Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, pages 40–47
with a Shared Task on Offensive Language Detection.
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Understanding and Detecting Dangerous Speech in Social Media

Ali Alshehri[1†], El Moatez Billah Nagoudi[2†] , Muhammad Abdul-Mageed[2]
[1] SUNY at Buffalo
[2] Natural Langauge Processing Lab, The University of British Columbia
alimoham@buffalo.edu, {moatez.nagoudi,muhammad.mageeed}@ubc.ca

Abstract

Social media communication has become a significant part of daily activity in modern societies. For this reason, ensuring safety in social media platforms is a necessity. Use of dangerous language such as physical threats in online environments is a somewhat rare, yet remains highly important. Although several works have been performed on the related issue of detecting offensive and hateful language, dangerous speech has not previously been treated in any significant way. Motivated by these observations, we report our efforts to build a labeled dataset for dangerous speech. We also exploit our dataset to develop highly effective models to detect dangerous content. Our best model performs at 59.60% macro F_1, significantly outperforming a competitive baseline.

1. Introduction

The proliferation of social media makes it necessary to ensure online safety. Unfortunately, offensive, hateful, aggressive, etc., language continues to be used online and put the well-being of millions of people at stake. In some cases, it has been reported that online incidents have caused not only mental and psychological trouble to some users but have indeed forced some to deactivate their accounts or, in extreme cases, even commit suicides (Hinduja and Patchin, 2010). Previous work has focused on detecting various types of negative online behavior, but not necessarily dangerous speech. In this work, our goal is to bridge this gap by investigating dangerous content. More specifically, we focus on direct threats in Arabic Twitter. A threat can be defined as "a statement of an intention to inflict pain, injury, damage, or other hostile action on someone in retribution for something done or not done."[1] This definition highlights two main aspects: (1) the speaker's intention of committing an act, which (2) he/she believes to be unfavorable to the addressee (Fraser, 1998). We especially direct our primary attention to threats of physical harm. We build a new dataset for training machine learning classifiers to detect dangerous speech. Clearly, resulting models can be beneficial in protecting online users and communities alike.

The fact that social media users can create fake accounts on online platforms makes it possible for such users to employ hostile and dangerous language without worrying about facing effective social nor legal consequences. This continues to put the responsibility on platforms such as Facebook and Twitter to maintain safe environments for their users. These networks have related guidelines and invest in fighting negative and dangerous content. Twitter, for example, prohibits any form of violence including threats of physical harm and promotion of terrorism. [2] However, due to the vast volume of communication on these platforms, it is not easy to detect harmful content manually. Our work aims at developing automated models to help alleviate this problem in the context of dangerous speech.

Our focus on Arabic is motivated by the wide use of social media in the Arab world (Lenze, 2017). Relatively recent estimates indicate that there are over $11M$ monthly active users as of March 2017, posting over $27M$ tweets each day (Salem, 2017). An Arabic country such as Saudi Arabia has the highest Twitter penetration level worldwide, with 37% (Iqbal, 2019). The Arabic language also presents interesting challenges primarily due to the dialectical variations cutting across all its linguistic levels: phonetic, phonological, morphological, semantic and syntactic (Farghaly and Shaalan, 2009). Our work caters for dialectal variations in that we collect data using multi-dialectal seeds (Section 3.3.). Overall, we make the following contributions:

1) We manually curate a multi-dialectal dictionary of *physical harm threats* that can be used to collect data for training dangerous language models.

2) We use our lexicon to collect a large dataset of threatening speech from Arabic Twitter, and manually annotate a subset of the data for dangerous speech. *Our datasets are freely available online.*[5]

3) We investigate and characterize threatening speech in Arabic Twitter.

4) We train effective models for detecting dangerous speech in Arabic.

The remainder of the paper is organized as follows: In Section 2., we review related literature. Building dangerous lexica used to collect our datasets is discussed in Section 3.3.. We describe our annotation in Section 4.1.. We present our models in Section 5., and conclude in Section 6..

2. Related work

Detection of offensive language in natural languages has recently attracted the interest of multiple researchers. However, the space of abusive language is vast and has its own nuances. Waseem et al. (2017) classify abusive

† Both authors contributed equally.

[1] https://en.oxforddictionaries.com/definition/threat

[2] https://help.twitter.com/en/rules-and-policies/twitter-rules

language along two dimensions: *directness* (the level to which it is directed to a specific person or organization or not) and *explicitness* (the degree to which it is explicit). Jay and Janschewitz (2008) categorize offensive language to three categories: *Vulgar*, *Pornographic*, and *Hateful*. The Hateful category includes offensive language such as threats as well as language pertaining to class, race, or religion, among others. In the literature, these concepts are sometimes confused or even ignored altogether. In the following, we explore some of the relevant work on each of these themes.

Offensive Language. The terms *offensive language* and *abusive language* are commonly used interchangeably. They are cover terms that usually include all types of undesirable language such as *hateful*, *racist*, *obscene*, and *dangerous* speech. We review some work looking at these types of language here, with no specific focus on any of its forms. GermEval 2018 is a shared task on the Identification of Offensive Language in German proposed by Wiegand et al. (2018). Their dataset consists of $8,500$ annotated tweets with two labels, "offensive" and "non-offensive". Another relevant shared task is the OffensEval (Zampieri et al., 2019), which focuses on identifying and categorizing offensive language in social media. Very recently, an Arabic offensive language shared task is included in the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4). [3]

Hate Speech. Hate speech is a type of language that is biased, hostile, and malicious targeting a person or a group of people because of some of their actual or perceived innate characteristics (Gitari et al., 2015). This type of harmful language received the most attention in the literature. Burnap and Williams (2014) investigate the manifestation and diffusion of hate speech and antagonistic content in Twitter in relation to situations that could be classified as 'trigger' events for hate crimes. Their dataset consists of $450K$ tweets collected during a two weeks window in the immediate aftermath of Drummer Lee Rigby's murder in Woolwich, UK. In Waseem (2016), issues of annotation reliability are discussed. Authors examine whether the expertise level of annotators (e.g expert or amateur) and/or the type of information provided to the annotators, can improve the classification of hate speech. For this purpose, they extend the dataset of (Waseem and Hovy, 2016) with a set of about $7K$ tweets annotated by two types of CrowdFlower users: expert and amateur. They find that hate speech detection models trained on expert annotations outperform those trained on amateur annotations. *This suggests that hate speech can be implicit and thus harder to detect by humans and machines alike*. Another work by (Davidson et al., 2017) builds a hate speech lexicon based on a list of words and phrases provided by *Hatebase.org*. Using Twitter API, they crawled a set of $85M$ tweets containing terms from the lexicon. Most recent works on detecting hate on Twitter are done as part of a SemEval2019 competition, HatEval (Òscar Garibo, 2019). This shared task addresses the problem of multilingual detection of hate speech against immigrants and women in Twitter.

Obscene Language. Obscene speech includes vulgar and pornographic speech. A few research papers have looked at this kind of speech in social media (Singh et al., 2016; Mubarak et al., 2017; Alshehri et al., 2018). Mubarak et al. (2017) present an automated method to create and expand a list of obscene words, for the purpose of detecting obscene language. Abozinadah (2015) build a dataset of over $1M$ tweets comprising the most recent 50 tweets of 255 users who has participated in swearing hashtags as well as the most recent 50 tweets of users in their network. As feature input to their classifiers, the authors extracted basic statistical measures from each tweet and reported 96% accuracy of adult content detection. Alshehri et al. (2018) build a dataset of adult content in Arabic twitter and their distributors. The work identifies geographical distribution of targets of adult content and develops models for detecting spreaders of such content.Alshehri et al. (2018) report 79% accuracy on detecting adult content.

Racism and Sexism. Kwok and Wang (2013) create a balanced dataset comprising $24,582$ of 'racist' and 'non-racist' tweets. Waseem and Hovy (2016) collect a set of $136K$ hate tweets based on a list of common terms and slurs pertaining ethnic minorities, gender, sexuality, and religion. Afterwards, a random set of $16K$ tweets are selected and manually annotated with three labels: 'racist', 'sexist', or "neither". Gambäck and Sikdar (2017) introduce a deep-learning-based Twitter hate speech text classification model. Using data from Waseem and Hovy (2016) with about $6.5K$ tweets, the model classifies tweets into four categories: 'sexist', 'racist', 'both sexist and racist', and 'neither'. Clarke and Grieve (2017), using the same list, explore differences among racist and sexist tweets along three dimensions: *interactiveness, antagonism*, and *attitude* and find an overall significant difference between them.

Dangerous Language. Little work has been dedicated to detection and classification of dangerous language and threats. They are usually part of work on abusive and hate speech. This is to say that dangerous language has only been indirectly investigated within the NLP community. However, there is some research that is not necessarily computational in nature. For example, Gales (2011) investigates the correlation between interpersonal stance and the realization of threats by analyzing a corpus of 470 authentic threats. Ultimately, the goal of Gale's work is to help predict violence before it occurs. Hardaker and McGlashan (2016), on the other hand, investigates the language surrounding threats of rape on Twitter. In their corpus, the authors find that women were the prime target of rape threats. In the rest of this paper, we explore the space and language of threats in Arabic Twitter. We now describe our lexicon and datasets.

[3] http://edinburghnlp.inf.ed.ac.uk/workshops/OSACT4/

Verb	Dialect	English	Verb	Dialect	English	Verb	Dialect	English
أباد	G,M,R	exterminate	رض	G,M	contuse	فجر	all	blow up
أتل	E,L	kill	سطر *	E,G	mark	فشق	G,L	split
أدى *	E,G	give	سلخ	all	skin	فقع*	E,G,L,R	burst
أعدم	all	execute	سلق	E,G,R	boil	فك*	E,G,L	disentangle
أفنى	G,M,R	exterminate	شج	M	slash	قتل	all	kill
أهلك	G,M,R	destroy	شرب **	E,G,L,R	drink	قرح	E	sound
إغتال	G,L,M,R	assassinate	شق	E,G,L,R	rip off	قسم	all	divide
إغتصب	all	rape	شوه	E,G,L,R	distort	قصف	G,R	smash
إقتلع *	G,L,R	pluck	صرم	G	cut off	قصم	G,M	smash
بطش	E,L,M	assault	صفق	G,L	slap	قضى	E,G,L,M	eliminate
جرح	all	wound	صلخ	G,L	skin	قطع	all	cut
جزر	G	cut off	ضرب	all	hit	قلع	E,G,L,R	pluck
جلد	all	whip	طخ	E,G	shoot	كسر	all	break
حرق	all	burn	طعن	all	stab	لخ	G	hit
حطم	E,L,M,R	smash	طير	E,G,L,R	make fly	محا **	E,G,L,R	erase
دك	E,G,L	demolish	عذب	E,G,M,R	torture	محق	M	destroy
دهك	G	run over	عزب	E	torture	نحر	E,G,M,R	slaughter
ذبح	all	slaughter	عقر	E,G	kill	نسف	E,G,M,R	blast
رجم	E,G,M,R	stone	فتك	E,G,L,M	destroy	هشم	G,L,R	smash

Table 1: Our list of dangerous verbs. All= all dialects. E= Egyptian. G= Gulf. L= Levantine. M= MSA. R= Maghrebi. * = metaphorical. ** = used idiomatically.

3. Dangerous Lexica and Dataset

3.1. Dangerous Language

We define dangerous language as *a statement of an intention to inflict physical pain, injury, or damage on someone in retribution for something done or not*. This definition excludes threats that do not reflect physical harm on the side of the receiver end of the threat. The definition also excludes *tongue in cheek* whose real intention is to tease. An example of this later category is a threat made in the context of sports where it is common among fans to tease one another using metaphorical, string language (see Example # 6 in Section 3.3.).

3.2. Dangerous Lexica

We came up with a list of 57 verbs in their basic form that can be used literally or metaphorically to indicate physical harm (see table 1). This list is by no means exhaustive, although we did our best to expand it as much as possible. As such, the list covers the frequent verbs used in the threatening domain in Arabic. [4] These verbs are used in one or more of the following varieties: Egyptian, Gulf, Levantine, Maghrebi, and MSA (see table 2 for more details). Most of these verbs (n=50 out of 57) literally indicate physical harm. Examples are طعن (*'to stap'*) and سلخ (*'to de-skin'*). The rest are used (sometimes metaphorically) to indicate threatening, such as اقتلع (*'to pluck'*) and سطر (*'to mark'*) usually with a body part such as وجه (*'face'*) or رأس (*'head'*). Finally, some of the verbs are used idiomatically, such as شرب من دم (*'to drink someone's blood'*) and محا من على وش الارض (*'to erase/eliminate from the face of the earth'*). Multiword expressions in our seed list can be found in Table 3.

Dialect	# of verbs
MSA	30
Gulf	50
Egyptian	39
Maghrebi	34
Levantine	34
All (unique)	57

Table 2: Distribution of threat verbs across Arabic dialects.

To be able to collect data, we used our manually curated list to construct threat phrases indicating physical harm such as اقتلك (*'I kill you'*) and يكسره (*'He breaks him/it'*). That is, each phrase consists of a physical harm verb, a singular or plural first or third person subject, and a plural or singular second or third person object. This gives us the following pattern:

1st/3rd (SG / PL) + threat verb + 2nd/3rd (SG / PL)

[4] The concept of frequency here is based on native speaker knowledge of the language. The list was developed by the 3 authors, all of whom are native speakers of Arabic with multi-dialectal fluency.

Some of the phrases only differ on the basis of spelling due to dialectical variations. For example, the body part وجه (*'face'*) can be spelled as وجيهكم or وجوهكم in the plural form depending on the dialect. Another example is the verb قتل (*'kill'*), which can also be spelled as اتل in Egyptian and some other Arabic dialects. Manual search of some of the seed tokens in twitter suggests that patterns involving 3rd person subject are almost always not threats. The following are two illustrating examples of this non-threatening use:

1) ميسي إذا لم يسجل فإنه يقتل فرحة بعض البشر
'If he doesn't score, Messi kills happiness in some people'

2) ما يكسر الخاطر ... سوى شخص غالي
'Only a dear friend can break one's heart'

Thus, we decided to limit our list of phrases to 'direct' dangerous threats, which are phrases involving a singular or plural first person subject and singular or plural second person object as follows:

1st (SG/PL) + threat verb + 2nd (SG/PL)

Examples of these direct threats include نغتصبك (*'We rape you'*) and احرقكم (*'I burn you'*). Less dangerous threats such as اجرحكم (*"We hurt you (all)"*) and ادفك (*'I push you'*) are also not considered. Our motivation for not including these latter phrases even though they involve direct threats is that they indicate less danger and (more crucially) are more likely to be used metaphorically in Arabic. This resulted in a set of 286 direct and dangerous phrases, which constitute our list of 'dangerous' seeds. We make the list of 286 direct threats phrases available to the research community. [5]

3.3. Dataset

We use the constructed 'dangerous' seed list to search Twitter using the REST API for two weeks resulting in a dataset of $2.8M$ tweets involving 'direct' threats as shown in Table 4. We then extract *user ids* from all users who contributed the REST API data ($n = 399K$ users) and crawled their timelines ($n = 705M$ tweets). We then acquire $107.5M$ tweets from the timelines, each of which carry one or more items from our 'dangerous' seed list. Combining these two datasets (the REST API dataset and dataset based on the timelines) results in a dataset consisting of $110.3M$ tweets as shown in Table 4. In this work, we focus on exploiting the REST API dataset exclusively, leaving the rest of the data to future research.

4. Data Annotation

4.1. Annotation

We first randomly sample $1K$ tweets from our REST API dataset.[5] Two of the authors annotated each tweet for being a threat ('dangerous') or not ('safe'). This sample annotation resulted in a Kappa (κ) score of 0.57, which is fair according to Landis and Koch's scale (Landis and Koch, 1977). The two annotators then held several discussion sessions to improve their mutual understanding of the problem and define some instructions as to how to label the data. We also added another random sample of $4K$ tweets (for a total size of $5K$) to the annotation pool. After extensive revisions of the disagreement cases by the two annotators, the κ score for the whole dataset ($5K$) was found to be at 0.90. The annotated dataset has a total of $1,375$ tweets in the 'dangerous' class and $3,636$ in the 'non-dangerous' class. Our overall agreed-upon instructions for annotations include the following:

- Textual threats combined with pleasant emojis such as 🙂 and 💙 are not dangerous, as opposed to threat combined with less pleasant emojis such as 🔪 🔫. Thus, tweet 3 below should be coded as 'safe' while tweet 4 should be tagged as 'dangerous'.

  3) @user المنطق يقول أنا باقتلك 🙂
  'It goes with logic that I kill you 🙂'

  4) @user @user @user قدامي بس لا اطعنك 🔪
  'Move forward [in front of me] or else I stab you 🔪'

- Mitigated threats with question marks or epistemic modals are dangerous unless they are combined with positive language or emojis such as Example 5 below. Note that the word *Touha* in Example 5 is an informal, friendly form for Arabic names such as *FatHi* or *MamdouH*.

  5) @user انا بفكر اقتلك يا توحه 🙂
  'I am thinking of killing you, Touha 🙂'

- Threats related to sports are not dangerous. That is because it is common to use verbs like نحر (*"slaughter"*) and اغتصب (*"rape"*) among fans of rival teams to describe wins and losses, as in the following example.

  6) @user حلاته نغتصبكم على ارضكم وبين جمهوركم
  'It's actually better that we 'rape' you in your stadium, among your fans'

- Ambiguous threats such as threats consisting of one word (as in Example 7 below) should be coded as 'dangerous':

  7) اقتلكوا
  'I kill you'

Below, we show examples of tweets that were annotated as 'dangerous':

8) @user @user
وددي احرقك و ارميك للكلاب
'I wish to burn you and throw you to dogs'

[5] https://github.com/UBC-NLP/ara_dangspeech.

Seed	English	Seed	English
اشرب من دمك	I drink from your blood	امحيك من على وش الارض	I erase you from the face of the earth
اشرب دمك	I drink your blood	امحيكم من على وش الارض	I erase you all from the face of the earth
اشرب من دمكم	I drink your blood all	نشرب من دمك	We drink from your blood
اشوه وجهك	I disfigure your face	نشرب دمك	We drink your blood
اطير راسك	I cut your head	نشرب من دمكم	We drink your blood all
اطير روسكم	I cut your head all	نطير راسك	I cut your head
افجر راسك	I blow up your head	نطير روسكم	I cut your head all
افقع وجهك	I burst your face	نفجر راسك	We blow up your head
افقع وجوهكم	I hit your face all	نفقع وجهك	We hit your face
افك وجهك	I disentangle your face	نقضي عليك	We finish you
اقضي عليك	I finish you	نقضي عليكم	We finish you all
اقضي عليكم	I finish you all	نمحيك من على وش الارض	I erase you from the face of the earth
اكسر وجهك	I break your face	نمحيكم من على وش الارض	We erase you all from the face of the earth

Table 3: Multiword expressions in our seed list.

Dataset	# of tweets
REST API	$2.8M$
Timelines	$107.5M$
ALL	$110.3M$

Table 4: Breakdown of our 'dangerous' dataset.

	Safe	Dangerous	Total
Safe	3,570	52	3,622
Dangerous	70	1319	1,389
Total	3,640	1,371	5,011

Table 5: Annotator Agreement of 5011-tweet sample.

9) @user ماتكلمنيش ب الطريقة دي لحسن اقوم اضربك
قال اربع نسوة قال 😒
'Don't talk to me in this way, or else I hit you! Talking of (marrying) four women!'

10) @user @user سوف تبدا الحرب ورب العرش العظيم سوف
نحرقكم حرق انتم ياخنثيين يا خنزير العرب يا خاونه
'The war will begin. By God, we will burn you down, you fags, you pigs, you traitors'

11) @user الحمار دايم حمار مااستفدتوا من الدرس هاذا لازم
نضربكم على قفاكم زي الجهال انتوا بشر ولاحيونات
'A donkey will always be a donkey. You didn't learn the lesson. We have to hit you on the back of you heads like kids. Are you humans or animals?'

12) @user عطيني كروكي بيتكم واجي اشرحك مو بس اقتلك
'Give me your address so I can come to you, and not only kill you but also dissect you'

Measure	Value
Avg. # timeline tweets	2,313
Avg. # dangerous tweets / user	3.97
St. dev.	3.64
25th percentile	1
50th percentile	4
75th percentile	6
Minimum	1
Maximum	23

Table 6: Descriptive statistics of the timeline data of 1,370 users who contributed tweets classified as 'dangerous' in our annotated dataset.

Seed	English	Emoji
اذبحك	I slaughter you	😅
اقتلك	I kill you	💔
اغتصبك	I rape you	😈
اضربك	I hit you	😡
اعذبك	I torture you	😭
اديك	I hit/give you	🖤
اجلدك	I lash you	😒
اطعنك	I stab you	✋
اجرحك	I hurt you	🔪
احرقك	I burn you	🔫

Table 7: Top 10 most frequent 'dangerous' seeds and emojis in our REST API dataset.

4.2. Data Analysis

The fact that 'dangerous' tweets are not frequent in the dataset suggests that *this phenomenon of dangerous speech is relatively rare in the Twitter domain*. To further investigate the commonality of such a phenomenon, we extract

Models	Datasets	Precision	Recall	Acc	F_1
Baseline	–	50.00	29.33	58.66	36.97
BERT	Dangerous	58.42	**60.10**	74.27	58.98
BERT	Dangerous + Offensive	53.80	53.44	66.11	53.52
BERT-Emotion	Dangerous	**60.06**	59.24	**77.97**	**59.60**
BERT-Emotion	Dangerous + Offensive	54.50	53.99	66.84	54.11

Table 8: Results from our models on TEST.

the timelines of the authors of tweets in the dangerous class in the annotated dataset. Table 6 shows some descriptive statistics of the occurrence of dangerous seeds in their timelines. We can see from Table 6 that timelines contain on average $2,313$ tweets for each user, and there are on average 3.97 tweets in each timeline containing a dangerous seed token. This represents $\sim 0.17\%$ of the tweets for each user. The average number of dangerous tweets is higher ($n = 6$) for users in the $75th$ percentile as opposed to $n = 1$ in the $25th$ percentile.

To further understand dangerous language, we also analyze all the $5,011$ tweets from our annotated dataset. We identify a number of patterns in the data, cutting across both the 'dangerous' and 'safe' classes. We explain each of these patterns next.

Conditional threats: One common threatening pattern involves conditional statements where the consequent involves a physical threat by the speaker toward the addressee, and the antecedent is a conditional phrase involving deterrence of an action that can possibly be carried out by the addressee or someone else. The following are two examples:

13) @user اذبحك اذا تسوين شي
'I slaughter you if you (F) do anything'

14) @user اذا انتقل سوف اطعنك طعناً مبرحاآ امام الملآ
'If he transfers, I will stab you hardly in front of the crowds'

It is clear from Examples 13 and 14 that the threats are directed to a twitter user mentioned in the tweet. So these tweets are potentially part of ongoing conversations between the person who posted the tweet and the user mentioned in the body of the tweet. As Table 9 shows, $\sim$ 71.2% of tweets in our annotated dataset (across the 'dangerous' and 'safe' classes) contain mentions of other Twitter users. This percentage is higher within the dangerous class (%= 78).

Threats accompanied with commands: Another common pattern involves a command accompanying the threat as in Example 15 below. These kinds of threats are more common in the dangerous than the safe class.

15) @user اقول انقلعي لا افقع وجهك
'I say get out before I hit your face'

Threats accompanied with questions: Another less common pattern is threats in the form of questions. This kind of threats occurs in about 5% of our dangerous data as compared to 2.8% in the safe class. Unlike the examples above, the reason behind most of the 'question' threats is not particularly clear as they tend to be short, sometimes of one word. Interpretation of these threats requires more context, beyond the level of the tweet itself. Examples 16-18 illustrate this category.

Phenomena	Freq.	Percentage (non-dangerous)	Percentage (dangerous class)
Mentions	3673	72.8%	78%
Questions	100	2.8%	5%
Emoji	$2,010$	45.5%	36%
Conditional	742	15.8%	11.4%
Body parts	378	6.6%	11.3%
Hahaha	355	9.9%	1.1%

Table 9: The frequency of some textual phenomena in our Annotated data.

16) ممكن اقتلك من الم الحبر؟
'can i kill you by the pen'

17) @user ينفع اغتصبك؟
'Does it work if I rape you?'

18) @user اذبحك؟
'I slaughter you?'

Threat accompanied by modality: Some threats carry *deontic modality* where modals such as 'would', 'probably', 'may' are employed. *Epistemic modality* are also found in some data points. Similar to the question types above, these tweets (Examples 19-21 below) are less threatening than Examples 13-18 above.

19) @user جعلني اغتصبك
'May I rape you?!' (***deontic modality***)

20) @user ودي اذبحك
'I would like to kill you' (***deontic modality***)

21) @user شكلي رح اقتلك مع صحبتك
'I am probably going to kill you with your friends' (***epistemic modality***)

Metaphorical threats: Many of the tweets involve metaphorical use of the phrases in our annotated data. The target domain of the majority of these metaphorical uses is either sports or relationships. Words such as 'kill', 'rape',

and 'slaughter' are used to indicate 'wining' in sport or 'burn' to mean 'pain' or 'longing' in romantic relationships. Examples 23-24 illustrate these cases:

20) احب قول لاخواني المانشستراويه بكره راح
نغتصبكم فلا داعي للذعر والضجر واتمااستمتعوا
'I would like to tell my Manchester (football club) fans that we will rape them tomorrow'

21) س احرقك عشقا واطفئك غراما
'I will burn you with love and put off (the fire on you) with affection'

Emojis: Another interesting phenomenon (see Table 9) is the frequent use of emojis, which are found in about 40% of the annotated dataset. This is not surprising as it helps participants mitigate (and hence better disambiguate the nature of) their threats. Table 7 shows the top most frequent emojis used in our REST API data. It is evident that most of the used emojis do not indicate friendliness, but rather have a threatening nature. This is also true of using expressive interjections such as *hahaha*, which is more common in the non-dangerous than the dangerous class. Additionally, as mentioned above, some expressions involve use of 'body parts' such as *eyes, head, face, nose*, etc.. These are found to occur significantly higher in the 'dangerous' class.

Conversational context: Finally, Table 7 also shows the top 10 most frequent seeds in our REST API dataset. All of these seeds involve a first singular person subject and a singular second person object, which indicate that many of these tweets containing dangerous seeds are part of one-to-one conversations on Twitter.

5. Deep Learning Models

Dangerous speech data. We use our $5,011$ annotated tweet dataset for training deep learning models on dangerous speech. The dataset comprises $3,570$ 'safe' tweets and $1,389$ 'dangerous' tweets. We first remove all the seeds in our lexicon since these were used in collecting the data. We then keep only tweets with at least two words, obtaining $4,445$ tweets with $3,225$ 'safe' labels and $1,220$ 'dangerous' tweet (see Table 10). We split this dataset into 80% training, 10% development, and 10% test.

Offensive speech data. In one of our settings, we also use the offensive dataset released via the Offensive Shared Task 2020.[6] This offensive content dataset consists of 8000 tweets ($1,590$ 'offensive' and $6,410$ 'non-offensive'). We use the offensive class data to augment our train split. Hence, we evaluate only on our test split where tweets are restricted to our dangerous gold tweets in the annotated dataset. We run this experiment as a way to test the utility of exploiting offensive tweets for enhancing dangerous language representation based on the assumption that dangerous speech is a subset of offensive language. However, as we see in Table 8, this measure did not result in any improvements on top of our dangerous models. In fact, it leads to model deterioration.

	Train	Dev	Test
#Safe	$2,727$	244	254
#Dangerous	852	189	179
Total	$3,579$	433	433

Table 10: Distribution of dangerous and safe classes in our annotated dataset after normalization by removing seeds and one-word tweets.

Models. For the purpose of training deep learning models for detecting dangerous speech, we exploit the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) model. For all our models, we use the BERT-Base Multilingual Cased (Multi-Cased) model.[7] It is trained on Wikipedia for 104 languages (including Arabic) with 12 layers, 12 attention heads, 768 hidden units each and 110M parameters. Additionally, we further fine-tune an off-the-shelf trained BERT Emotion (BERT-EMO) from AraNet (Abdul-Mageed et al., 2019) on our dangerous speech task. BERT-EMO is trained with Google's BERT-Base Multilingual Cased model on 8 emotion classes exploiting Arabic Twitter data. We train all BERT models for 20 epochs with a batch size of 32, maximum sequence size of 50 tokens and learning rate up to $2e^{-5}$. We identify best results on the development set, and report final results on the blind test set. As our baseline, we use the majority class in our training split. Note that since our dataset is not balanced, the majority class baseline is competitive (63.97% macro F_1 score). Also, importantly, due to the imbalance in class distribution, the macro F_1 score (the harmonic mean of precision and recall) is our metric of choice as it is more balanced than accuracy.

Results & Discussion. As Table 8 shows, the results demonstrate that all the models outperform the baseline and succeed in detecting the dangerous speech with F_1 scores between 53.42% and 59.60%. We also observe that training on the offensive dataset did not improve the results. On the contrary, augmenting training data with the offensive task tweets cause deterioration to 53.52% F_1 for BERT and 54.11% F_1 for BERT-Emotion.

The best model for detecting dangerous tweets is BERT-Emotion when fine-tuned on our gold dangerous dataset. It obtains an accuracy level of 77.97% and F_1 score of 59.60%. We note that *both* accuracy and F_1 are *significantly higher then the the baseline*. As mentioned earlier, since our dataset is highly imbalanced, F_1, rather than accuracy, should be used as the metric of choice for evaluation. As such, our models are significantly better than our baseline.

[6] http://edinburghnlp.inf.ed.ac.uk/workshops/OSACT4/

[7] https://github.com/google-research/bert/blob/master/multilingual.md.

6. Conclusion

We have described our efforts to collect and manually label a dangerous speech dataset from a range of Arabic varieties. Our work shows that dangerous speech is rare online, thus making it difficult to find data for training machine learning classifiers. However, we were able to collect and annotate a sizeable dataset. To accelerate research, we will make our data available upon request. Another contribution we made is developing a number of models exploiting our data. Our best models are effective, and can be deployed for detecting the rare, yet highly serious, phenomenon of dangerous speech. For future work, we plan to further explore contexts of use of dangerous language in social media. We also plan to explore other deep learning methods on the task.

7. Bibliographical References

Abdul-Mageed, M., Zhang, C., Hashemi, A., and Nagoudi, E. M. B. (2019). Aranet: A deep learning toolkit for arabic social media. *arXiv preprint arXiv:1912.13072*.

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

Alshehri, A., Nagoudi, A., Hassan, A., and Abdul-Mageed, M. (2018). Think before your click: Data and models for adult content in arabic twitter. *The 2nd Text Analytics for Cybersecurity and Online Safety (TA-COS-2018), LREC*.

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.

Clarke, I. and Grieve, J. (2017). Dimensions of abusive language on twitter. Association for Computational Linguistics.

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*.

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

Farghaly, A. and Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–22.

Fraser, B. (1998). Threatening revisited. *Forensic linguistics*, 5(2).

Gales, T. (2011). Identifying interpersonal stance in threatening discourse: An appraisal analysis. *Discourse Studies*, 13(1):27–46.

Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Hardaker, C. and McGlashan, M. (2016). 'real men don't hate women': Twitter rape threats and group identity. *Journal of Pragmatics*, 91:80 – 93.

Hinduja, S. and Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221.

Iqbal, M. (2019). Twitter revenue and usage statistics in 2019. November.

Jay, T. and Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *AAAI*.

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

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.

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.

Òscar Garibo, i. O. (2019). Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 460–463.

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.

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.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *CoRR*, abs/1705.09899.

Waseem, Z. (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*, pages 138–142.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.