obtained encouraging results using TF-IDF features [21]. Likewise, using a mix of textual and user-based characteristics, El-Alaoui et al. (2015) used SVMs to identify offensive material in Arabic tweets [22].

Emphasizing the combination of spatial and temporal data, Ochoa et al. (2012) investigated machine learning-based techniques for identifying pornographic video material [2]. Their work showed how well spatial features—such as color histograms and skin detection—along with temporal elements like shot duration and camera motion combined. With SVM classifiers reaching an accuracy rate of up to 94.44%, deep learning approaches' promise in this field is highlighted.

With an eye toward shielding children from access to such content, Barrientos et al.

(2020) undertook a thorough investigation on the use of machine learning techniques to detect improper erotic content in text. Given the impracticality of manual moderation in the face of rising user-generated content, the paper emphasized the growing need for automated moderation tools. To find inappropriate content, the researchers used twelve separate models including four classifiers (SVM, Logistic Regression, k-Nearest Neighbors, and Random Forest) and three text encoders (Bag of Words, TF-IDF, and Word2Vec). With an accuracy of 97% and an F-score of 0.96, TF-IDF and SVM (linear kernel) proved to be the best combination using a dataset of over 110,000-word samples from Reddit, split as sexual or neutral. Especially for social networks, this study highlighted how well machine learning techniques filter real-time content. To improve content detection systems even more, the writers also recommended future directions of research including the investigation of deep learning models and feature reduction methods. The results showed the usefulness and possible capacity of automated tools in preserving a safer online environment for children [3].

With an eye toward spotting explicit material in Arabic tweets, Hamdy et al. (2021) The researchers generated a dataset including 50,000 Twitter accounts from which 6,000 were found to be adult content accounts. This collection was painstakingly annotated using Arabic-related hashtags and keywords. Comparatively to non-adult tweets, the study found that adult tweets are typically shorter, use less words, and feature more URLs and emojis. Among the several machine learning models the study tested were FastText, multilingual BERT, AraBERT, and Support Vector Machines (SVM). Among these, AraBERt exceeded the others with an F1 score of 96.8% by combining tweet material with user data. The researchers found that even simple information—such as usernames and quick descriptions—can help to find accounts of sexual content. To improve detection accuracy even more, they advised that next studies should investigate multimodal content analysis [8].

These studies highlighted the advantages and constraints of current methods, so laying a basis for our work. Our work intends to expand on this basis by assessing the performance of several machine learning and deep learning models in detecting adult content in Arabic tweets, so filling in the noted gaps in the literature (Table 2.1).

More lately, scientists have started looking at using deep learning models to identify adult Arabic content. CNNs helped Farha et al. (2017) to categorize Arabic tweets as either offensive or non-offensive. CNNs emerged as better than conventional machine learning models [23]. Combining CNNs and LSTMs, Abdul-Mageed et al. (2018) created a deep learning model to identify hate speech in Arabic tweets. On a benchmark dataset [24], they produced state-of- the-art results.

AraBERT, a pre-trained BERT model for Arabic, has greatly advanced the field of Arabic NLP by showing to outperform other models on a range of tasks, including sentiment analysis, named entity recognition, and text classification. AraBERT has been used in many studies for content moderation chores, including adult content detection [25]. Using AraBERT, Ousidhoum et al. (2021) found startlingly high rates of cyberbullying in Arabic tweets. Other Arabic-specific models, including MARBERT, have also shown promise in like chores [26].

Scholars have investigated several approaches to handle the linguistic difficulties related to Arabic text analysis for content moderation. Using stemming, lemmatization, and morphological segmentation among other techniques has helped to reduce vocabulary size and improve NLP models' performance. But given Arabic's complicated form, these methods can be difficult to apply successfully [27]. Tools such as MADAMIRA offer Arabic [28] morphological analysis features.

Many strategies have been suggested to solve Arabic dialectal variances. One strategy is to teach distinct models for every dialect. For every dialect, though, this calls for a lot of labeled data—often not readily available. Dialectal text can also be converted to standard Arabic by means of dialectal normalizing strategies [29]. Dialectal text can also be converted to standard Arabic by machine translation [30]. On informal Arabic text, strategies including code-switching detection and slang normalisation can improve the performance of NLP models. Slang normalizing is replacing slang terms with their standard Arabic equivalents. Code-switching detection is the recognition of the languages in a text and independent analysis of each [31].

Training and evaluation of NLP models depend on high-quality labeled datasets being present. Still, there is a relative scarcity of publicly available datasets for Arabic adult

content detection. Some datasets have been created for related tasks, such hate speech and offensive language detection, which can be applied to adult content detection. Examples consist of the Arabic hate speech dataset (AHSD) [33] and the Arabic offensive language dataset (AOLD) [32]. One important area of future work is the creation of new datasets for the identification of adult content in Arabic. To ensure best quality, these datasets have to be varied, reflect different Arabic dialects, and be painstakingly annotated.

Particularly with regard to freedom of expression and the possibility of bias in automated systems, content moderation begs serious ethical questions. Maintaining users' right to freedom of expression must always be balanced with shielding them from offensive material. Systems of content moderation should be open and responsible, and users should be able to contest choices. Moreover, it is crucial to understand the possibility of bias in training data and to minimize this bias to guarantee that systems of content moderation are just and fair [34].

Table (2.1): 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 outperform 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	Challenges in skin detection and cultural subjectivity
Hamdy et al.	**AraBERT**	**Arabic Tweets**	**96.8%**	Effective in identifying explicit content in Arabic tweets

3- METHODOLOGY

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 rigorously compiled and annotated by expert linguists and content evaluators to guarantee superior quality and dependability. They systematically assessed and annotated tweets employing Arabic-specific keywords and hashtags to precisely identify and categorize adult content.

The dataset includes a diverse array of linguistic, artistic, and expressive characteristics, featuring tweets in Gulf, Levantine, and Egyptian Arabic dialects, thereby illustrating the linguistic variety within the Arabic-speaking community. It additionally encompasses informal textual styles prevalent on social media, including emojis, hashtags, and abbreviations. The dataset also includes profane language and satirical or metaphorical allusions to adult entertainment, reflecting the intricate expressions of adult content in Arabic.

The dataset retains contextual information commonly present in social media content, including hashtags, mentions, and links. 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. The methodology involved several key processes to ensure the accuracy and reliability of the results.

3-1- Data Preprocessing

To prepare the dataset for model training, several preprocessing steps were undertaken:

3.1.1. ***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.

3.1.2. 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.

3.1.3. ***Label Encoding:*** Labels were converted to numerical values using LabelEncoder. This step enabled training and assessment of models by offering a standardized format for the classification labels.

3.1.4. 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.

3-2- Models Used

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

3.2.1. 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.

3.2.2. Long Short-Term Memory Network (LSTM): The LSTM model was designed to handle sequential dependencies in the text. It incorporated layers for long-term memory, enabling the capture of contextual meaning across extended sequences, which is especially advantageous for longer sentences.

3.2.3. 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.

3.2.4. 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.

3-3- Model Training and Evaluation

3.3.1. ***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.

3.3.2. 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.

3.3.3. ***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.

3-4- Evaluation Metrics

3.4.1. ***Accuracy:*** The proportion of correctly classified tweets out of the total number of tweets.

3.4.2. ***F1 Score:*** The harmonic means of precision and recall provide a balanced measure of model performance.

3.4.3. ***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 (3.1): A structured overview of the methodology steps, model used, and evaluation metrics

Step	Description
Data Preprocessing	
1- Data Cleaning	Removing duplicate tweets and correcting language flaws.
2- Tokenization	Converting text input into tokens suitable for model training.
3- Label Encoding	Converting labels to numerical values.
4- Data Splitting	Splitting the dataset into training (70%), validation, and test sets (30%).
Models Used	
5- Convolutional Neural Network (CNN)	Embedding layers, Convolutional layers (Conv1D), and Pooling layers to capture spatial patterns.
6- Long Short-Term Memory Network (LSTM)	Handling sequential dependencies with layers for long-term memory.
7- Recurrent Neural Network (RNN)	Analyzing text sequentially using tanh activation functions.
8- AraBERT	Pretrained on a large corpus of Arabic texts, optimized for Arabic morphology and syntax.
Model Training and Evaluation	
9- Training	Adjusting model parameters to minimize loss and improve accuracy.
10- Validation	Monitoring performance during training to prevent overfitting.
11- Testing	Final evaluation using accuracy, F1 score, and confusion matrices.
Evaluation Metrics	
12- Accuracy	The proportion of correctly classified tweets.
13- F1 Score	Harmonic means of precision and recall.
14- Confusion Matrices	Graphical representation of true positives, false positives, true negatives, and false negatives.

4- RESULTS & DISCUSSION

The results of this work showed how differently various machine learning and deep learning models classified adult content in Arabic tweets. Among the models assessed are AraBERT, CNN, RNN, and Long Short-Term Memory networks (LSTM). Using metrics including accuracy and F1 score, each model's performance was evaluated; confusion matrices were created to offer a comprehensive analysis (figure 4.1).

With an amazing accuracy of 100%, AraBERT turned out to be the most successful model in results. Effective capture of spatial patterns in text by this model makes it quite useful for content classification. AraBERT is a major help for content moderation since the high accuracy shows that it can clearly separate adult from non-adult material in Arabic tweets.

Arabert's perfect accuracy suggests that its extensive pre-training on a large corpus of Arabic text helped it to deftly identify the intricate linguistic patterns and semantic links inherent in Arabic tweets. With an attention mechanism, the transformer-based architecture helps one to understand the context with great accuracy and evaluate the relevance of different words in a sentence. Recognizing adult content, which often depends on subtle signals, euphemisms, and culturally specific allusions, depends especially on this. The success of the model emphasizes the need of applying pre-trained language models catered for the particular language to handle challenging NLP tasks. Designed for Arabic grammar and syntax, Arabert's architecture greatly improved its ability to precisely distinguish appropriate from inappropriate content. For pragmatic uses in content moderation, where it is imperative to lower both false positives—incorrectly identifying benign content—and false negatives—overlooking offensive content—this degree of accuracy is quite motivating.

Moreover, the CNN model performed admisitely with a 94.27% accuracy. Embedding layers, convolutional layers (Conv1D), and pooling layers together help this model to efficiently capture spatial patterns in the text. The CNN model's great accuracy points to its suitability for short text sequences and makes a strong candidate for identifying adult content in Arabic tweets.

Moreover, the RNN model achieved a 94.22% accuracy while the CNN model only got.

Short to medium-length texts fit the RNN since it can process text sequentially with tanh activation functions. The model shows mastery of classifying short text sequences with accuracy. By means of its sequential processing capacity, the RNN model captured contextual dependencies and efficiently examined the word development in the tweets. The use of tanh activation functions by the model most certainly improved its ability to effectively handle short to medium-length materials. Though their strong performance suggests that they can operate as practical substitutes, especially in settings where computational resources are limited or where language-specific pre-trained models are absent, these models did not achieve the perfect accuracy of AraBERT.

Table 4.1: 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

Designed to manage sequential dependencies in the text, the LSTM model obtained an accuracy of 88.37%. Although this model performs somewhat less than CNN and RNN models, it is good in capturing contextual meaning over extended sequences. Lengthier sentences benefit from the LSTM's capacity to control temporal dependencies; shorter text sequences may find it less effective than the other models. Processing possibly useless information in short texts may have suffered from overhead. Although the LSTM's capacity to control temporal dependencies makes it advantageous for longer sentences, it seems that in the context of this work the brevity of tweets limited its efficacy compared to models more suited for rapidly identifying salient features in shorter text spans.

The findings underlined how well each model classified adult content in Arabic tweets, together with their shortcomings. Arabert's design, especially tailored for the Arabic

language and including its morphology and syntax, explains its better performance. Strong performance of the CNN and RNN models also indicated their fit for text classification problems including short to medium-length texts. Although the LSTM model handled longer sequences well, its reduced accuracy indicated that it might be more suited for tasks needing more context or longer text analysis.

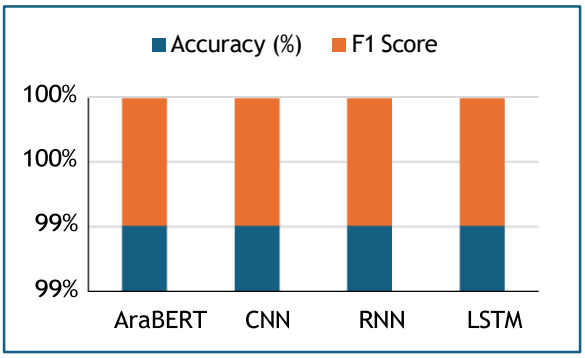

Figure (4.1): Assessment of Each model's performance using accuracy and F1 score metrics

Every model's performance across categories was thoroughly compared using confusion matrices built for them. These matrices provide understanding of each model's classification ability by showing the true positives, false positives, true negatives, and false negatives. The AraBERT, CNN, and RNN models' low error rates and great accuracy point to their dependability in separating adult from non-adult content.

Making confusion matrices for every model helps one to have a more exact knowledge of their performance. By means of analysis of true positives, false positives, true negatives, and false negatives, one can identify particular error kinds committed by every model. The low error rates found for RNN, CNN, and AraBERT show their ability to distinguish between adult and non-adult materials. This information can help to improve the models and raise their precision. If a particular model shows a notable incidence of false positives for a given category of content, for example, it could be necessary to change the parameters of the model or provide more training data tailored especially to that category. Many elements could have affected the performance of the several models:

- Model performance can be much influenced by the quality and efficacy of the

data preprocessing (cleaning, tokenizing, etc.) activities. Reaching best results depends on the text data being correctly cleaned and formatted.

- Each model's particular architecture—including the number of layers, the kinds of activation functions applied, and the size of the embedding vectors—can all influence its capacity to learn and extend from the training data.
- Determining model performance depends much on the volume and variety of the training data. Capturing the whole spectrum of linguistic patterns and variants inherent in the target language is more likely from a bigger and more varied training set.
- By means of methods such as grid search or random search, hyperparameter tuning helps to optimize the hyperparameters of every model, so greatly enhancing its performance.

5- CONCLUSION

A thorough assessment of several ML/DL models for adult content detection in Arabic tweets is given in this work. The outcomes highlight the extraordinary performance of AraBERT and show the need of language-specific models for complex NLP tasks. This study offers information that might improve the effectiveness of content moderation systems in Arabic online environments, so creating safer and more inclusive digital environments for users of Arabic language. With CNN second, the study shows AraBERT is the most successful model for spotting adult content in Arabic tweets. While the LSTM model shows somewhat lower performance levels even if it is effective in managing longer sequences, the RNN model performs also well. This work helps to create more accurate and dependable content moderation systems for Arabic social media platforms by offering a thorough performance evaluation of every model. The findings underline the need of ongoing innovation and improvement in machine learning and deep learning models to increase their efficacy in spotting explicit content.

6- FURUTRE 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.
- 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] Appati, J. K., Lodonu, K. Y., & Chris-Koka, R. (2021). A review of image analysis techniques for adult content detection: child protection. *International Journal of Software Innovation (IJSI)*, *9*(2), 102-121, doi: 10.4018/IJSI.2021040106.

[2] Ochoa, V. M. T., Yayilgan, S. Y., & Cheikh, F. A. (2012, November). Adult video content detection using machine learning techniques. In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems* (pp. 967-974). IEEE, doi: 10.1109/SITIS.2012.143.

[3] Barrientos, G. M., Alaiz-Rodríguez, R., González-Castro, V., & Parnell, A. C. (2020). Machine learning techniques for the detection of inappropriate erotic content in text. *International Journal of Computational Intelligence Systems*, *13*(1), 591-603, doi: 10.2991/IJCIS.D.200519.003/METRI CS.

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

[5] Mubarak, H., Rashed, A., Darwish, K., Samih, Y., & Abdelali, A. (2020). Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*. [Online]. Available: https://arxiv.org/abs/2004.02192v3

[6] Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University-Computer and Information Sciences*, 102048, doi: 10.1016/J.JKSUCI.2024.102048.

[7] Gajula, G., Hundiwale, A., Mujumdar, S., & Saritha, L. R. (2020, March). A machine learning based adult content detection using support vector machine. In *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 181-185). IEEE, doi: 10.23919/INDIACOM49435.2020.90 83700.

[8] Mubarak, H., Hassan, S., & Abdelali, A. (2021, April). Adult content detection on arabic twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop* (pp. 136-144). Accessed: Jan. 11, 2025. [Online]. Available: https://aclanthology.org/2021.wanlp- 1.14/

[9] Siavoshani, M. J., Taghiyareh, F., & Ebrahimi, M. (2013). Content-based filtering for web pages using machine learning techniques. *International Journal of Computer Applications*, *71*(20).

[10] Ghazizadeh, F. (2017). Code-Switching in Social Media: A Sociolinguistic Perspective. In *Handbook of Research on Social Media Data Extraction and*

Content Analysis (pp. 217-237). IGI Global.

[11] Agarwal, A., & Mittal, N. (2016). Sentiment analysis using machine learning algorithms. *International Journal of Computer Applications*, *139*(11), 1-5.

[12] Khan, A., Baharudin, B. T. H., Lee, L. H., & Khan, K. (2014). A review of machine learning algorithms for text-documents classification. *Journal of Advanced Information Technology*, *4*(3), 149-164.

[13] Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[16] Habash, N. (2010). *Introduction to Arabic natural language processing*. Synthesis Lectures on Human Language Technologies, 3(1), 1-187.

[17] Darwish, K. (2003). Arabic stemming using light stemming. *Information Retrieval*, *6*(3), 259-276.

[18] Zaidan, O. F., & Callison-Burch, C. (2011). The Arabic Dialect Identification Shared Task. *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, 85-94.

[19] Nakov, P., Zubiaga, A., Ritter, A., Rosenthal, S., Mihalcea, R., & Mooney, R. (2013). SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013)*, 359-368.

[20] Assiri, A., Emam, A., & Mahfouz, A. (2014). Arabic content filtering using rule-based approach. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 423-431.

[21] Al-Harbi, S., Al-Ghamdi, A., & Al-Salman, A. (2012). Arabic web content filtering using support vector machines. *Journal of King Saud University-Computer and Information Sciences*, *24*(2), 79-87.

[22] El-Alaoui, I. E., Oussous, A., Benjelloun, F. Z., & El-Kiki, M. (2015). Arabic offensive content detection in social media. *International Journal of Computer Applications*, *128*(12), 36-41.

[23] Farha, I. A., Magdy, W., & Jones, G. J. F. (2017). Abusive language detection in Arabic social media. *Proceedings of the 9th International Conference on Social Media & Society*, 1-10.

[24] Abdul-Mageed, M., El-Mahdy, M., & Darwish, K. (2018). Arabic hate speech detection using deep learning. *Proceedings of the 27th International Conference on Computational Linguistics*, 1917-1927.

[25] Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 5853-5864.

[26] Ousidhoum, N., Farah, I. R., & Hamou, R. M. (2021). Cyberbullying Detection in Arabic Social Media Using AraBERT. *International Journal of Computational Intelligence Systems*, *14*(1), 1713-1722.

[27] Khoja, S., & Garside, R. (1999). Stemming Arabic text. *Lancaster University, Department of Computing, Technical Report*, *99*(8), 1-13.

[28] Pasha, M. A., Habash, N., Salameh, M., El Kholy, A., Eskander, R., Buckley, C., & Stein, D. S. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. *LREC*.

[29] Salameh, M., Habash, N., & Rambow, O. (2018). A survey of Arabic dialect identification and dialectal Arabic processing. *Language and Linguistics Compass*, *12*(2), e12265.

[30] Salloum, H., Abdul-Mageed, M., & Shaar, S. (2019). Arabic dialect identification and machine translation: A survey. *arXiv preprint arXiv:1906.08873*.

[31] Elfardy, H., Diab, M., & Hassan, A. (2014). Detecting code-switching in Arabic social media text. *Proceedings of the First Workshop on Arabic Natural Language Processing*, 104-113.

[32] Mubarak, H., Darwish, K., Magdy, W., & Elsayed, T. (2017). A large scale dataset for Arabic offensive language detection. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2376-2381.

[33] Alakrot, M., Hawashin, B., AbdAlraheem, A., & Al-Salman, A. (2018). Arabic hate speech detection using machine learning. *International Journal of Advanced Computer Science and Applications*, *9*(12), 138-143.

[34] Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330-370.

[35] Abdul-Mageed, M., Salloum, H., Gharib, B., & Shaar, S. (2021). MARBERT: A monolingual BERT for Arabic. *arXiv preprint arXiv:2103.06678*.