Abusive Language Detection on Arabic Social Media

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the First Workshop on Abusive Language Online

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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”).

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...
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 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 أبن آل .. يا ولد آل (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.

<table>
<thead>
<tr>
<th>List</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeedWords (SW)</td>
<td>0.97</td>
<td>0.43</td>
<td>0.59</td>
</tr>
<tr>
<td>SW + LOR (unigrams)</td>
<td>0.97</td>
<td>0.44</td>
<td><strong>0.60</strong></td>
</tr>
<tr>
<td>SW + LOR (bigrams)</td>
<td>0.89</td>
<td>0.45</td>
<td>0.60</td>
</tr>
<tr>
<td>LOR (unigrams)</td>
<td><strong>0.98</strong></td>
<td>0.41</td>
<td>0.58</td>
</tr>
<tr>
<td>LOR (bigrams)</td>
<td>0.89</td>
<td>0.44</td>
<td>0.59</td>
</tr>
</tbody>
</table>

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 genres 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.
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

<table>
<thead>
<tr>
<th>Comment</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>كاذب ابن لديّ ابن ق***</td>
<td>Obscene</td>
</tr>
<tr>
<td>Liar, son of the ***</td>
<td></td>
</tr>
<tr>
<td>الإرهابيّ انت وأجدادك</td>
<td>Attack</td>
</tr>
<tr>
<td>You and grandparents are terrorists</td>
<td></td>
</tr>
<tr>
<td>كلب وجب قتله</td>
<td>Violence</td>
</tr>
<tr>
<td>A dog who must be killed</td>
<td></td>
</tr>
<tr>
<td>لأن العرب عبد وسيفون كذلك</td>
<td>Racism</td>
</tr>
<tr>
<td>Arabs are slaves and will remain</td>
<td></td>
</tr>
<tr>
<td>عيب كهكنا واحدة ست</td>
<td>Sexism</td>
</tr>
<tr>
<td>Shameful to be ruled by a woman</td>
<td></td>
</tr>
</tbody>
</table>

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


