Towards a corpus of violence acts in Arabic social media

Ayman Alhelbawy,[♣,♠,◇] Udo Kruschwitz,[♠] Massimo Poesio [♠]
[♣] Minority Rights Group, London, UK [♠] University of Essex, Colchester, UK [◇] Fayoum University, Fayoum, Egypt
aalhel, udo, poesio@essex.ac.uk

Abstract

In this paper we present a new corpus of Arabic tweets that mention some form of violent event, developed to support the automatic identification of human rights abuses and different violent acts. The dataset was manually labelled for seven classes of violence using crowdsourcing. Only tweets classified with a high degree of agreement were included in the final dataset.

Keywords: Arabic NLP, Human Rights Abuse, Violence, Arabic Twitter Corpus, Classification

1. Introduction

Unfortunately, a great number of Human Rights Abuse (HRA) incidents take place every day in Arab countries in a conflict or quasi-conflict situation like Syria, Iraq, and Libya. Human Right Organizations do their best to find out about such incidents, but they have very few representatives and these cannot be everywhere. Social media, together with text messages and emails, provide an alternative channel to report such abuse. And Human Language Technology could provide a means to retrieve large quantities of such reports from a variety of media and assess their plausibility. In order to support such activity, we created a corpus of tweets reporting violent activities, including HRAs as well as other types of crime and conflict. To our knowledge, there is no resource to support such research either in English or Arabic. A further issue is that people posting on Twitter normally do not use Modern Standard Arabic (MSA)–only news organizations do so–but the 'dialectal' forms of Arabic used in everyday life. This means, for instance, that different words are used to express the same sense depending on where the posters are located.

In this paper, we will describe the corpus and the methodology we adopted to create it. In Section 2. we discuss how we collected an initial 557,576 tweets using a semi-automatically prepared list of violence words and a twitter word filter. Additional filters were deployed to filter posts that contained violence words but do not really report violent accidents. In Section 3. we discuss the classification scheme we developed to classify the tweets reporting violence. This scheme was used to manually annotate a dataset of 20,151 tweets using crowdsourcing. Finally, we investigate the degree of agreement achieved using the classification scheme, and report that we observe a substantial level of inter-rater agreement.

2. Data Collection

As we are only interested in Arabic tweets, we use the Twitter language filter to receive only Arabic tweets. But there are still millions of tweets in Arabic, a small fraction of which may mention some kind of violence. So, further filters are required to find an initial set of tweets that may mention some kind of violence to use for manual classification. We developed two filters for this purpose that we discuss in this Section.

Arabic Word	English Translation
قتل	Kill
مصابين	Injured
هجوم	Attack
يفجر	Blowing
يدمر	Destroys
أختطاف	Abduction, or kidnapping

Table 1: Examples of 'Violence Words'

2.1. Basic Twitter Filter

The first filter uses the Twitter API to receive Arabic tweets which contain any violence words. We prepared a word list containing 237 violence words in Arabic. Table 1 shows examples of some of the words that are used for the basic filter. We used the different stems of the word because the Twitter filter does not have an Arabic morphological analyser. So the different stems of the word are added manually to the list. The following examples show different stems of the word kill (قتل , kill),(قاتل , killer),(قتيل , one had been killed), (قتلى , many have been killed), and (مقتل , report killing action).

The Twitter word filter only selects tweets containing the word exactly in the form in which it occurs. In Arabic however there are three sets of 'confusing letters' which people may confuse in writing because they are very close, and the correct one to use in a word depends always on the position in the word. The first set includes { ا ، أ ، إ } ; the second set includes { ى ، ي }, and the confusion happens if it comes at the end of the word. The last set includes letters { ه ، ة } and confusion also happens when it comes at the end of the word. Darwish (2002) proposed a letter normalization technique to handle the first two sets of letters by normalizing them into one letter. We reversed his work to extend our violence word list with other possible forms of the same stem. For each stem in the list beginning with one of the letters in the first set, we automatically generate two other stems starting with the other possible letters e.g.

word like (أحتجاز , Detention) is extended with two more words (إحتجاز ، احتجاز). The same technique is applied to all stems that end by any letter in second and third sets.
In addition to these manually generated stems, we used some morphological prefixes and suffixes i.e ال ، وال ، بال ، ون ، ات ، هم to generate other forms to extend the word list. After adding the suffixes and prefixes, we noticed that some of the generated word forms turned out to be incorrect. So, the list was reviewed manually and all incorrect forms were removed. Note that not removing these 'words' should not actually affect the quality of the retrieved tweets as they will not appear in the text. It is like adding English words to a filter list used to filter Arabic text.
We refer to the dataset we obtained in this process as 'Raw Tweets Text' (RTT).

2.2. Advanced Filter

The Twitter filter is very basic and linguistically limited. The Raw Tweets Text (RTT) which is collected from Twitter using the previously discussed processing steps is still very noisy. So many tweets which contain a mention to one or more of the violence words are not real violence reportings. So, we introduced a second, more advanced filter to filter the following types of tweets:

Redundant Tweets There are two types of redundancy in RTT. The first one is explicit redundancy where the same tweet is posted by different users with the same text or re-tweeted by other users, the second type is implicit redundancy where the tweet text is modified by adding some hash tags or some other minor modifications in formatting. To reduce redundancy all tweet text is normalized and hashed. All of RTT is normalized by removing hash tags, line breaks, separators, and extra spaces. Also, Darwish's letter normalization (Darwish, 2002) is applied to normalize confusing letter sets which have been described in section 2.1.. Secure Hash Algorithm (SHA-1) (NIST, April 1995) is used to hash every normalized tweet text and any other tweet text that has the same hash value is removed.

Emotional tweets We explicitly exclude tweets that contain 'emoticons' as we see the reporting of different types of violence as a practical action not an emotional one. A list of emoticons used in our filters is attached to the dataset.

Short Tweets We make the simplifying assumption that short tweets which contain fewer than 5 tokens do not tend to report a violence incidence. So, all tweets with length less than 5 tokens are filtered out.

Sexual Adverts One of the most frequent types of tweets on Twitter are sexual adverts. Advertisers tend to add a list of hash tags and keywords to help users find their adverts through Twitter search. Unfortunately, this list contains the word (اغتصاب, Rape) which is also used to describe one type of human rights abuse. The following two examples show different uses for the same word i.e. rape where is the first example used as a kind of HRA and in the second one used as an sexual movie advert.

اغتصاب السوريات في سجون النظام.....فيلم وثائقي ١٤دقيقة

"Rape in the Syrian regime's prisons ..... documentary film 14 minutes"

الي يبي افلام اغتصاب يسوي رتويت

"Who wants rape movies , he has to make re-tweet"

We used the list of words and hash tags that are used by advertisers to filter out that type of tweets. It is a short list of 15 words i.e. raping, sex, ..etc. This list works efficiently in filtering out sexual advert tweets from the RTT.

3. Classification Scheme

The tweets collected as discussed in the previous Section were manually classified. We started with a pilot scheme containing five manually defined classes. We used 31 test questions distributed over the five classes. Then, we launched a pilot classification job with 1,025 tweets asking for 3 judgements for each tweet. After analysing the results and studying the missed questions in addition to contributor contentions, two more classes were added to the class list, and the class descriptions were modified. The result were seven violence classes. The test questions were also increased to 206 to get a more accurate evaluation of the contributors and give every contributor enough training to understand the task. The number of required annotators for each tweet text was also increased to five instead of three and more annotators are involved incase of low inter-rater agreement (Snow et al., 2008). Also, such increase helps in reducing the chances of accidental personal biases (Artstein and Poesio, 2008). The class definitions were then reviewed by experts in the Minority Rights Group (MRG). The result is the following class definition, that was used in the actual crowdsourced annotation. For each class we present examples, first in Arabic and then in English translation.

HRA

A tweet which describes (an) incident(s) which violates internationally recognised Human Rights Laws. Such abuses are distinguished from general criminal behaviour, in that they are usually committed by members of an overriding authority or group with an agenda.

جيش الأسد في دمشق يرتكب مجزرة مروعة راح ضحيتها عشرات الأطفال داخل مدرستهم فيديو صور

"Assad's army in Damascus commits horrific massacre in which dozens of children die inside their school (video images)"

اعتقال سائحة سعودية في الإمارات لانتقادها بعض مظاهر الفساد

"Saudi woman tourist arrested in UAE for criticising signs of corruption"

Political opinion

A tweet which illustrates a subjective view or judgement on a particular political issue that is not based on fact or objective knowledge. These texts may contain emotive words that signify a clear bias towards or against a party, situation, law etc.

المضحك في الأمر أن بعض الدّول تصدر وتحارب الإرهاب في نفس الوقت!

"It's funny how some countries export terrorism and wage war on it at the same time!"

Accident

A tweet which reports an unfortunate incident that happens unexpectedly. That incident has to be unintentional and resulted in damage or injury.

٢٤ قتيلا على الأقل في غرق مركب لمهاجرين غير شرعيين قبالة سواحل اسطنبول

"At least 24 killed as illegal migrants' boat sinks off the coast of Istanbul"

Crime

A tweet which describes a criminal act on persons or property. These actions can also take the form of negligence by state actors that infringe upon the welfare and rights of individuals, and which is legally prohibited by the laws of the country.

قتل ابن عمه ووضعه في كيس شوال داخل احد البنايات قيد الانشاء بعمان

"Killed his cousin and left him in a sack inside an unfinished building in Amman"

Conflict

A tweet which describes a physical altercation (often with weapons) between two or more belligerents.

مقتل العشرات في اشتباكات بين حوثيين وقبليين وأنصار الشريعة في اليمن

"Dozens killed in clashes between Houthis, tribal forces and Ansar al-Sharia in Yemen"

Crisis

A tweet that suggests a disaster. This does not necessarily have to arise from direct acts of violence, conflict or abuse, but could also be due to the indirect implications of these and other factors which consequently leads to a humanitarian situation.

انتهاء إضراب العاملين في مطار رئيسي بسريلانكا

"Strike ends at Sri Lanka's main airport"

غضب بين طلاب معهد النور للمكفوفين بسبب إضراب النقل المدرسي

"Students' anger over school bus strike at Noor Institute for the Blind"

الحكومة الأردنية تقرر استدعاء السفير الأردني في إسرائيل على خلفية الاعتداءات الإسرائيلية في القدس

"Jordan's government decides to recall its ambassador in Israel as Israeli attacks continue in Jerusalem"

Violence

A tweet which describes to some degree of detail a conscious act of physical aggression. These texts may contain information regarding the belligerent(s) and/or victim and the nature of aggression as well as other relevant information such as number affected, date, location et cetera.

اصابة أربعة أشخاص أثر مشاجرة بدرب الأربعين

"Four injured in brawl in Darb Al-Arbaeen"

مظاهرات ضد السلفيين في ألمانيا تتحول إلى أعمال شغب

"Anti-Salafi demonstrations in Germany turn into a riot"

Other

A tweet that does not fall under either of the above classes and is clearly irrelevant in the context of the work described here.

الإهمال بعد الإهتمام هو قتل نفس بريئه بغير حق !

"To lose interest in someone you once cared for is to kill an innocent soul unjustly!"

اجمل جريمه وافضل الجرائم ! هي قتل الاشتياق بسلاح اللقاء !

"The best and most beautiful of crimes is to put an end to yearning by meeting!"

تكلم وبنفس الوقت تقتل اللي يرفع ضغطك

"Speak, and at the same time you will kill the person who's raising your blood pressure"

4. Annotation

We launched our dataset for classification on the popular crowdsourcing platform CrowdFlower[1]. Each tweet text was classified by at least five different contributors. In fact, some tweets were annotated by up to ten contributors. A set of 206 tweets manually classified by an expert was used as test questions to train new contributors and filter them to be sure they understood the task. The test questions were also used for monitoring contributor performance; every contributor had to maintain a 70% level of accuracy to continue working on our task. CrowdFlower assigns an accuracy score called Trust Score (TS) for every contributor working in the job. This score may be reduced if the contributor answers one of the test questions incorrectly. After finishing the job, all judgments on each tweet text are aggregated with a Confidence Score (CS). It is a quantitative measure of the level of agreement between contributors weighted by their trust scores as shown in Equation 1.

[1] http://www.crowdflower.com/

Violence Class	Individuals		Aggregate	
	Count	%	Count	%
crises	4,066	3	274	1
violence	6,823	4	487	2
accident	5,679	4	558	3
crime	10,942	7	1,331	7
HRA	19,079	12	2,367	12
conflict	23,555	15	3,189	16
opinion	29,556	19	4,261	21
other	56,834	36	7,684	38

Table 2: Individual and Aggregated Classes Distribution Over the Dataset

C_i The Class i where $1 < i < 8$

K The set of all contributors judging a certain tweet.

M The set of contributors classifying a certain tweet to a certain class C_i.

TS_j The Trust Score for a contributor j where $0 < j < k$ and $0 < TS_j < 1$.

$$CS(C_i) = \frac{\sum_{m \subset M} TS_m}{\sum_{k \subset K} TS_k} \quad (1)$$

The final assigned class for each tweet text is chosen based on the greatest confidence score.

5. Evaluation

We initially collected 557,576 tweets using the basic filter and that number is reduced to 76,619 tweets after applying the advanced filter. The currently annotated dataset comprises 20,151 tweets.

Because the dataset was launched for annotation in different batches, we decided to release two datasets. The first one is Crowdflower's job report which contains some duplicates (redundantly annotated tweets). The second dataset is the cleaned version with unique tweets together with their annotations. A new confidence score is calculated using different annotations from different jobs if any.

A basic analysis of the annotations obtained has been carried out and Table 2 shows the frequency counts of each class considering the aggregated class, i.e. the class that has the highest confidence score across all annotates; and individual classes. Considering HRA as an example, there are 2,367 i.e. 12% tweets have been classified as HRA using the highest confidence score. Also, there are 19,079 annotations of HRA distributed over different tweets. To visualize this information, Figure 1 shows the class distribution by aggregating the class and individual contributors' classification. The difference between the aggregate classifications and individual judgements is only around $\pm 2\%$. It is a preliminary indicator about the annotations' homogeneity.

Further analysis for the inter-rater agreement across different classes is carried out. We used Fleiss' kappa to assess the agreement on the annotations of these tweets (Fleiss, 1971). Because of the overlap between different classes it is not always easy to decide on the category of the tweet.

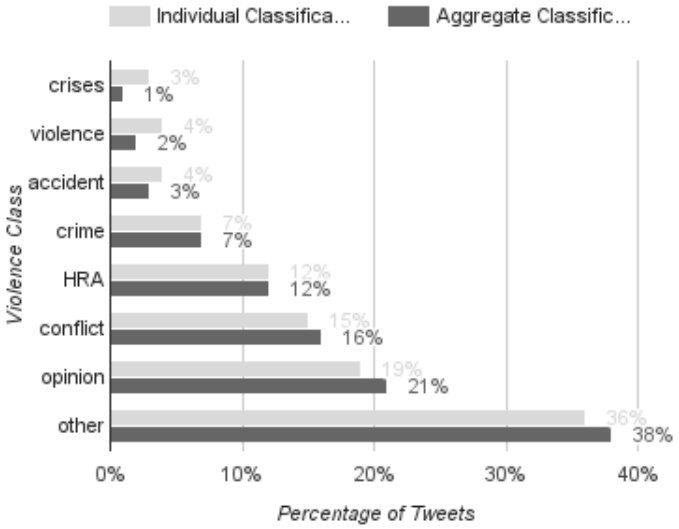

Figure 1: Classification Distribution

So, we used the confidence score to evaluate different subsets of the dataset based on the assigned confidence score for each tweet. The number of tweets is different for each confidence score as shown in Figure 2. Table 3 shows the k score for each class and the overall k score. The first subset which contains all tweets has confidence score ≥ 0.2. This dataset includes all annotated tweets. Other subsets are created and evaluated based on their confidence score i.e. 0.4, 0.6, 0.7, 0.8, 0.9, 1.0.

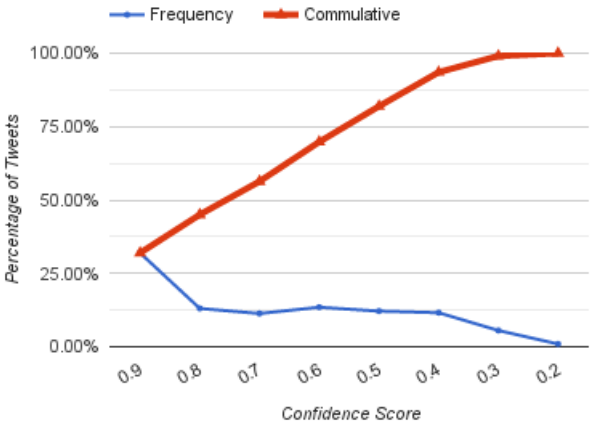

Figure 2: Tweets Histogram Over Confidence Scores

Figure 3 shows the Fleiss' kappa scores for each class. The inter-rater agreement for Crime and Conflict classes are quite low even with tweets that show high confidence. While the inter-rater agreement for the 'Crime' category dramatically increased when considering tweets with $CS \geq 0.8$, the class 'Conflict' still has low inter-raters agreement. For all other classes, the inter-rater agreement is getting higher as we limit the dataset to those tweets with $CS \geq 0.7$. In general, the Kappa scores could be inter-

Class	0.2	0.4	0.6	0.7	0.8	0.9	1.0
crises	0.33	0.37	0.51	0.64	0.71	0.94	0.93
violence	0.53	0.57	0.72	0.83	0.88	0.99	0.99
accident	0.69	0.74	0.82	0.89	0.93	0.99	0.99
crime	0.17	0.19	0.21	0.25	0.27	0.91	0.98
HRA	0.55	0.57	0.67	0.78	0.85	0.99	0.99
conflict	0.13	0.14	0.15	0.12	0.10	0.37	0.37
opinion	0.46	0.50	0.64	0.76	0.83	0.99	0.99
other	0.61	0.63	0.74	0.83	0.89	0.99	0.99
Overall k	0.50	0.54	0.67	0.78	0.85	0.99	0.99

Table 3: kappa Analysis for for different datasets each of them is selected based on the CS score

preted as Substantial agreement (Landis and Koch, 1977). The most confusing classes are 'HRA' , 'Crime' and 'Conflict'. Most human rights abuses are a subset of all violence classes as it could be seen as a violence under certain conditions. The following two example show the use of the word ("شهيد", martyr) to refer toŁa person who was killed. The first example shows that person is killed in a battlefield, however, in the second example it refers to someone who was killed during an attack against a civilian. The first one is classified as 'Conflict' while the second one classified as 'HRA'. To do that distinction, some knowledge of the human rights declaration is required.

Example 1

ألف شهيد لتنظيم الدولة بغارات التحالف في ثلاثة أشهر

"1000 martyr for ISIS in a coalition strikes in three months"

Example 2

ارتفاع حصيلة مجزرة القابون التي ارتكبتها قوات الأسد الى ١٧ شهيداً معظمهم من الأطفال.

"Death toll rises to 17, mostly children, in Kaboun massacre by Assad forces"

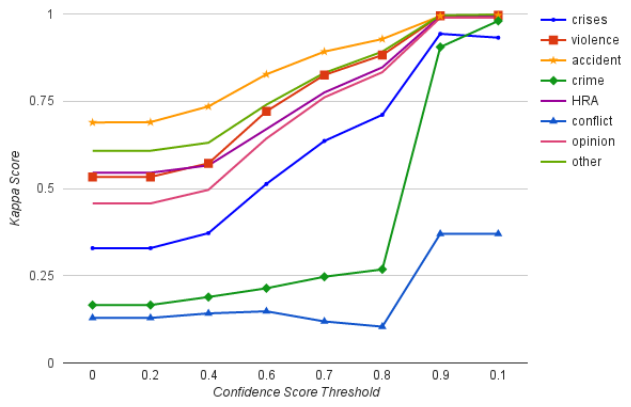

Figure 3: Kappa Analysis for different classes with different CS score

6. Conclusion

In this paper we described a dataset of Arabic tweets reporting violence, and the classification scheme covering seven types of violence we used to classify them. The tweets were collected from Twitter using a series of filters, and then presented on CrowdFlower for manual classification using crowdsourcing. Our evaluation shows that crowd classification is overall reliable[2] and could be used for further research on violence on social media.[3] In future we plan to enrich the dataset by training an automatic classifier to enrich the advanced filter. The dataset is available for free download.

Acknowledgements

We would like to acknowledge support from InnovateUK through a Knowledge Transfer Partnership (KTP) project between MRG and the University of Essex. Also, a special thanks to Mahmoud Swid for his valuable suggestion in designing the violence classes and classifying the test questions.

7. References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Darwish, K. (2002). Building a shallow arabic morphological analyzer in one day. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, SEMITIC '02, pages 1–8. ACL.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

NIST. (April 1995). Secure hash standard. In *Federal Information Processing Standard*. FIBS Publication.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *EMNLP*, pages 254–263. ACL.

[2]Not for all individual classes though

[3]Note that it is impossible to have a truly objective judgement in many of these cases.