Abusive Language Detection on Arabic Social Media

Hamdy Mubarak, Kareem Darwish
Qatar Computing Research Institute,
HBKU, Doha, Qatar
{hmubarak,kdarwish}@qf.org.qa

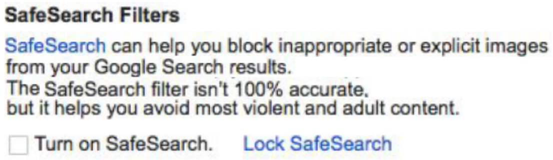
Walid Magdy
School of Informatics,
The University of Edinburgh, UK.
wmagdy@inf.ed.ac.uk

Abstract

In this paper, we present our work on detecting abusive language on Arabic social media. We extract a list of obscene words and hashtags using common patterns used in offensive and rude communications. We also classify Twitter users according to whether they use any of these words or not in their tweets. We expand the list of obscene words using this classification, and we report results on a newly created dataset of classified Arabic tweets (obscene, offensive, and clean). We make this dataset freely available for research, in addition to the list of obscene words and hashtags. We are also publicly releasing a large corpus of classified user comments that were deleted from a popular Arabic news site due to violations the site's rules and guidelines.

1 Introduction

Social media is a popular medium for discussion, expression of views, sharing of content, and promotion of ideas and products. Like any other medium of communication, the content may be clean or obscene/profane or and cordial/polite or offensive/rude. Identification of profane and offensive exchanges on social media can be useful for a variety of applications. For example, users may be interested in filtering out obscenities or indecent content from their social media stream or in filtering out such content for their children. Further, detecting obscene or offensive language in a social media exchange may indicate the discussion of contentious/controversial subjects/content or the presence of hate speech that may be connected to or promoting hate crimes (Watch, 2014). Some sites such as Facebook allows users to filter out content based on a word list that users provide. Similarly, as shown in Figure 1, popular web search engines, such as Google and Bing, and media sharing sites, such as YouTube, have settings for "safe search" that filters out obscenities and pornographic contents. On way to filter out such desirable content is to maintain a list of obscene words to filter content against. However, the manual construction and maintenance of such lists is arduous. This is due to the fact that list curators may not cover all words, particularly country/culture specific ones (written in local dialects or understood in certain cultures) and users may coin new words or alter the spelling of existing words (ex. by replacing letters with similarly looking characters, such as "0" instead of "O").



SafeSearch Filters

SafeSearch can help you block inappropriate or explicit images from your Google Search results.
The SafeSearch filter isn't 100% accurate,
but it helps you avoid most violent and adult content.

□ Turn on SafeSearch. Lock SafeSearch

Figure 1: Google "safe search" setting

Jay and Janschewitz (2008) identified three categories of offensive speech, namely: **Vulgar**, which include explicit and rude sexual references, **Pornographic**, and **Hateful**, which includes offensive remarks concerning peoples race, religion, country, etc. The goal of this work is to detect vulgar and pornographic obscene speech in Arabic social media without the need for manually curating word lists. The detection of offensive language that includes personal attacks, demeaning comments, or hateful language is left for future work. Unlike previous work on obscenity and offensive language detection for different languages, such as

Proceedings of the First Workshop on Abusive Language Online, pages 52–56,
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics

English (Mahmud et al., 2008; Spertus, 2007; Xiang et al., 2012) and German (Ross et al., 2016), very limited previous work for this task was done for Arabic (Abozinadah et al., 2016).

Arabic poses interesting challenges primarily due to the lexical variations of different Arabic dialects. Our approach is concerned with an automated approach to construct of an offensive word list. The approach mines tweets to nominate new obscene words, which can be provided to judges who would either add them to the word list (if obscene) or not. Our approach is based on the intuition that if we can identify users who often use obscene words from a seed word list of obscenities, then by contrasting these users against other users who never use words from the list, we can net additional obscenities. We also introduce two new datasets for this task. The first contains 1,100 manually labeled tweets, and the second contains 32K user comments that the moderators of a popular Arabic news site deemed inappropriate. We are publicly releasing the datasets along with the lexicons we created.

2 Approach

In our approach, we created a set of obscene words to work as our seeding list. We extracted the list from a large set of tweets containing 175 million tweets that we obtained from Twitter during March 2014 using the Twitter streaming API with language filter set to Arabic "lang:ar". We searched the tweets for some patterns that are usually used in offensive communications, such as: يا ابن ال..، يا ولاد ال .. (You, son(s) of, daughter(s) of, .. etc.) along with their variant spellings. The words appearing after these patterns were then collected and manually assessed for being obscene or not. The final list after manual assessment contained obscene 288 words and phrases. Additionally, we added the 127 hashtags that are used to screen pornographic pages in an online tweet aggregator TweetMogaz (Elsawy et al., 2014; Magdy, 2013)). The list can be downloaded from: `http://alt.qcri.org/~hmubarak/offensive/ObsceneWords.txt`

Next, given our tweet set of 175 million tweets, we obtained a list of Twitter users, aka tweeps, who authored at least 100 tweets along with their tweets. The text of the tweets was cleaned and normalized in the manner described in (Darwish et al., 2012). This included the normalization of different shapes of hamza, yaa, and taa marbuta, normalization of decorative characters, and proper segmentation of hashtags and URLs. Given the list of tweeps, we divided them into two groups, namely: those who authored tweets that did not include a single obscene word from our aforementioned list (clean group) and those who used at least one of the words from our list at least once (obscene group). Our hypothesis is that those who use at least one of the words in our list are likely to use other obscenities that may not be included in our list. The size of the clean and obscene groups were 166K tweeps, who authored 86M tweets, and 23K tweeps, who authored 16M tweets, respectively.

Given the tweets of the two groups, we computed unigram and bigram counts in both of them. Given these counts, we computed the Log Odds Ratio (LOR) (Forman, 2008) for each word unigram and bigram that appeared at least 10 times. The tweets authored by the clean tweeps are used as a background corpus, and the tweets authored by the obscene tweeps are used as a foreground corpus. The computation of the LOR is as follows:

$$LOR = \log\left[\frac{tp\cdot(pos-tp)}{fp\cdot(neg-fp)}\right]$$

where tp and fp are the counts in the foreground and background corpora respectively, and pos and neg are the tweet counts in the foreground and background corpora respectively. We retained unigrams and bigrams that yielded an LOR equals to infinity which means that they appeared in the foreground corpus only (obscene) but didnt appear in the background corpus (clean), and we added them to our original list of words and phrases. This enhanced the precision, and in future we will consider other ranges of LOR to enhance the recall without affecting the precision. This process can be done iteratively. We performed one iteration and we generated 3,430 word unigrams and bigrams. We refer to list of words generated using this method as the LOR list.

3 Experimental Setup

To measure the effectiveness of our approach, we used intrinsic as well as extrinsic evaluation. For intrinsic evaluation, we randomly selected 100 words (unigrams or bigrams) from the list of generated words with LOR equals to infinity. We marked the words as either obscene or not. Of the 100 words, 59 were found to be obscene.

For extrinsic evaluation, we built a test set for the obscene and offensive language detection that contains 100 highly discussed tweets that each had at least 10 replies. Specifically, we collected the 100 tweets by identifying 10 controversial tweeps from the top tweeps in Egypt, according to SocialBakers.com. For each of the tweeps, we randomly selected 10 tweets that have 10 or more comments/replies. In all, we had 100 original tweets plus 1,000 comment/reply tweets – 1,100 tweets all together. For the tweets, we submitted each tweet along with its context (thread of replies) to CrowdFlower.com to be judged by 3 different annotators from Egypt. The annotators could mark the tweets as: obscene, offensive (but not obscene), or clean. Figure 1 shows three tweets and the their output judgments. Of the judged tweets, the percentages of obscene, offensive (but not obscene), and clean tweets were 19.1%, 40.3%, and 40.6% respectively. The average inter-annotator agreement was 85%. In the context of this paper, we are only considering obscene tweets in our evaluation. Offensive tweets are left for future work. The 1,100 annotated tweets can be downloaded from http://alt.qcri.org/~hmubarak/offensive/TweetClassification-Summary.xlsx. Given the annotated test set and our list of obscene words, we automatically tagged each tweet in the test set as obscene if it contained a word in the list. We experimented with several lists namely: the SeedWords list, the LOR list (word unigrams only), the LOR list (word bigrams only), combined LOR (unigrams only) + SeedWords lists, and combined LOR (bigrams only) + SeedWords lists. Table 1 shows the results (Precision, Recall, and F1) using the different lists. As can be seen in the results, using word unigrams is superior to using word bigrams. The results suggests that the initial seed word list yields high precision with relatively low recall. Combining SeedWords and LOR (unigram) lists yielded slightly improved recall, while maintaining the precision.

Using list-based methods to detect abusive language is proved to be good and robust (Sood et al., 2012b; Chen et al., 2012a). However, this approach is limited by its reliance on lists. This is shown also in our results in the form of high precision and low recall. Chen et al. (Chen et al., 2012a) suggest using lexical and syntactical features along with automatically generated black lists. We plan to explore such features to account for the complexities and richness of Arabic and its dialects. We also plan to look at morphological features to account for the rich morphology of Arabic. Breaking Arabic words into constituent clitics can be useful in generating appropriate morphological features.

List	P	R	F1
SeedWords (SW)	0.97	0.43	0.59
SW + LOR (unigrams)	0.97	0.44	**0.60**
SW + LOR (bigrams)	0.89	**0.45**	0.60
LOR (unigrams)	**0.98**	0.41	0.58
LOR (bigrams)	0.89	0.44	0.59

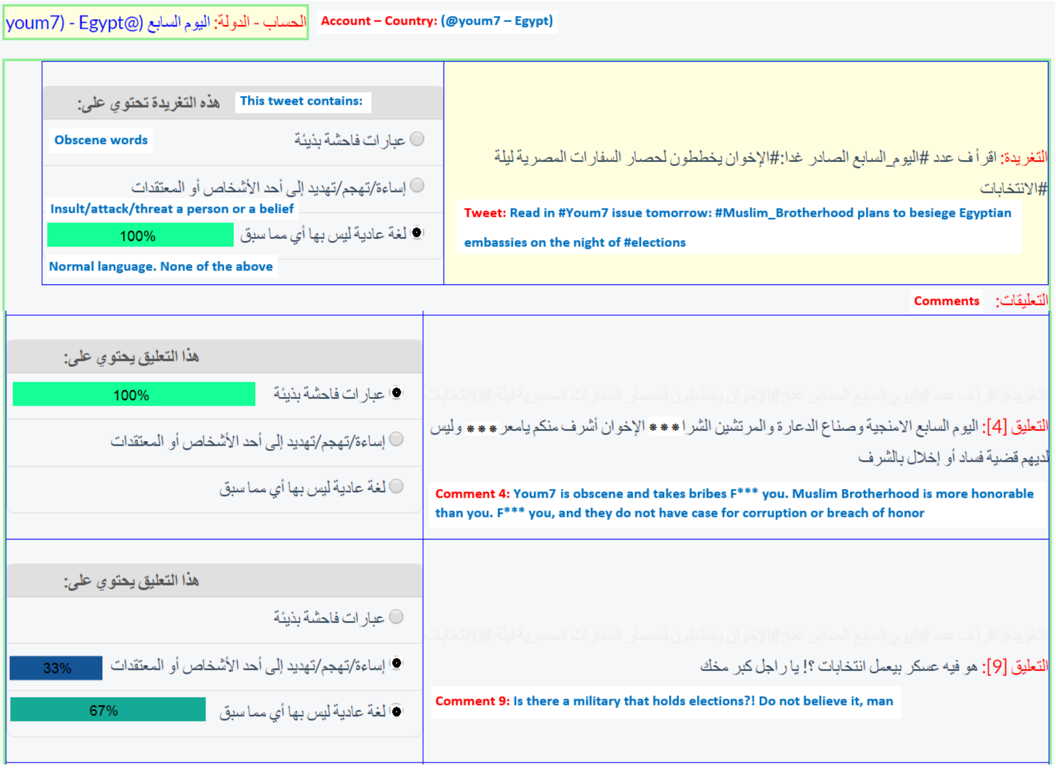
Table 1: Extrinsic evaluation results

4 Aljazeera Deleted Comments

In the interest of the research community, we are also releasing a dataset of 32K deleted comments from Aljazeera.net[1]. Aljazeera.net, a popular Arabic news channel, moderates all the comments that appear on their site. According to the site's "Community Rules and Guidelines" (http://www.aljazeera.com/aboutus/2011/01/201111681520872288.html), a user comment is not accepted if it is a personal attack, racist, sexist, or otherwise offensive, inciting violence, non-relevant, advertising, etc.

Initially we obtained a corpus of 400K comments on approximately 10K articles that cover many gneres such as politics, economy, society, and science. From these comments, we selected 32K comments whose lengths are between 3 and 200 characters to ease subsequent annotation. We annotated the selected comments using CrowdFlower, where three annotators were asked to classify comments as obscene, offensive, or clean. The annotators were also given article titles as we did not have the entire thread of comments. The breakdown of the annotation is as follows: 2% obscene, 79% offensive, and 19% clean. The inter-annotator agreement was 87%. Low percentage of obscene comments may be attributed to the fact that

[1] We would like to thank Aljazeera for courteously agreeing to release the data



الحساب - الدولة: اليوم السابع (@youm7) - Egypt | Account – Country: (@youm7 – Egypt)

التغريدة: اقرأ ف عدد #اليوم_السابع الصادر غدا:#الإخوان يخططون لحصار السفارات المصرية ليلة #الانتخابات

Tweet: Read in #Youm7 issue tomorrow: #Muslim_Brotherhood plans to besiege Egyptian embassies on the night of #elections

هذه التغريدة تحتوي على: | This tweet contains:

- ○ عبارات فاحشة بذيئة — Obscene words
- ○ إساءة/تهجم/تهديد إلى أحد الأشخاص أو المعتقدات — Insult/attack/threat a person or a belief
- ● لغة عادية ليس بها أي مما سبق — 100% — Normal language. None of the above

التعليقات: Comments

التعليق [4]: اليوم السابع الامنجية وصناع الدعارة والمرتشين الشرا *** الإخوان أشرف منكم يامعر *** وليس لديهم قضية فساد أو إخلال بالشرف

Comment 4: Youm7 is obscene and takes bribes F*** you. Muslim Brotherhood is more honorable than you. F*** you, and they do not have case for corruption or breach of honor

هذا التعليق يحتوي على:

- ● عبارات فاحشة بذيئة — 100%
- ○ إساءة/تهجم/تهديد إلى أحد الأشخاص أو المعتقدات
- ○ لغة عادية ليس بها أي مما سبق

التعليق [9]: هو فيه عسكر بيعمل انتخابات ؟! يا راجل كبر مخك

Comment 9: Is there a military that holds elections?! Do not believe it, man

هذا التعليق يحتوي على:

- ○ عبارات فاحشة بذيئة
- ● إساءة/تهجم/تهديد إلى أحد الأشخاص أو المعتقدات — 33%
- ● لغة عادية ليس بها أي مما سبق — 67%

Figure 2: CrowdFlower judgment screen (translations are added for clarification)

users know in advance that their comments on news agencies are subject to moderation, which is not the case when they post freely on social media.

The comments are written in Modern Standard Arabic (MSA) and different dialects. Examples of different types of offensive comments are shown in Table 2. We plan to use this corpus to detect offensive language for attacking people and hate speech. The data can be downloaded from: http://alt.qcri.org/~hmubarak/offensive/AJCommentsClassification-CF.xlsx

Comment	Type
كذاب ابن ك** ابن ق*** Liar, son of the ***	Obscene
الارهابى انت و اجدادک You and grandparents are terrorists	Attack
كلب وجب قتله A dog who must be killed	Violence
لأن العرب عبيد وسيبقون كذلك Arabs are slaves and will remain	Racism
عيب تحكمنا واحدة ست Shameful to be ruled by a woman	Sexism

Table 2: Examples of offensive user comments

5 Conclusion and Future Work

In this paper we present an automated method to create and expand a list of obscene words. We also introduce a new test set for the task, which we plan to make publicly available in addition to the list of obscene words and a large corpus of annotated user comments for obscene and offensive language detection.

For future work, we plan to enhance the recall by applying different algorithms, and to expand the test set to include tweets from multiple regions (Egypt, Gulf, Levant, Maghreb, and Iraq) to cover different dialects and cultures. Further, the work in this paper focused on identifying obscene tweets, and we plan to expand it to cover offensive language and hate speech. Additionally, we plan to study different levels of morphological and syntactic analysis, and using character n-grams as suggested by (Waseem, 2016) in addition to unigrams and bigrams to deal with the rich morphology of Arabic and its dialects. Hopefully, morphological processing can lead to improved recall.

References

Ehab A Abozinadah, Jr Jones, and H James. 2016. Improved microblog classification for detecting abusive arabic twitter accounts. In *International Journal of Data Mining and Knowledge Management Process*. IJDKP.

Ying Chen, Yilu Zhou, Sencun Zhu, and Xu Heng. 2012a. Detecting offensive language in social media to protect adolescent online safety. In *In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*. IEEE, pages 71–80.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, pages 2427–2430.

Eslam Elsawy, Moamen Mokhtar, and Walid Magdy. 2014. Tweetmogaz v2: Identifying news stories in social media. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM.

George Forman. 2008. Bns feature scaling: an improved representation over tfidf for svm text classification. In *In Proceedings of the 17th ACM conference on Information and knowledge management*. ACM.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. In *Journal of Politeness Research. Language, Behaviour, Culture 4, no. 2*.

Walid Magdy. 2013. Tweetmogaz: a news portal of tweets. In *In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM.

Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. 2008. Detecting flames and insults in text. In *In Proceedings of the Sixth International Conference on Natural Language Processing*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Michael Beißwenger, Michael Wojatzki, and Torsten Zesch, editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*. Bochum, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9.

Sara Sood, Judd Antin, and Elizabeth Churchill. 2012b. Using crowdsourcing to improve profanity detection. In *In AAAI Spring Symposium: Wisdom of the Crowd, volume SS-12-06 of AAAI Technical Report*. AAAI.

Ellen Spertus. 2007. Smokey: Automatic recognition of hostile messages. In *In Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, pages 138–142. http://aclweb.org/anthology/W16-5618.

Hate Speech Watch. 2014. Hate crimes: Consequences of hate speech. In *http://www.nohatespeechmovement. org/hate-speech-watch/focus/ consequences-of-hate-speech*.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *In Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM.