Think Before You Click: Data and Models for Adult Content in Arabic Twitter

Ali Alshehri[1], **El Moatez Billah Nagoudi**[2], **Hassan Alhuzali**[3], **Muhammad Abdul-Mageed**[3]
[1] SUNY at Buffalo
[2] Laboratoire d'Informatique et de Mathmatique LIM, Amar Telidji University
[3] Natural Langauge Processing Lab, The University of British Columbia
alimoham@buffalo.edu, e.nagoudi@lagh-univ.dz, halhuzali@alumni.ubc.ca, muhammad.mageeed@ubc.ca

Abstract

Given the widespread use of social media and their increasingly impactful role in our lives today, there is a pressing need to ensure their safety of use. In particular, various social groups view the spread of adult content in social networks as undesirable. This content may even pose a serious threat to other vulnerable groups (e.g. children). In this work, we develop a unique, large-scale dataset of adult content in Arabic Twitter and provide in-depth analyses of the data. The dataset enables us to study the scope and distribution of adult content in the Arabic version of the network, thus possibly uncovering target phic locales. In addition, computationally exploit the data to learn a large lexicon specific to the topic and detect spreaders of adult content on the microblogging platform. Our models achieve promising results, reaching 79% accuracy on the task (24% higher than a competitive baseline).

1. Introduction

Social media continues to play an increasingly important role in our lives, making it necessary to keep these platforms safe and free from 'undesirable' content. Undesirable postings come in many forms, including deceptive (Westerman et al., 2014), hateful (Williams and Burnap, 2015), abusive (Mubarak et al., 2017), dangerous (Fuchs, 2017; Sikkens et al., 2017), and adult content (Abozinadah, 2015). Identification of spreaders of unsolicited content is beneficial not only for user satisfaction, but also for the safety of individuals and communities alike.

In the Arab world, social media are widely used (Lenze, 2017). This is especially the case for the Twitter platform where, according to some estimates (Salem, 2017), the number of monthly active users was expected to be 11.1 million as of March 2017. These Arab users send 27.4 million tweets per day, almost doubling up from 5.8 million in 2014 (Salem, 2017). Twitter has also been a very influential tool in the Arab world, as is evident from its role in the waves of uprisings the region. In the contexts of the political and social transformations the Arab world has witnessed, activists have heavily used the platform for disseminating views antagonistic to several Arab governments (Khondker, 2011; Gerbaudo, 2012). Similarly, governments themselves are increasingly using Twitter to spread content supporting their causes (i.e., propaganda) (Mejova, 2017).

Twitter prohibits the promotion of adult or sexual products, services, and content, whether in images, videos, or text. [1]However, spreaders of undesirable content are exploiting Twitter's popularity, and it is not uncommon to even witness advertising and adult content hashtags trending (Herzallah et al., 2017).

Popular search engines such as Google and Yahoo provide "safe search" options to filter out unwanted content. Social media platforms (e.g. Twitter, Facebook, YouTube) also offer similar options, yet seem to be fighting a more difficult battle. Efforts to combat unsolicited content, however, does not seem to be very successful thus far, as we will show. Depending on manually curated lists of words for use in filtering out adult content is no longer sufficient since language and techniques employed by spreaders of these content are constantly evolving. For example, spreaders of adult content often intentionally employ misspelled and/or slang words. Misspellings can be as simple as replacing the letter 'o' with the digit '0' in a word, which can enable these users to bypass Twitter's algorithmic filters.

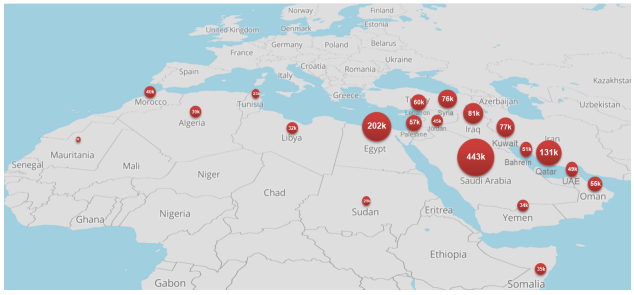

Figure 1: Geographical distribution of adult content in the Arab world.

Filtering out adult content is perhaps especially valuable in the Arab world, due to religious and cultural sensitivities. In this work, we seek to alleviate this bottleneck for Arabic social media. We make the following contributions: (1) we build a large-scale dataset of Arabic adult content; (2) we learn large-scale lexica (based on hashtags, unigrams, and bigrams) correlated with adult content from the data; (3) we perform an in-depth analysis of the data, thus affording a better understanding of the dynamics of adult content sharing and the behavior of its users on Twitter; and (4) we develop successful predictive models for detecting spreaders of adult content.

The remainder of the paper is organized as follows: In Section 2, we review related literature. We describe our dataset in Section 3, we perform several textual analyses of the data and describe learning a lexicon of adult content in Section 4. In Section 5, we describe our models for detecting adult content. Section 6 concludes the paper with our main find-

[1] https://support.twitter.com/articles/20170427?lang=en

ings.

2. Related work

Unsolicited Content on Twitter. Undesirable content can be prevalent in Twitter. The network is indeed vulnerable to misuse through posting of undesirable content such as spams, racist content, hateful speech, threats, and adult content. This is due to the fact that creating and maintaining an account on Twitter is fairly easy. Unlike Facebook, where anonymity is at least theoretically not possible, anonymity is easier on Twitter. This possibly translates to more undesirable content. The work of Grier et al. (2010) is relevant to the scope of unsolicited or spam content on Twitter. The authors studied 25 million URLs posted on Twitter and found that 8% of content in these URLs are spam. Analyzing the click-through rate of those spam tweets, they found that around 0.13% of them generate a site visit. This rate is much higher than the click-through rate reported for spam emails (Kanich et al., 2008). This implies that the number of spammers on Twitter is increasing over time.

Racist and Hateful Speech. A number of studies have attempted to investigate racists and hateful speech in the web as well as Twitter. For example, Burnap and Williams (2014) look at the manifestation and diffusion of hate speech and antagonistic content in social media in relation to events that could be classified as 'trigger' events for hate crimes. Their dataset consists of $450k$ tweets collected a two weeks window in the immediate aftermath of Drummer Lee Rigbys murder in Woolwich, UK. Using n-gram and type-dependency features, they implemented probabilistic, rulebased, and spatial classifiers. The authors reported a best F-score of 0.77 using the probabilistic classifier. Similarly, Davidson et al. (2017) created a hate speech lexicon based on a list of phrases and words provided by *Hatebase.org*. Using this list, they crawled a set of $85m$ tweets containing terms from the lexicon. Then, a random set of $25k$ tweets were manually annotated by CrowdFlower users on three categories: hate speech, offensive, and neither. They used Logistic Regression and a dictionary to construct a predictive hate and offensive language model, which achieved an F1-score of 90%.

Adult Content. Some studies were also devoted to investigating and detecting adult content online. For example, Coletto et al. (2016) analyzed 169 million data points on Tumblr and Flickr and found that although the community of adult content producers is small, adult content is spread widely in the networks. While producers of adult content are clustered in semi-isolated communities on these platforms, they are linked with the rest of the network with a very high number of what Coletto et al. (2016) called "consumers" (users who do not post new adult content but follow producers of such content, share and like their posts). The authors maintained that, due to the fact that users in the network are enabled to see what other users 're-post' or 'like,' over a quarter of the all Tumblr users were unintentionally exposed to adult content. The case is no different in Twitter where users are able to see recently liked tweets by users they follow. Singh et al. (2016) estimated at least 10 million accounts tweeting and spreading adult content according as of May 2015.

Singh et al. (2016) employ graph- and content-based features extracted from $74k$ tweets posted by $18k$ Twitter users on the same task, reporting 91.96% accuracy. Their analysis shows that adult content users fulfill the characteristics of spammers as stated by the rules and guidelines of Twitter[2]. These pioneering works, however, focused on detecting adult or spam content, without providing analyses of the content itself. Our work exploits a much bigger dataset (e.g., our dataset is about eight times bigger than (Abozinadah, 2015)), and pays attention to especially the geographical distribution of targets of the adult content.

Twitter Spam. What increases Twitter users's exposure to pornographic tweets is also the fact that trending hashtags are usually exploited by spammers (Abozinadah, 2015; El-Mawass and Alaboodi, 2016). This vulnerability of Twitter users has recently led to a number of studies focusing on analyzing and detecting Twitter 'spams' (e.g. (Lin and Huang, 2013; Yang et al., 2013; Wahsheh et al., 2012b; Wahsheh et al., 2013; Herzallah et al., 2017; Chu et al., 2012; Grier et al., 2010; El-Mawass and Alaboodi, 2016; Singh et al., 2016)). A few of these studies were dedicated to spam detection in Arabic social media (e.g. (Wahsheh et al., 2012a; Wahsheh et al., 2012b)).

Adult Content in Arabic. Early work on Arabic social media has focused on developing corpora and systems for detecting sentiment (Abdul-Mageed and Diab, 2012; Abdul-Mageed and Diab, 2011; Abdul-Mageed et al., 2014), aided by automatic processing tools developed for the language like ASMA (Abdul-Mageed et al., 2013), and later emotion (Abdul-Mageed et al., 2016). More related to our work is research by Abozinadah (2015) and Singh et al. (2016) who focused on detecting adult content on Arabic and English Twitter, respectively. Abozinadah (2015) and Abozinadah and Jones (2017) built a dataset of $1,000,300$ tweets comprising the most recent 50 tweets of 255 users as well as the most recent 50 tweets of users in their network. The authors then develop a machine learning classifier using different feature sets. They found that lexical features yield the best performance. As feature input to their classifiers, the authors extracted basic statistical measures from each tweet (e.g., average, minimum, maximum, standard deviation, and the total number of URLs, hashtags, picture, mentions, and characters). They reported 96% accuracy of adult content detection.

3. Dataset

We collect a large dataset of tweets with adult content. In addition, we identify a large network of adult content producers (who are also usually spreaders). We explain our data collection methods in terms of the following steps[3]:

1. **Hashtag seeds:** We start by collecting a list of hashtags[4]. associated with adult content by manually in-

[2] https://support.twitter.com/articles/64986.

[3] Due to the nature of this work, in various places of the paper, we provide examples that involve language that are related to adult content. Although we use academic norms to present the content in appropriate way, reader discretion is advised.

[4] This list can be downloaded from: https://goo.gl/Qcc1wW.

specting several relevant tweets. We iteratively expand the list by adding co-occurring hashtags that clearly communicate adult content. Our final list is composed of 100 hashtags that we manually judge as highly connected to adult content. Example hashtags from this list include سكس (Eng. "sex"), مولعه (Eng. "horny"), and مومس (Eng. "prostitute").

2. **Tweet-level dataset:** We use both the Twitter rest and streaming APIs to crawl tweets employing items from this list of 100 manually developed hashtags described above. Using these crawlers, we acquired a dataset of $\sim$ 27 million tweets. We refer to this dataset as **`main`**. After filtering out retweets and duplicates, we ended up with a total of $200K$ tweets. We refer to this dataset as **`unique`**.

3. **User-level dataset:** We extract all the users who posted one or more of the tweets in the **`main`** dataset and acquire a total of $20,621$ users. We then crawl the timelines of these users, possibly fetching up to $3,200$ tweets from each user. We are able to obtain the timelines of $11,648$ of these users, making the total number of tweets from these timelines around 8.6 million. We could not fetch the tweets of the remaining $8,973$ users for a number of reasons: First, $2,456$ users were suspended during the period between crawling the **`main`** dataset and the timelines. These users represents $\sim$ 11% of all users. Second, 629 users were not found at the time of user data crawling at all. These users most likely have deleted their accounts. The remaining $5,888$ users were found active, but our crawlers failed to fetch their data due either to (a) their accounts being protected[5] or (b) have no tweets at the time of crawling. We call this dataset **`timelines`**. See Table 1 for a summary of the datasets and Table2 for a summary of users in our datasets.

Dataset	Size (tweets)
Main	27 M
Unique	200 K
Timelines	8.6 M

Table 1: Datasets in the study. **Main:** All the tweets we have initially crawled. **Unique:** Tweets from main after deduplication and removal of retweets. **Timelines:** Tweets from our list of unique list of $11,648$ users' timelines.

4. Understanding Adult Content

We use our dataset as a basis for understanding adult content in various ways. First, we build lexica of adult content in the form of hashtags and n-grams (unigrams and bigrams). These can provide a summary of what the involved lexical content is like, but can also be used for collecting adult content in the future for building classifiers. Related results are presented in Section 4.1. Second, we study the posting behaviors of adult content users by aggregating important frequencies from their content. We also present a description of their network structure via simple follower-followee statistics (Section 4.2.). The types of media employed in adult content is another significant aspect of sharing pornography online and hence we also study this aspect of content in Section 4.4. Adult content users also seem to have specific practices as to choosing their screen names on the network. In an attempt to understand these practices, we analyze a sample from our data in Section 4.3. Finally, a question that arises is related to the locales this particular type of business might be targeting and/or most thriving in. In Section 4.5., we perform an analysis that answers this exact question. We now turn to describing our findings related to each of these user and content attributes.

Type of user	Freq.	%
Active (collected)	$11,648$	56.5%
Active (not collected)	$5,888$	28.5%
Suspended	$2,456$	11.9%
Not found	629	3.1%

Table 2: Types, counts, and percentages of users in our **`timelines`** datasets.

4.1. Lexica of Adult Content

4.1.1. Hashtags

We extract all the hashtags with frequency $>$ 20 in the dataset, acquiring a total of $21,907$ hashtags. A sample from the extracted hashtags is in Table 3. The range of hashtags are related to descriptions of explicit content that may be accessible via a shared URL in a tweet, a range of pornographic activities, and references to individuals with different sexual orientations. The lexicon can be used as a basis for monitoring online adult content and collecting even larger data for detecting pornography.

4.1.2. N-grams

We also extract all unigrams and bigrams with frequencies $>$ 20 from the dataset, acquiring a total of $128,625$ unigrams and $243,953$ bigrams. Table 3 shows a sample of each of these types[6]. Similar to the hashtag lexicon, the N-gram lexicon exposes a range of activities related to adult content, but also clickbait where users are asked to click on a link to watch adult video or see an explicit photo. This clearly paints a picture of adult content marketing as a business.

4.2. User Timelines

For a deeper understanding of the behaviour of adult content spreaders, we calculate several measures based on our **`timelines`** dataset. These measures include the average, median, and mode of (1) total tweets posted per user, (2) total pornographic hashtags employed by a user, (3) avergae

[5]Protected users can only be crawled when the authenticated user crawling the data either "owns" the timeline or is an approved follower of the owner. None of these applied to us.

[6]The lists of all hashtags, unigrams and bigrams with their frequencies can be downloaded from: https://goo.gl/LVig9g.

Hashtag		Uingram		Bigram	
AR	**EN**	**AR**	**EN**	**AR**	**EN**
#سكس	#sex	هنا	here	هنا #سكس	#sex here
#ز*ك	#f*ck	ز*ك	f*ch	الفيلم كامل	full movie
#فحل	#bull	سكس	sex	شاهد وحمل	watch and download
#ممحونه	#divorced	الفيلم	movie	الفيلم هنا	movie here
#ديوث	#cuckold	اضغط	click	كامل هنا	full here
#محارم	#incest	شاهد	watch	ثم اضغط	then click
#افلام_سكس	#sex_movies	خاص	private	#روابط_سكس #سكس	#sex #sex_links
#ز*	#pe*is	كامل	full	اضغط الرابط	ckick the link
#سالب	#bottom	الرابط	link	اضغط على	click on
#ق*ه	#b*tch	ين*ك	fu*king	#سكس #ز*ك	#sex #fu*k

Table 3: A Sample of our Adult Content Lexica. Hashtags (left), unigrams (middle), and bigrams (right).

hashtags used per tweet, and (4) number of friends and followers per user. As Table 4 shows, an average adult content user posts ~ 914 tweets, uses 1.45 hashtags per tweets, and has $\sim 7,489$ friends and 850 followers in their network. These statistics show that spreaders of adult content not only employ hashtags as a mechanism of reaching wider audiences, but also as a way to adhere to Twitter regulation about pornographic content. The analysis also reveals that these users are not silos in the network, but rather have friends and followers.

	Mean	**Median**	**Mode**
Total tweets	914.20	235	10
Total hashtags used	1,370.91	525.50	28
Hashtags per tweet	1.45	0.35	0
Friends	7,488.70	252	0
Followers	850.30	72	0

Table 4: Descriptive statistics of adult content and user network in our data.

4.3. Screen Names Analysis

We wish to investigate screen names used by adult content users. To do so, we first randomly sampled 100 adult users and manually analyzed their screen names. We found out a number of interesting patterns. As shown in Table 5, the most common screen name pattern consists of one or more (e.g., age, physical) attributes. For example, in عشريني وسيم (EN: "a handsome twenties aged guy") there are two adjectives describing both the age and physical attributes of the user. For another example, in المتغطرس (EN: "the arrogant one"), the user chooses to describe his psychological attributes that imply power and pride. In addition, about 60% of those include more pronounced physical attributes with clear sexual meanings and an indication of user gender. Examples include جادة ممحونة (EN: "horny and serious female"), مربرب مشعر (EN: "chubby and hairy male"), and فحل عنيف (EN: "violent and potent male"). Other common screen names are person names, some of which also contain attributes such as أمل ك* مفتوح (EN: "Amal open vag*na") and مجودي بوث (EN: "Majoodi bisexual"). It is also not uncommon for screen names to have city or country names such as سالب مصرى القاهره (EN: "Egyptian bottom Cairo") and سالب الرياض (EN: "bottom from Riyadh"). Some users use their email, phone, or social media account addresses as their screen names. Finally, some screen names do not seem to follow any specific patterns. Instead, they contain numbers, commas, underscores, symbols or mixture of these without any apparent meaning such as '-' and '//-//'. To further analyze adult users screen names, we extract unigrams, bigrams and emoji from all screen names. Table 6 provides a list of the top 10 unigrams, bigrams, and emoji employed by these users. It is clear from the Table that adult content users tend to employ screen names with sexual connotations. We also investigated which exact language is used in screen names. We found that about 66% of these names consist of either Arabic alphabet exclusively or a mixture of Arabic and Roman alphabet. About 29% employ Roman alphabet only. The rest 5% consists of emojis, numbers, symbols, or/and alphabet other than Arabic and Roman.

4.4. Tweet media

We also analyze the use of media in the tweets posted by adult content spreaders. This helps us answer questions like: "What is the rate of tweets that contain URLs?" and "Which is the most URL type (web page, photo or videos) used?". Table 7 summarizes the results of this analysis. We have noticed that many of the adult content tweets contain links, many of which do not actually lead to what they are advertised to be, specifically adult content (%59.68), but rather other sites but such as news sites or ones related to health and beauty content (e.g., http://healthwabeauty.com/). Interestingly, some links lead to blogs that do not seem to originate from the Arab world. For example, the blog

Type	percentage	Example	English
Attribute	34%	فحل عنيف	Violent and potent
Attribute + city/country	9%	سالب الرياض	Bottom of Riyadh
Email address	2%	a-sa**@**.com	–
Emoji	19%	👅🔥	–
Hashtag	1%	#مقاطع	#clips
Person name	25%	خالد	Khalid
Person name + attribute	5%	مجودي بوث	Majoodi bisexual
Others	19%	//-//	–

Table 5: Types of screen names in a sample of 100 pornographic users

Uingram	EN	Bigram	EN	Emoji
سكس	Sex	ك* مفتوح	Open pus*y (unvirgin)	🔥
مطلقه	Divorced (F)	طي* كبيرة	Big As*	🔞
مفتوح	Open	افلام سكس	Sex movies	🖤
متحرره	Emancipated (F)	من المغرب	From Morocco	✨
هايجة	Horny (F)	سكس محارم	Incest sex	👅
مولعة	Horny (F)	سكس عربي	Arab sex	❤️
فحل	potent (M)	سكس في	Sex in	🫧
طي*	As*	مقاطع سكس	Sex clips	🌈
كبيرة	Big (F)	سكس فون	Phone sex	💋
افلام	Movies	وسيط زواج	Marriage broker	💝

Table 6: Top 10 unigram, bigram, and emojis in screen names used by users (**F:** female; **M:** male).

	Count	%
Web link URLs	6.754M	59.68%
URLs refer to photo	3.166M	27.98%
URLs refer to video	1.310M	11.57%
URLs refer to animated gif	86.973M	0.77%
Total URLs (web link+media)	**11.318M**	**100%**

Table 7: Types of media in tweet URLs in the data.

at https://ecoinsnews.blogspot.com/ focuses on Bitcoin and the encryption market mostly likely directed to English speaking-audience. We also observed that only a small fraction of these sites are ones that solicit subscriptions for one or another of a sex 'service' or sexual content.

4.5. Geographical Distribution

Using our dataset, we analyze the geographical distribution of adult content across the Arab world. For the purpose, we follow a simple method:

1. Initially, we automatically generate a list of Arab countries and cities (we refer to the list as **autocities**) from Google map API[7]. The **autocities** list pertains 22 countries and has a total of 361 cities. **autocities** had several errors (e.g., names in English and Hebrew, neighborhood names instead of the a specific city name, GPS coordinates cities).

2. For this reason, we manually correct this list using the following procedure: For each country in the **autocities**, we keep only Arabic city names and

Country	Freq.	City	Freq.
KSA	443,112	Riyadh	89,232
Egypt	202,795	Jeddah	66,944
Qatar	131,707	Amman	27,651
Iraq	81,517	Makkah	16,133
Kuwait	81,517	Qassim	14,344
Syria	76,948	Dammam	14,251
Lebanon	76,290	Madinah	10,365
Palestine	57,029	Jerusalem	9,345
Oman	55,735	Tabuk	8,690
Bahrain	51,956	Gaza	8,256

Table 8: Top 10 Arab countries and cities matched in the adult content.

[7] https://developers.google.com/maps/?hl=fr.

				regular_content			adult_content		
	#data_points	**acc**	**avg-f**	**prec**	**rec**	**f**	**prec**	**rec**	**f**
BOW	10	0.54	0.42	0.64	0.07	0.12	0.54	0.97	0.69
	50	0.54	0.41	0.77	0.04	0.08	0.54	0.99	0.70
	100	0.55	0.43	0.83	0.06	0.12	0.54	0.99	0.70
	250	0.53	0.38	0.50	0.01	0.02	0.53	0.99	0.69
	500	0.53	0.38	1.00	0.01	0.02	0.53	1.00	0.69
BOM	10	0.76	0.76	0.69	0.92	0.79	0.90	0.63	0.74
	50	0.77	0.77	0.69	0.94	0.80	0.92	0.63	0.75
	100	0.78	0.78	0.70	0.94	0.80	0.92	0.64	0.76
	250	**0.79**	**0.78**	**0.70**	0.93	**0.80**	**0.92**	**0.65**	**0.76**
	500	0.78	**0.78**	**0.70**	**0.94**	**0.80**	**0.92**	0.64	**0.76**

Table 9: Results from our models for detecting spreaders of adult content on Twitter. We use SVMs in our experiments. **BOW:** bag-of-words models. **BOM:** bag-of-means models.

manually add other cities (replacing, e.g., the English and Hebrew names with Arabic counterparts, and substituting GPS co-oridnates with corresponding cities). For this step, we use Wikipedia[8]. We also search Wikipedia for Arabic city names that are not in the original **autocities** list and add cities we find. The new list covers 22 countries and a total of 488 cities. We call this list **goldcities**[9].

3. Finally, we use **goldcities** to identify the names of countries and cities targeted in the adult dataset, based on simple matching between our goldcities and tweets' unigrams. This allows for identifying the most targeted Arab countries and cities by adult content users. Figure 1 maps the geographical distribution of targets in adult content by country. Table 8 shows the top 10 Arab countries as well as top 10 cities matched in the data. The top two countries are *KSA*[10] and *Egypt*. The city list in Table 8 contains "Qassim" which is a KSA province rather than a city. Observably, 7 cities out of the 10 top mentioned cities are KSA cities. This shows very heavy targeting of KSA cities. The findings about KSA and Egypt is not surprising as these two countries have large Twitter populations, although there may be other reasons these countries are targeted most. Any such potential reasons are outside the scope of our current work, but form the basis of important research questions.

5. Classification

We build supervised models for detecting adult users exploiting the data of these users. For the purpose, we identify $2,500$ users in the adult data such that each has at least 500 tweets. For the negative class (i.e., regular users), we use an equal number of users' data where each user has at least 500 tweets.

5.1. Pre-processing, Data splits, and settings

We randomize the user data from both the positive and the negative classes and remove all the hashtag seeds used to collect the data. For this work, we choose our hyper-parameters beforehand from a small set of choices as we describe next. To facilitate replication and future work under more sophisticated conditions, we split the data into 80% training, 10% development, and 10 % testing so that development data can be used to tune parameters with more advanced experiments. We employ simple SVM classifiers with a fixed vocabulary size of $20K$ words, under two classification conditions:

Bag-of-Words: Where each vector simply represents each word existing in a tweets with a binary value (0 or 1).

Bag-of-Means: We build a word embedding model (Mikolov et al., 2013) exploiting a large in-house dataset of Arabic tweets totaling $> 100m$ data points. For this purpose, we adopt the pre-processing pipeline of (Zahran et al., 2015; Abdul-Mageed et al., 2018), in that we remove any non-unicode characters, normalize *Alif maksura* to *Ya*, reduce all *hamzated Alif* to plain *Alif*, remove all non-Arabic characters. To clean noise, we reduce all letter repetition of > 2 characters to only 2. We build a skip-gram model with 300 dimensions, a minimal word count $= 100$ words, and a window size of 5 words on each side of a target word. For vectorization, we average the word vectors of each tweet, acquiring a 300-dimension bag of means for each data point.

Settings: We develop the classifiers under a number of conditions, pertaining the number of tweets exploited from each user. We use numbers of tweets according to values from the set $\{$*10, 50, 100, 250, 500*$\}$. For these simple classifiers, we use the **scikit learn** [11] SVC implementation.

5.2. Evaluation:

We report in terms of accuracy (acc), precision (prec), recall (rec), and F-score (f). We use a random baseline of 50%, which is also equal to each of the two classes in the data, given that the two classes are balanced. We first performed the experiments on both Dev and Test under the same conditions, but only report on Test here. As mentioned earlier,

[8] https://en.wikipedia.org/wiki/Arab_world.

[9] The **goldcities** list can be downloaded from: https://goo.gl/s3xzpB

[10] Kingdom of Saudi Arabia

[11] http://scikit-learn.org/stable.

we choose to set aside a development set for future replicability and comparisons under more sophisticated experimental conditions.

Table9 presents the results of our model. As the Table shows, the **BOM** conditions perform better, with best accuracy reaching 79% with 250 tweets, significantly (i.e., $p< 0.03$) exceeding the random baseline of 50%. The best **BOM** (250 tweets) classifier reaches 92% of precision on the adult/positive class, with a reasonable recall of 65%. These results show the utility of the simple SVM **BOM** classifier on this task, as opposed to a **BOW**. Even with 10 tweets, the **BOM** classifier performs at 76% acc, reaching a high precision of 90% on the adult users class.

6. Conclusion

In this work, we described a method for collecting a large-scale dataset of adult content in Arabic Twitter. We also described the data we acquired using this method and used the data to understand the tweeting behavior in this safety-important area of online behavior. We also extracted three lexica involving hashtags, unigrams, and bigrams, which we also make available to the community. Analyzing our data also gave us an opportunity to identify the geographical distribution of targets of adult content, which may lead to future important discoveries about the dynamics and market of adult content production and spread. We finally developed simple, yet quite successful, models for detecting spreaders of adult contents on the microblogging platform. Our models achieve 79% accuracy on the task. In the future, we plan to improve our classification models and further investigate the network structure of the adult content spreaders.

7. Acknowledgement

This research was enabled in part by support provided by WestGrid (`https://www.westgrid.ca/`) and Compute Canada (`www.computecanada.ca`).

8. Bibliographical References

Abdul-Mageed, M. and Diab, M. T. (2011). Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th LAW*, pages 110–118. ACL.

Abdul-Mageed, M. and Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914.

Abdul-Mageed, M., Diab, M., and Kübler, S. (2013). Asma: A system for automatic segmentation and morpho-syntactic disambiguation of modern standard arabic. In *Proceedings of RANLP 2013*, pages 1–8.

Abdul-Mageed, M., Diab, M., and Kübler, S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Abdul-Mageed, M., AlHuzli, H., and DuaaAbu Elhija, M. D. (2016). Dina: A multi-dialect dataset for arabic emotion analysis. In *The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*, page 29.

Abdul-Mageed, M., Alhuzali, H., and Elaraby, M. (2018). You tweet what you speak: A city-level dataset of arabic dialects. In *LREC*.

Abozinadah, E. A. and Jones, Jr., J. H. (2017). A statistical learning approach to detect abusive twitter accounts. In *Proceedings of the International Conference on Compute and Data Analysis*, ICCDA '17, pages 6–13, New York, NY, USA. ACM.

Abozinadah, A., M. A. a. J. J. (2015). Detection of abusive accounts with arabic tweets. *International Journal of Knowledge Engineering*, Vol. 1, No. 2.

Burnap, P. and Williams, M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.

Chu, Z., Widjaja, I., and Wang, H., (2012). *Detecting Social Spam Campaigns on Twitter*, pages 455–472. Springer Berlin Heidelberg, Berlin, Heidelberg.

Coletto, M., Aiello, L. M., Lucchese, C., and Silvestri, F. (2016). Pornography consumption in social media. *CoRR*, abs/1612.08157.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

El-Mawass, N. and Alaboodi, S. (2016). Detecting arabic spammers and content polluters on twitter. In *2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*, pages 53–58, April.

Fuchs, C. (2017). *Social media: A critical introduction*. Sage.

Gerbaudo, P. (2012). *Tweets and the streets: Social media and contemporary activism*. Pluto Press.

Grier, C., Thomas, K., Paxson, V., and Zhang, M. (2010). @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pages 27–37, New York, NY, USA. ACM.

Herzallah, W., Faris, H., and Adwan, O. (2017). Feature engineering for detecting spammers on twitter: Modelling and analysis. *Journal of Information Science*, page 0165551516684296.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G., Paxsonf, V., and Savage, S. (2008). Spamalytics: An empirical analysis of spam marketing conversion. In *Proceedings of the ACM Conference on Computer and Communications Security*.

Khondker, H. H. (2011). Role of the new media in the arab spring. *Globalizations*, 8(5):675–679.

Lenze, N. (2017). Social media in the arab world: Communication and public opinion in the gulf states. *European Journal of Communication*, 32(1):77–79.

Lin, P.-C. and Huang, P.-M. (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, Jan.

Mejova, Y. (2017). Seminar users in the arabic twitter

sphere. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings*, volume 10539, page 91. Springer.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Salem, F. (2017). The arab social media report 2017: Social media and the internet of things: Towards data-driven policymaking in the arab world. Vol. 7.

Sikkens, E., van San, M., Sieckelinck, S., Boeije, H., and de Winter, M. (2017). Participant recruitment through social media: Lessons learned from a qualitative radicalization study using facebook. *Field Methods*, 29(2):130–139.

Singh, M., Bansal, D., and Sofat, S. (2016). Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining*, 6(1):41, Jun.

Wahsheh, H., Alsmadi, I., and Al-Kabi, M. (2012a). Analyzing the popular words to evaluate spam in arabic web pages. *IJJ: The Research Bulletin of JORDAN ACM–ISWSA*, 2(2):22–26.

Wahsheh, H. A., Al-kabi, M. N., and Alsmadi, I. M. (2012b). Evaluating arabic spam classifiers using link analysis. In *Proceedings of the 3rd International Conference on Information and Communication Systems*, ICICS '12, pages 12:1–12:5, New York, NY, USA. ACM.

Wahsheh, H. A., Al-Kabi, M. N., and Alsmadi, I. M. (2013). Spar: A system to detect spam in arabic opinions. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, Dec.

Westerman, D., Spence, P. R., and Van Der Heide, B. (2014). Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, 19(2):171–183.

Williams, M. L. and Burnap, P. (2015). Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2):211–238.

Yang, C., Harkreader, R., and Gu, G. (2013). Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, 8, Aug.

Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H. M., Rashwan, M., and Atyia, A. (2015). Word representations in vector space and their applications for arabic. In *CICLing (1)*, pages 430–443.