Adult Content Detection on Arabic Twitter: Analysis and Experiments

Hamdy Mubarak, Sabit Hassan and **Ahmed Abdelali**
Qatar Computing Research Institute
Doha, Qatar
{hmubarak,sahassan2,aabdelali}@hbku.edu.qa

Abstract

With Twitter being one of the most popular social media platforms in the Arab region, it is not surprising to find accounts that post adult content in Arabic tweets; despite the fact that these platforms dissuade users from such content. In this paper, we present a dataset of Twitter accounts that post adult content. We perform an in-depth analysis of the nature of this data and contrast it with normal tweet content. Additionally, we present extensive experiments with traditional machine learning models, deep neural networks and contextual embeddings to identify such accounts. We show that from user information alone, we can identify such accounts with F1 score of 94.7% (macro average). With the addition of only one tweet as input, the F1 score rises to 96.8%.

1 Introduction

Disclaimer: Due to the nature of this research, we provide examples that contain adult language. We follow academic norms to present them in an appropriate form, however the discretion of the reader is cautioned.

In recent years, Twitter has become one of the most popular social media platforms in the Arab region (Abdelali et al., 2020). On average, Arab users post more than 27 million tweets per day (Alshehri et al., 2018). Such popularity has also spawned a number of spammers who exploit the popularity to post malicious content. Such malicious content may contain pornographic references or advertisement. We refer to such content as adult content. Adult content may have deliberating effects on many, particularly among those of younger age groups. Users who fall for the pornographic advertisements are at risk of losing money and sensitive information to the spammers. Due to the massive amount of user-generated content on Twitter, it is impossible to detect such accounts manually and this calls for automatic detection —the focus of this paper.

Twitter's policy prohibits users from posting adult content[1]. However, the methods deployed for detecting spams such as adult content are mostly expanded from English and are not very effective for detecting accounts who post adult content in other languages as such as the case of Arabic (Abozinadah et al., 2015). Traditional methods such as filtering by list of words are not effective since spammers use smart ways such as intentional spelling mistakes to evade such filtering (Alshehri et al., 2018).

Despite the dire need of eliminating adult content from Arabic social media, there has been a very few notable works (Alshehri et al., 2018; Abozinadah et al., 2015; Abozinadah and Jones, 2017) in the field. In contrast to the existing work (Alshehri et al., 2018; Abozinadah et al., 2015; Abozinadah and Jones, 2017) that rely on extracting collection of tweets from each account to classify whether they post adult content, we present a dataset and several models aimed at classifying accounts based on minimal information. By minimal information, we mean user information such as username, user description or just one random tweet from each account. In our study, we focus on using textual features to detect adult content and we leave multimedia (e.g. images) and social network features for future work.

Our dataset consists of 6k manually annotated Twitter accounts who post adult content and 44k ordinary Twitter accounts in addition to a tweet from each account (a total of 50k accounts and tweets). We perform extensive analysis of the data to identify characteristics of these accounts. Lastly, we experiment extensively with traditional machine learning models such as Support Vector

[1] https://help.twitter.com/en/rules-and-policies/media-policy

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 136–144
Kyiv, Ukraine (Virtual), April 19, 2021.

Machines (SVM) and Multinomial Naive Bayes (MNB), Deep Learning models such as FastText and Contextual Embedding models (BERT). We analyze contribution of each information available (username, user description, or single tweet) to the performance of the models.

Since accounts that post adult-content want to attract others, their usernames and user descriptions are often catchy and contain references that are indicators of them posting adult content. We demonstrate that with just username and user description, we can detect these accounts with macro-averaged F1 score of 94.7%. With addition of single tweet as available information, we achieve macro-averaged F1 score of 96.8%. Detecting accounts who post adult content with minimal information (e.g. from username or description) will allow such accounts to be detected early and possible warning messages can be sent to users to protect them from potential harm or inappropriateness.

The contribution of this work can be summed as: 1) Providing the largest dataset of Twitter accounts that is manually annotated for adult content detection in Arabic, and we make it available for researchers. 2) Exploring the dataset to learn silent features used in the domain as well as features related to users and their profiles. We show that user information can be used for early detection of adult accounts even before tweeting, and when they are combined with tweet text, results are improved. 3) Evaluating a number of machine learning and deep neural approaches for classification of adult content.

The paper is structured as follows: In section 2, we discuss related work in the field. In section 3, we describe the data collection method and present analysis of the data. In section 4, we present our experimental setups and results. In section 5, we examine features learned by our best model and perform error analysis that provides insight on how to improve the data and models in the future. Lastly, in section 6, we present conclusions of our work.

2 Related Work

Despite the fact that many social media platforms enforce rules and conditions about the content shared on their platforms, malicious users attempt to circumvent these rules and guidelines. Researchers have attempted different approaches for exposing malignant content. Spam detection in particular has gained a lot interest among researchers (e.g., (Po-Ching Lin and Po-Min Huang, 2013; Yang et al., 2013; Herzallah et al., 2018; Grier et al., 2010; Mubarak et al., 2020)). Spam detection is a generalized approach for detecting unsolicited messages. Our focus in this paper is on the more concentrated field of detecting adult-content, which categorically includes pornographic references.

For English language, there is a number of works devoted to adult-content detection in terms of analyzing the social networks or the content itself. Mitchell et al. (Mitchell et al., 2003) study exposure to adult content and its relation to age/gender. Singh et al. (Singh et al., 2016) propose Random Forest classifier to detect pornographic spammers on Twitter. Cheng et al. (Cheng et al., 2015) propose an iterative graph classification technique for detecting Twitter accounts who post adult content. Harish et al. (Yenala et al., 2017) study deep learning based methods for detecting inappropriate content in text.

In Arabic, however, the field of adult-content detection is still relatively unexplored. A related field that has been explored recently in Arabic is abusive/hate-speech detection. There has been a few recent works (Mubarak et al., 2017; Albadi et al., 2018; Hassan et al., 2020a,b) in the areas of offensive and hate-speech detection. However, offensive language and hate-speech have few fundamental differences with adult-content. While offensive language and hate-speech typically consist of profanity and attack on individuals or groups, adult-content may contain profanity but primarily consist of pornographic references. More concentrated work on adult-content detection have been conducted by (Alshehri et al., 2018; Abozinadah et al., 2015; Abozinadah and Jones, 2017). In (Alshehri et al., 2018), a list of hashtags was used to automatically construct dataset of tweets that contain adult content. In (Abozinadah et al., 2015), 500 Twitter accounts were manually annotated for adult-content posts. Both (Abozinadah et al., 2015; Alshehri et al., 2018) use traditional machine learning models such as Support Vector Machine (SVM) or Multinomial Naive Bayes (MNB) for classification. Using statistical features of tweet text for classification was proposed in (Abozinadah and Jones, 2017). Although (Alshehri et al., 2018) perform some analysis of screennames, they do not use them or any other user information for classification. While (Abozinadah et al., 2015) explore number of tweets, followers and following by the

accounts, they do not utilize username or user description either. These works rely on collection of tweets from each user for classification.

3 Data Description

We describe the method used to collect the dataset, some statistics and observations about it including most frequent words, emojis and hashtags. We show also the geographical distribution of Adult accounts and some differences between our dataset and previous datasets.

3.1 Data Collection

It is common for users on Twitter to describe themselves by providing a header (username), a short bio (description) and a location in their profiles. We noticed that many Arabic speaking profiles that post adult content declare their location in terms of the country or the city that they are from. They use this information mainly to describe themselves and/or to communicate with other users. This information can be found in username, user location or user description. Alshehri et. al in (Alshehri et al., 2018) reported that it's common for user names to have city or country names (ex: سالب الرياض (bottom from Riyadh)) but in fact this is observed in other profile fields as well. Figure 1 shows sample of profiles for artificial adult accounts where city or country names frequently appear in any of profile fields.

To build a list of country and city names, we obtained all Arabic country names written in either Arabic, English, or French and their major cities from Wikipedia[2], and we added adjectives specifying nationalities in masculine and feminine forms, for example: مصر، بيروت، عراقي، مغربية (Egypt, Beirut, Iraqi (m.), Moroccan (f.)) and so on. We call this list "CountryList".

We used Twitter API to crawl Arabic tweets in March and April 2018 using language filter ("lang:ar"). During this period, we collected 25M tweets from which we identified all users who posted these tweets. We considered only accounts that contain any entry from CountryList in their profile fields. By doing so, we obtained a list of 60k accounts and one random tweet from each user. As an initial classification, we provided the result as obtained from the best system reported by (Mubarak and Darwish, 2019) for detecting vulgar tweets. Then we asked an Arabic native speaker who is familiar with different dialects to judge whether an account can be considered as adult or not based on all available textual information: user profile information, a sample tweet, and the automatic initial classification. Profile pictures or network features (e.g. followers and followees) were not used during annotation and this can be explored in the future. The annotator was allowed to check Twitter accounts in case of ambiguous cases.

Final annotation showed that 6k accounts can be considered as Adult while the rest can be considered as Non-Adult. While the system reported by (Mubarak and Darwish, 2019) achieved F1 = 90 in detecting vulgar language on Egyptian tweets used in communication between users, its performance dropped dramatically due to dialect mismatch and the big differences between vulgar communication and adult content[3].

To conform with Twitter policy that allows sharing up to 50k public tweets and user objects[4], we took all Adult accounts and 44k from the Non-Adult accounts to have a total of 50k accounts and tweets. To verify annotation quality, Two annotators reviewed a random sample of 100 accounts and tweets (50 Adult and 50 Non-Adult), and agreement was 100% and 94% in the Adult and Non-Adult classes respectively. Cohen's kappa (κ) was used to measure the Inter-Annotator Agreement (IAA). The Cohen's κ value was 0.94 ($p\text{-}value < 10e\text{-}5$) which indicates an "Almost Perfect" agreement according to the interpretation of the Kappa value (Landis and Koch, 1977) . Preliminary statistics about the dataset are shown in Table 1, and it can be downloaded from this link: https://alt.qcri.org/resources/AdultContentDetection.zip.

3.2 Analysis

In this subsection, we report some observations about length of Adult and Non-Adult tweets, existence of user mentions, URLs and emojis in both classes, distinguishing words, emojis, and hashtags, etc.

Figure 2-(up) shows that Adult tweets are normally shorter than Non-Adult tweets (9 words ver-

[2] https://en.wikipedia.org/wiki/List_of_countries_by_largest_and_second_largest_cities

[3] Out of 5,854 accounts classified automatically as vulgar, only 825 accounts are manually classified as adult (14%).

[4] https://developer.twitter.com/en/developer-terms/policy

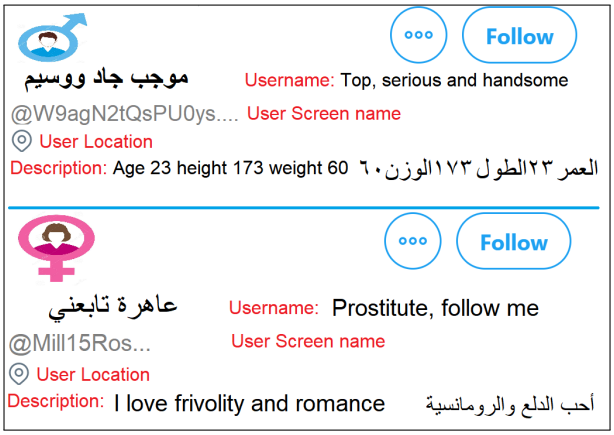

Figure 1: User profile on Twitter for male and female artificial adult accounts

Table 1: Dataset statistics. Tokens and Types (unique Tokens) are calculated for tweet text.

	Tweets (also Accounts)	%	Tokens	Types
Adult	6k	12%	59k	19k
Not Adult	44k	88%	707k	195k
Total	50k		766k	201k

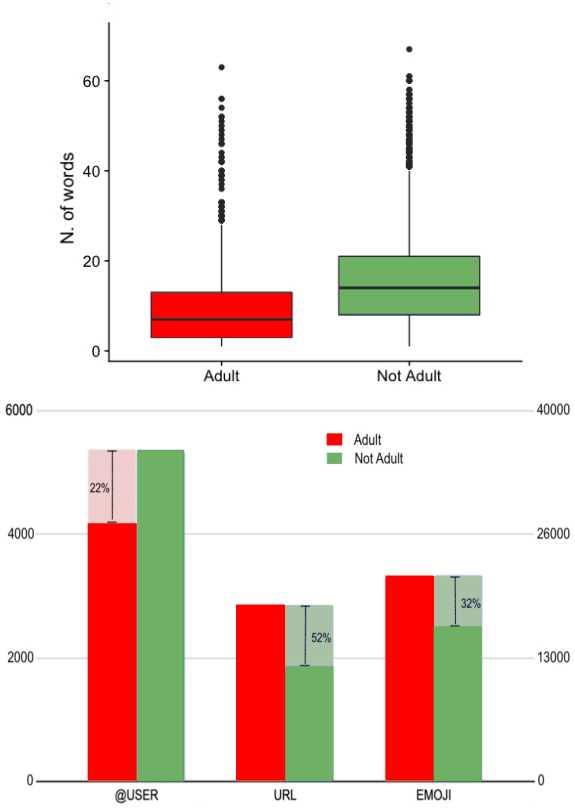

Figure 2: Features comparison between Adult and Non-Adult classes. Boxplot (up) shows the length distributions in both classes. Barplots (down) show the counts for @USER mentions, URLs and emojis per class.

sus 15 words). While Adult tweets tend to have few words to tell users to contact in private or to look at external movies or pictures, Non-Adult tweets normally talk about news, stories or opinions that need more words to describe. This is a significant difference that could highlight the differences in writing style.

This is also confirmed by Figure 2 for the "@USER" mentions. They are less common in the Adult tweets. Typically these tweets are not directed to specific persons but are more an attempt to reach a broad audience. In contrast to Adult tweets, there are a large number of Non-Adult tweets that reference specific @USER either in a response or as a mention. We also observe that, in contrast to Non-Adult tweets, Adult tweets use almost 52% more URLs and 32% more emojis.

Diving further in our analysis, we would like to investigate the different words and emojis that discriminate each class; for such, we will employ the valence score (Conover et al., 2011) in this analysis. The valence score $\vartheta(t)_C$ helps determine the importance of a given word/symbol t in a given class C while considering its presence or absence in other classes. This includes all tokenized words and symbols. Given $freq(t, AD)$ and $freq(t, NA)$ representing the frequency of the term t in Adult and Non-Adult classes respectively, the valence is computed as follows:

Figure 3: Top 10 emojis for Adult class with valence score $\vartheta(.) >= 0.98$.

Figure 4: Word cloud for Adult (left) and Non-Adult (right) user information. Most Adult words are related to genitals and sexual actions while most of the Non-Adult words are related to religion, politics, sports, etc.

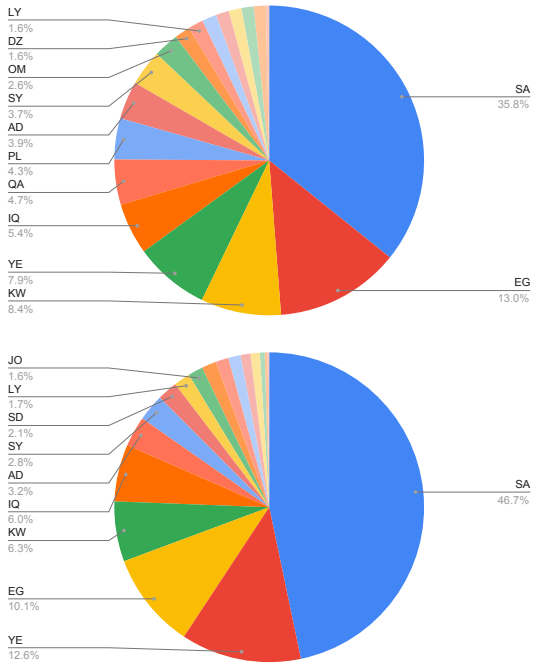

Figure 5: Distribution of the countries for all accounts (up) and for Adult Accounts (down).

$$\vartheta(t)_{AD} = 2\frac{\frac{freq(t,AD)}{N(AD)}}{\frac{freq(t,AD)}{N(AD)} + \frac{freq(t,NA)}{N(NA)}} - 1 \quad (1)$$

Where $N(AD)$ and $(N(NA))$ are the total number of occurrences of all vocabulary in the Adult and Non-Adult classes respectively.

Using Equation 1, we computed the prominence - valence score- for emojis and words in both classes. Figure 3 shows the top most frequent emojis in Adult class, and Figure 4 shows top most frequent words in both classes.

Figure 5 shows the geographical distribution of all accounts in the dataset and Adult accounts as obtained from self-declaration in user profile (user location, username, or user description). We use ISO 3166-1 alpha-2 for country codes[5]. As 36% of all accounts in our dataset are from Saudi Arabia (SA), it was expected also to find the largest number of Adult accounts (2801 accounts, 47% of all Adult accounts) to come from the same country.

[5] https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

We extracted the hashtags that have a valence score of 1 (appear only in the Adult class). The top 150 hashtags list can be downloaded from the same data link: https://alt.qcri.org/resources/AdultContentDetection.zip.

It is worth mentioning that from the 100 seed hashtags used in (Alshehri et al., 2018) to collect adult tweets, we found 37 hashtags that are common between the two lists. We found some noisy hashtags (not necessary to be used in adult tweets) in the seed hashtags from (Alshehri et al., 2018) such as: #طفلة, #عمانيه, #عرب, #زواج, #سكايب (#baby_girl ,#Skype, #marriage, #Arabs, #Omani(f.)), while very strong hashtags such as: #جنس, #ليزبيان (#sex, #lesbian) are missed. We believe that our obtained list of adult hashtags are more accurate and diverse and can be used to extract larger and accurate adult tweets.

4 Experiments

To train classifiers for automatic detection of Adult tweets, we split the data into training set (70%), development set (10%), and a test set (20%).

We experiment extensively with 1) different classifiers, and 2) dataset variants with different degree of information available about accounts. Although we conducted experiments on different preprocessing techniques such as removing diacritics or normalizing Arabic text, we did not notice any significant improvement (between 0.1%-0.2%) in performance. We omit these experiments to make room for more significant results.

4.1 Classifier Description

We conduct our experiments with traditional machine learning classifiers Support Vector Machines (SVM) and Multinomial Naive Bayes (MNB), Deep learning based model FastText (Joulin et al., 2016), and contextual embedding models BERT-Multilingual (Devlin et al., 2019) and AraBERT (Antoun et al., 2020). FastText and MNB were seen to be outperformed by the other three classifiers. For compactness, we only include the top three classifiers along with the baseline model in our discussion and results.

4.1.1 Baseline Model

Our baseline model simply predicts the majority class, Non-Adult, for every instance. Purpose of the baseline model is to simply act as a reference point for the other classifiers we experimented with.

4.1.2 Support Vector Machines (SVM)

To train our SVM models, we use scikit-learn library[6].We transform the input text to character and word n-grams vectors using term frequencies (tf)-inverse document frequencies (idf) vectorizer. We experiment with different ranges of n-grams for character and words. We experiment by training the SVM 1) on only character n-gram vectors, 2) on only word n-gram vectors, and 3) on both character and word n-gram vectors stacked together. We experimented with ranges from [2-2] to [2-6] for character n-grams and from [1-1] to [1-5] for word n-grams. We found that the results did not improve beyond [2-4] for character n-grams and [1-2] for word n-grams. Only the best results are reported in Table 2 and Table 3.

We also experimented with the pre-trained Mazajak word embeddings (skip-gram model trained on 250M tweets) (Abu Farha and Magdy, 2019) as input features for the SVM. Due to its lower performance compared to the n-gram features, we omit these results from the paper.

4.1.3 Multilingual BERT

Deep contextual embedding models such as BERT (Devlin et al., 2019) have been seen to outperform many other models for Natural Language Processing (NLP) tasks. Multilingual BERT is a BERT-based model pre-trained on Wikipedia text of 104 languages that includes Arabic. We fine-tune the model for the task of adult content detection by running it for 4 epochs on the training data with learning rate of 8e-5 using ktrain library (Maiya, 2020).

4.1.4 AraBERT

AraBERT (Antoun et al., 2020) is a BERT-based model specifically trained for Arabic language. The model is pre-trained on Arabic Wikipedia and news articles from various sources. Similar to Multilingual BERT, we fine-tune AraBERT for 4 epochs with learning rate of 8e-5 using ktrain library (Maiya, 2020).

4.2 Dataset Variants

One of our primary goals is to understand how much information is required to detect accounts who post Adult tweets. To achieve this, we examine contribution of different information available about the Twitter accounts. We also evaluate the

[6]https://scikit-learn.org/

Table 2: Performance on user information

model	feats.	screen_name				username				user description			
		P	R	F1	mF1	P	R	F1	mF1	P	R	F1	mF1
baseline	-	0.0	0.0	0.0	46.7	0.0	0.0	0.0	46.7	0.0	0.0	0.0	46.7
SVM	c[2-4]	53.1	**12.6**	**20.4**	**56.9**	**93.8**	65.3	**77**	**87.1**	94.8	61.6	74.7	85.9
SVM	w[1-2]	0.0	0.0	0.0	46.7	89.0	61.8	72.9	84.9	93.0	58.1	71.5	84.2
SVM	c[2-4], w[1-2]	**81.5**	7.9	14.3	54.1	91.2	**66.2**	76.7	87.0	**94.8**	62.3	75.2	86.2
Multi-BERT	-	0.0	0.0	0.0	46.7	85.6	64.2	73.4	85.1	90.1	**65.8**	76	86.6
Ara-BERT	-	0.0	0.0	0.0	46.7	85.1	65.34	73.9	85.4	92.1	64.8	**76**	**86.6**

Table 3: Performance on tweet and combination of tweet + user information

model	feats.	username+user description				tweet				username+user description+tweet			
		P	R	F1	mF1	P	R	F1	mF1	P	R	F1	mF1
baseline	-	0.0	0.0	0.0	46.7	0.0	0.0	0.0	46.7	0.0	0.0	0.0	46.7
SVM	c[2-4]	**96.6**	84.7	90.3	94.5	**88.5**	70.9	78.7	88.0	**96.3**	91.7	94.0	96.6
SVM	w[1-2]	92.2	82.5	87.1	92.7	85.5	71.3	77.8	87.5	93.5	90.1	91.8	95.3
SVM	c[2-4], w[1-2]	96.1	85.8	**90.7**	**94.7**	87.4	74.5	**80.4**	**88.9**	95.3	93.0	94.1	96.6
Multi-BERT	-	93.4	86.4	89.8	94.2	83.4	73.8	78.3	87.7	94.4	92.6	93.5	96.3
Ara-BERT	-	91.1	**88.3**	89.7	94.1	82.2	**76.1**	79.1	88.1	94.7	**94.0**	**94.4**	**96.8**

classifier when combinations of these information are made available.

4.2.1 Individual Information

We compare performance of the classifiers when they have access to only i)username, ii) screen_name, iii) user description, or iv) single tweet from an account.

4.2.2 Combination of Information

We give the classifiers access to increasingly more information to evaluate how their performance change. We notice that addition of screen_name does not contribute to any improvement in performance, and thus, it is excluded from our discussion. We discuss change in performance when i) other user information (username and user description) is combined and, ii) the user information is combined with a single tweet from the account. To combine information, we concatenate the strings representing user information and the single tweet.

4.3 Experiment Results

In Table 2 and Table 3, we present results of different models on variants of information available. We report precision (P), recall (R), and F1 for the Adult class on the test set. We also report the macro-averaged F1 (mF1), i.e. average of F1 for the Adult and Non-Adult classes because the data is not balanced. We use mF1 metric for comparison in our discussion. The key findings are listed below.

- Among the different individual user information available (username, screen_name, description), usernames of Twitter accounts carry most importance. From usernames alone, SVMs trained with character n-gram features achieve mF1 score of **87.1**, an increase of **40.4** from baseline (46.7). Screen_name has very little importance as it increases mF1 by only 10.2 from baseline. User description alone results in mF1 score of 86.6 with AraBERT model.
- When username and user description are combined, we get a notable spike in performance —mF1 score of **94.7**, an increase of **48** from baseline. This is achieved by SVM when character and word n-grams are combined.
- From a single tweet, the maximum mF1 score achieved is **88.9**, an increase of **42.2** from baseline. This is also achieved by SVM with character and word n-gram vectors as features.
- When a single tweet is added to username and user description, the maximum mF1 score achieved is **96.8**, an increase of **50.1** from baseline and an increase of **2.1** from user infor-

Table 4: Confusion matrix of AraBERT model

		Predicted	
		Adult	**Non-Adult**
Reference	**Adult**	1161	74
	Non-Adult	65	8700

mation alone. This is achieved by AraBERT model and is our best-performing model.

- SVM trained on word n-gram features alone is outperformed by other classifiers in all cases. It's behind by about **2** in mF1 score compared to the best system in each case. This suggests character-level/contextual information are important for detecting adult content.
- SVMs trained on character n-gram, combination of character and word n-gram, MultiBERT and AraBERT are very close to each other. For example, in the case of username+user description+tweet, the maximum difference between their performances is **0.5**.

5 Error Analysis

The confusion matrix of predictions by our best system, AraBERT trained on user information+tweet, is shown in Table 4. We manually analyzed all classification errors and these errors can be summarized as follows:

Non-Adult accounts that are detected as Adult: this occurred 65 times. We found that in 70% of these cases, they were annotated incorrectly in the reference, for example when an account has كاش وجاد (cash and serious) in either user information or tweet text, this account should be marked as Adult as such term is commonly used by Adult accounts. This suggests that automatic classification can be used iteratively to detect possible annotation errors. The rest of the errors were due to the existence of frequently-used words in Adult accounts such as مساج (massage) but these words can be used also by Non-Adult accounts.

Adult accounts that are detected as Non-Adult: this occurred 74 times. Only 7 of these cases were due to errors in the reference annotation while majority of errors were due to: i) using unseen words in the training data (ex: creative spelling of some dialectal adult words); ii) complex cases where combining features in user profiles can intuitively reveal adult accounts to human annotators, e.g. when a screen_name is "K3Eut8i8t3pFMy..." and the user described himself as رومانسي فوق العادة (extraordinary romantic) and the tweet is an invitation to come in private. For classifiers, it maybe difficult to capture such complex intuition.

6 Conclusion

We presented a dataset for detecting Twitter accounts who post adult content in Arabic tweets. We performed extensive analysis of the data to identify characteristics of such accounts. In our experiments, we have shown that Support Vector Machines and contextual embedding models AraBERT and Multilingual BERT can detect these accounts with impressive reliability while having access to minimal information about the accounts. In the future, we aim to explore if similar methods can be adopted to identify accounts who post other variants of undesirable content such as unsolicited advertisement. Also, we plan to experiment tools that detect adult content in multimedia (e.g. in images) and compare performance with our model that depends only on textual information.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild. *arXiv preprint arXiv:2005.06557*.

Ehab Abozinadah, Alex Mbaziira, and James Jr. 2015. Detection of abusive accounts with arabic tweets. volume 1.

Ehab A. Abozinadah and James H. Jones. 2017. A statistical learning approach to detect abusive twitter accounts. In *Proceedings of the International Conference on Compute and Data Analysis*, ICCDA '17, page 6–13, New York, NY, USA. Association for Computing Machinery.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Florence, Italy. Association for Computational Linguistics.

N. Albadi, M. Kurdi, and S. Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

Ali Alshehri, El Moatez Billah Nagoudi, Hassan Alhuzali, and Muhammad Abdul-Mageed. 2018. Think before your click: Data and models for adult content in arabic twitter.

Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *ArXiv*, abs/2003.00104.

H. Cheng, X. Xing, X. Liu, and Q. Lv. 2015. Isc: An iterative social based classifier for adult account detection on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1045–1056.

Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. *ICWSM*, 133:89–96.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, page 27–37, New York, NY, USA. Association for Computing Machinery.

Sabit Hassan, Younes Samih, Hamdy Mubarak, and Ahmed Abdelali. 2020a. ALT at SemEval-2020 task 12: Arabic and English offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1891–1897, Barcelona (online). International Committee for Computational Linguistics.

Sabit Hassan, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Ammar Rashed, and Shammur Chowdhury. 2020b. Alt submission for osact shared task on offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (LREC 2020)*, page 61–65.

Wafa Herzallah, Hossam Faris, and Omar Adwan. 2018. Feature engineering for detecting spammers on twitter: Modelling and analysis. *Journal of Information Science*, 44(2):230–247.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv*, arXiv:2004.10703 [cs.LG].

Kimberly J Mitchell, David Finkelhor, and Janis Wolak. 2003. The exposure of youth to unwanted sexual material on the internet: A national survey of risk, impact, and prevention. *Youth & Society*, 34(3):330–358.

Hamdy Mubarak, Ahmed Abdelali, Sabit Hassan, and Kareem Darwish. 2020. Spam detection on arabic twitter. In *Social Informatics*, pages 237–251, Cham. Springer International Publishing.

Hamdy Mubarak and Kareem Darwish. 2019. Arabic offensive language classification on twitter. In *International Conference on Social Informatics*, pages 269–276. Springer.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Po-Ching Lin and Po-Min Huang. 2013. A study of effective features for detecting long-surviving twitter spam accounts. In *2013 15th International Conference on Advanced Communications Technology (ICACT)*, pages 841–846.

Monika Singh, Divya Bansal, and Sanjeev Sofat. 2016. Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining*, 6.

Chao Yang, Robert Harkreader, and Guofei Gu. 2013. Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, 8(8):1280–1293.

Harish Yenala, Ashish Jhanwar, Manoj Chinnakotla, and Jay Goyal. 2017. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*.