An NLP-Powered Human Rights Monitoring Platform

Ayman Alhelbawy[a,c,d], Mark Lattimer[d], Udo Kruschwitz[a], Massimo Poesio[b], Chris Fox[a]

[a] *University of Essex, Wivenhoe Park, Colchester CO4 3SQ*
[b] *Queen Mary University of London, Mile End Rd, London E1 4NS*
[c] *Fayoum University, Qesm Al Fayoum, Al Fayoum, Faiyum Governorate, Egypt*
[d] *Minority Rights Group International, 54 Commercial Street, London E1 6LT, UK*

Abstract

Human rights organisations have been working for decades on monitoring human rights and their violations across the world. However, monitoring of violations of international humanitarian law more specifically has not developed as strongly and this has been compounded in recent years by a growing lack of access to zones of conflict. Part of the reason is that it can be risky for victims, witnesses, and activists to get interviewed by researchers working on behalf of human rights organisations. As a result, there is an actual practical need for tools that support human rights monitoring in conflict zones, be it for specific organisations or as an instrument for reporting this to the general public. This paper presents a platform that fills this gap deploying state-of-the-art natural language processing and social media mining. More specifically, we report on mining and classifying Arabic Twitter in order to identify potential human rights abuse incidents in a continuous stream of social media data within a specified geographical region. We find that deep learning approaches outperform more traditional classifiers (SVMs with feature engineering). This classification pipeline is part of an analyst's tool bench that also collates reports submitted by experts and individual witnesses within the same region. Apart from the scientific insights we demonstrate the viability of the framework which has been deployed as the *Ceasefire Iraq* portal for three years and which has collected thousands of witness reports from within Iraq.

Keywords: Crowdsourcing, Human Rights Monitoring, Machine Learning, Natural Language Processing, Social Media, Twitter, Ceasefire, Applications

1. Introduction

Ever since the *Universal Declaration of Human Rights* in 1948 (The United Nations, 1948), many human rights organisations have been established with a core mission of monitoring human rights and their violations in different countries across the world. Until recently this work was conducted using largely the same underlying methodology (Alston and Crawford, 2000)

More recently, technological advances have made it possible to deploy frameworks that allow the recording of potential human rights violations through Web services allowing organisations to conduct their mission in more productive and efficient ways. A prime example of this trend is the fast-growing deployment of the open-source platform *Ushahidi*[1], initially developed for collecting eyewitness statements to map reports of violence in Kenya after the post-election violence in 2008. By employing a crowdsourcing approach, i.e. anyone can contribute, the platform can tap into communities and witnesses that were previously difficult to reach out to. Reports can be submitted anonymously and the platform offers a high level of application-side security. Ushahidi has since been deployed in a wide range of human rights reporting, election monitoring and crisis response projects. Note however that any such application can only be a tool to *assist* the monitoring of human rights abuses as none of them actually *replace* the human analyst.

Apart from simple technological progress, there have been two further major developments that offer new ways of working for human rights organisations – progress in Artificial Intelligence (AI) and the ever-growing availability of data. Rapid progress in AI (Russell and Norvig, 2016; Müller and Bostrom, 2016) means that AI applications using machine learning are now ubiquitous, be it to rank the results of a Web search engine, to control the electronics of a car or to classify social media feeds into categories which could include the identification of potential human rights violations. In particular the shift from sparsely available data (of high quality) collected by a team of experts to massive streams of potential input signals in social media (of variable qual-

[1] `http://www.ushahidi.com`

ity) offers completely new opportunities but also comes with caveats. Such data does not have to be textual but also includes images, videos and other formats. For example, satellite imagery is now being employed by human rights organisations. A recent example is the satellite image analysis in a project conducted by *Human Rights Watch* to demonstrate the near total destruction of 214 villages in Burmas Rakhine State[2].

Any technical solution proposed in the general problem area of human rights monitoring does however face a range of challenges which vary depending on the actual application. For a crowdsourcing application the main challenge is to reach out to the target audience in the first place in addition to providing a platform that users can trust and easily use. Furthermore, there is the inherent problem of assessing how reliable each individual report is. Looking at AI-powered approaches that typically aim at classifying massive amounts of data into pre-defined categories, the main challenge lies in having enough reliable training data and employing suitable machine learning algorithms that offer sufficiently high-quality classification.

This paper proposes a platform that brings together the two strands. It can be seen as an analyst's tool bench offering the monitoring of human rights violations within a specified geographical region with, on the one hand, reports being submitted by experts as well as individual witnesses through a dedicated, structured reporting system and, on the other hand, a continuous stream of social media data that have been classified as signals of potential human rights violations within the same region. Moreover, the tool serves a dual purpose as in addition to its use for human rights monitoring within an organisation it is also an instrument for reporting this to the general public.

To the best of our knowledge, this is the first portal of its kind that combines the two strands, and we demonstrate the viability of the framework which has been deployed as the *Ceasefire Iraq* portal[3] for three years and which has collected thousands of witness reports from within Iraq. The analysis of these reports has led to a series of publications (policy documents) by *Minority Rights Group International*. The active response in social media, e.g. via tens of thousands of shares on Facebook, demonstrates that it

[2] `https://www.hrw.org/news/2017/09/19/burma-satellite-imagery-shows-mass-destruction`
[3] `http://iraq.ceasefire.org`

also serves the second intended purpose, offering a reporting tool to the general public. Our immediate next steps include the deployment of the framework in the wider Middle East and North Africa region.

This paper is organised as follows. Section 2 provides a detailed discussion of related work. It provides an overview of how human rights organisations traditionally operate and how recent technological progress and advances in AI and natural language processing (NLP) have impacted their work. We also look at existing tools and frameworks. This discussion will conclude with the identification of shortcomings in existing approaches and motivate our contribution.

Section 3 introduces our human rights monitoring platform that emerges from the identification of the gaps in the landscape of existing solutions. The practical deployment as the *Ceasefire Iraq* portal and added organisational structure and user security models are discussed in detail in Section 4. Section 5 discusses our NLP-based approach to automatically identify potential Human Rights Abuse (HRA) posts on Twitter. It also provides the experimental results achieved by the approach, and the field results.

Section 6 will reflect on the results and impact that have emerged from the deployment of the system. We will also provide some insight into lessons learned that should be of interest to our readers. Finally, our conclusions and suggestions for future directions are presented in Section 7.

2. Background

In the decades following the adoption of the Universal Declaration of Human Rights in 1948 (The United Nations, 1948), a global movement for human rights has taken shape across the member states of the United Nations. Organisations across different sectors have pursued a wide range of approaches to the challenge of respecting, protecting and fulfilling human rights, in which the monitoring of violations has formed an essential element. The persistence of human rights violations including gross violations in every world region today is evidence of the size and complexity of that challenge. In recent decades technological tools have rapidly developed to assist in this task, but to understand their relevance and application it would be helpful to review briefly the evolution of human rights monitoring in general as shown in Section 2.1. The growing contribution of technology to support human rights monitoring in different countries is discussed in Section 2.2. Section 2.3 will highlight a specific platform, *Ushahidi*, that has emerged as a viable tool that we also adopt as a backend in our approach. Previous work on using NLP technology to support human rights monitoring and related tasks is reviewed in Section 2.4 The main challenges faced by organisations working within the broader scope of human rights monitoring are summarised in Section 2.5.

2.1. Development of Human Rights Monitoring

International concern for atrocities committed in other parts of the world is arguably as old as recorded history, but modern campaigns for human rights abroad are often traced back to the movement against the international slave trade in the 19th century (Hochschild, 1999, 2005). Whether it was detailing the abuses committed by slavers or highlighting the appalling conditions in European colonies, such movements for change followed a familiar pattern: the presentation of documentary and photographic evidence by activist investigators or official fact-finders to a wide audience to expose the nature of abuses being committed, elicit sympathy for the victims, but also increasingly arouse a sense of injustice based on their status as holders of rights. In many respects this fundamental set of techniques still forms the basis for much human rights work today, with UN special rapporteurs and international NGO investigators despatched from Geneva, New York or London to spend a week or two in a country under scrutiny, interview victims and civil society, and return to present a report some months later to the UN Human Rights Council, national authorities or the international media.

The further development of international legal standards on human rights following the 1948 Universal Declaration and the growing professionalisation of human rights work led in turn to the development of related approaches to monitoring and documenting human rights observance, such as:

- Monitoring the application of national laws and practices to ascertain their effect on human rights;
- Undertaking statistical and social science research to analyse the fulfilment of human rights in given populations and the prevalence of discrimination on a range of grounds;
- Monitoring news reports and records to identify both specific violations and to build a picture of emerging patterns of violation;
- Using the outputs of monitoring and documentation to substantiate claims for redress before national courts or international human rights courts or monitoring bodies (Puttick, 2017).

While inequalities in development and application of the rule of law across world regions meant that the state of human rights monitoring and documentation itself displayed marked disparities between states, a particular problem was presented by armed conflict. Broadly speaking, the monitoring of violations of international humanitarian law (IHL) or the law of armed conflict has not developed as strongly as human rights monitoring (Lattimer and Sands, 2018) and this has been compounded in recent years by a growing lack of access to zones of conflict (Raad Al Hussein, 2016). Our approach is focusing on finding solutions for the lack of access to zones of conflict. We identify crowdsourcing as an effective paradigm for monitoring human rights in conflict areas bringing together automatic social media mining and online reporting allowing civilians and researchers on the ground to directly report observed incidents.

2.2. Growing Contribution of Technology

The development of the internet and the spread of mobile telephony have accelerated the pace of change in human rights monitoring and, in some respects, altered its character. However, monikers such as the Facebook revolution or the Twitter revolution applied to socio-political movements, including in the Middle East, are misleading with regards to human rights developments. Changes cannot be attributed to one application, or even to social media as a whole, but are rather due to larger, generalized effects that come from a confluence of technologies, in the context of wider social awareness and human rights education, including in developing countries.

Specific examples of the contribution of new technologies relevant to human rights monitoring and documentation include:

- Digital collection of monitoring information to facilitate statistical analysis, and digital storage off-site to protect security of information and human rights defenders from repressive measures;
- Availability of sophisticated encryption techniques to safeguard security of human rights communications;
- Crowdsourcing and geo-mapping platforms to pool monitoring information from users and support analysis;
- Analysis of satellite imagery to provide evidence of certain large-scale violations, including destruction of buildings, villages or habitats, or to facilitate location of mass graves;
- Software enabling meta-data to be embedded in digital documents, photographs and videos, assisting in the verification of evidence and chain-of-custody procedures required in legal proceedings.

The significance of any particular technological development is perhaps less important, however, than the huge expansion of internet access and smartphone usage in the developing world. This marks a transformation in which human rights monitoring is no longer the exclusive domain of professionals from the developed world but is now increasingly a practice also owned by activists from communities directly affected. The work discussed in this paper is aimed at exploiting this opportunity, without losing track of the fact that the positive advances promised by each technological innovation are inevitably accompanied by potential threats or negative implications.

2.3. Ushahidi

One modern development that has already had a very beneficial impact on human rights monitoring is the development of technology to collect data from non-experts.

Ushahidi is a good example of the new tools that have become available. It is an open-source crowdsourcing platforms that was initially developed to map reports of violence in Kenya after 2008 post-election violence (Bailard and Livingston, 2014). It has been widely used to monitor elections in different countries, e.g. in Kenya again in the 2017 elections[4], but also, for example, to document post-election violence following the US elections in 2016.[5] It has also been deployed for crisis response and advocacy & human rights and such applications range from recording violations of media freedom and threats to media workers in countries of the European Union[6] to mapping technology-based violence against women.[7]

Its maturity, open-source nature and large user community were the main factors for us to adopt Ushahidi as the backbone for our human rights monitoring platform. We should note however that we had to develop additional layers of security and provide support to collaborating organisations and will expand on these issues later on.

2.4. NLP Technology and Human Rights

Although machine learning and natural language processing are well-established research areas with steady progress in a variety of fields and applications over several decades, we have recently witnessed a paradigm shift when neural networks have started outperforming many more traditional machine learning applications. The most notable evidence for that is the proportion of research papers dedicated to neural networks and reporting significant advances over alternative methods at top academic conferences such as ACL[8], EMNLP[9], WSDM[10] and NeurIPS[11].

There has however only been limited interest in applying NLP and ML technologies for human rights monitoring, even in the broadest sense. There are nevertheless related areas that did attract the interest of researchers, much of it applied to mining

[4] https://uchaguzi.or.ke/
[5] https://documenthate.ushahidi.io/
[6] https://mappingmediafreedom.ushahidi.io/
[7] https://www.takebackthetech.net/mapit/
[8] http://www.acl2018.org
[9] https://emnlp2018.org/
[10] http://www.wsdm-conference.org/2019/
[11] https://nips.cc/

and analyzing social media in one way or another, and we will provide a brief overview here. Note that we will drill down further into the separate area of Arabic NLP when we discuss our approach to identifying potential human rights violoations in Twitter in Section 5.

NLP technology has been used successfully to identify cybercrime, cyberbullying, and violence detection (Whittaker and Kowalski, 2015; Kontostathis et al., 2010; Reynolds et al., 2011). We can distinguish two main lines of research in detecting violence on the Web. The first is to analyse videos using computer vision techniques (Nievas et al., 2011; Datta et al., 2002); the second is using text mining techniques (Nobata et al., 2016; Chandrasekharan et al., 2017). There has been much research on violent content detection in English social media but much less so on Arabic although there is now a growing body of research that starts building up, e.g. work on abusive language detection on Arabic social media, e.g. (Mubarak et al., 2017), as resources for Arabic in general and applied to social media more specifically have grown substantially, e.g. (Diab et al., 2018; Zirikly and Diab, 2015; Abdul-Mageed et al., 2014; Awad et al., 2018; Aldayel and Azmi, 2016).

A probabilistic violence detection model to identify text containing violent content based on word prior knowledge about whether the word indicates violence or not was proposed by Cano Basave et al. (2013). To build a training corpus, they used *OpenCalais* and *Wikipedia* documents, as well as *Wikipedia* and *YAGO* categories. The dataset was built to classify a set of categories including *Crimes, Accidents, War* and *Conflict*. Everything else, e.g., documents on *Education* and *Sports*, was tagged as *Non-violence related*. We considered the use of these datasets for our purposes; but unfortunately, *OpenCalais* does not support Arabic, and the number of documents corresponding to violence in Arabic Wikipedia is very small making the source dataset very sparse.

An offensive content detection model was proposed by Chen et al. (2012) to detect offensive language in social media. They introduced a set of lexical features like simple bag-of-words and n-grams, in addition to hand-written syntactic rules to identify name-calling harassments. They used traditional machine learning techniques including Naïve Bayes and SVM to learn a classifier. Their proposed system employs a user profile capturing the user's English writing style.

Harassment detection on the Web is another area of application of NLP techniques. Yin et al. (2009) proposed a model for harassment detection on the Web using both local features and contextual features. Local features are n-grams weighed using TF-IDF. Contextual features are also used, under the assumption that each post is surrounded by other posts from the community; chat-rooms and forums post style.

In summary, a variety of approaches have been used to tackle related problems for English, but to the best of our knowledge there is no previous work on the specific issue of human right violation detection, let alone work applied to the Arabic language in this context. Also, the accuracy achieved in previous work still tends to be rather modest. We will present our own approach to the problem in Section 5.

2.5. Challenges in Human Rights Monitoring

The traditional approach of human rights organisations is to use highly trained professionals (researchers) to gather and verify information. These researchers visit sites of human rights abuse and conduct detailed interviews with victims and witnesses (Heinzelman and Meier, 2015). To the existing challenges for the practice of human rights, referenced earlier, can therefore be added a new set of challenges for monitoring presented by advances in technology. In conflict situations, or in states with authoritarian governments, the democratization of human rights monitoring enabled by contemporary technology potentially places at risk a large number of monitors who might be targeted because of their activism. During the conflict in Syria, for example, media activists who sought to record the effects of bombing and other attacks in their neighbourhoods suffered high rates of fatality or injury. So, new challenges of verification, information security, and users awareness are raised.

Puttick identifies four categories of challenges for civilian-led monitoring in addition to digital and physical security risks (Puttick, 2017, 24-31):

- **Information deluge:** Data-mining techniques in particular, as well as crowdsourcing, have to deal with the huge and ever-growing mass of information presented online, most of it irrelevant to the purpose at hand.

- **Quality control:** multiplying the number of monitors can lead to inconsistencies, duplication of effort, and much greater variances in the quality of information produced.

- **Verification:** more fundamentally, there is a perception that crowd-sourced information is unreliable or untrustworthy. Although the reliability of human rights claims made by official bodies, including governments, is often exaggerated, there is no doubt that information gathered from a very wide range of different sources is likely to include some information that is falsified or misrepresented, deliberately or otherwise.

- **Ethical issues:** finally, a wide range of ethical challenges includes threats to privacy in the use of big data technology, and the safeguarding of interviewees and other human rights victims. Non-professional monitors, not schooled in the principle of 'do no harm', may be less rigorous about seeking informed consent and more inclined to share personal information online. Another problem with sharing content online "is that the platforms on which activists rely – such as Facebook, Twitter and YouTube – are private companies governed by corporate interest, whose terms of service are not necessarily tailored towards protecting human rights."(Puttick, 2017, 29)

2.6. Concluding Remarks

There are a number of conclusions we can draw from this discussion which will motivate our work. First of all, we conclude that the traditional approach to human rights monitoring has changed in recent years and that commonly applied methods are often simply no longer possible to apply. At the same time we observe that technology has made significant progress and that in particular advances in machine learning to mine social media for text clasification have been made. This goes hand in hand with a better understanding of how to exploit crowdsourcing methods to extract meaningful information from social media streams. We also witnessed the emergence of dedicated crowdsourcing platforms that can be deployed for online reporting allowing civilians and monitors on the ground to directly report incidents of human rights violations anonymously.

The gaps identified in addition to recent developments discussed motivates a framework that serves the dual purpose of reporting human rights violations to the general public as well as a practical workbench for analysis within human rights organisations. After all, such platform cannot operate without the *human in the loop*. Mining social media using NLP technology may help in finding early signals of potential human rights violation providing analysts with more evidence and incidents and possibly links to new witnesses. However, online reports will still need to be manually assessed and anonymized before they can be placed online for anyone to see. Apart from preserving the anonymity of witnesses this protects victims and activists and allows the collection of additional evidence without the need for personal interviews.

3. Ceasefire: A Platform to Support Grassroots Involvement in Human Rights Reporting

The exponential growth of data on the Web and, more specifically, in social media has contributed to the perception that we no longer deal with simply larger-scale data but with what is commonly referred to as *Big Data*.[12] Tapping into this resource offers insights into a wide range of patterns and we argue that this will also benefit human rights monitoring.

The platform presented in this paper aims at painting a picture of human rights violation and abuse within a specific geographical region by bringing together two streams of information: actual reports by witnesses, monitors and any civilian accessing the system, and relevant information identified in a continuous stream of social media. In other words, it supports user involvement, the merging of information coming from users and information coming from social media, and human rights organisations reporting in one place, as shown in Figure 1. It consists of two main components which

[12] The term *Big Data* is not well-defined and is used with different meanings, but most typically to refer to large data sets which are very hard to process using traditional approaches due to their size and complexity, e.g. see (Manyika et al., 2011).

we will discuss in the following section: an online reporting tool, and an NLP-based social media monitor (in this specific instance we use Twitter).

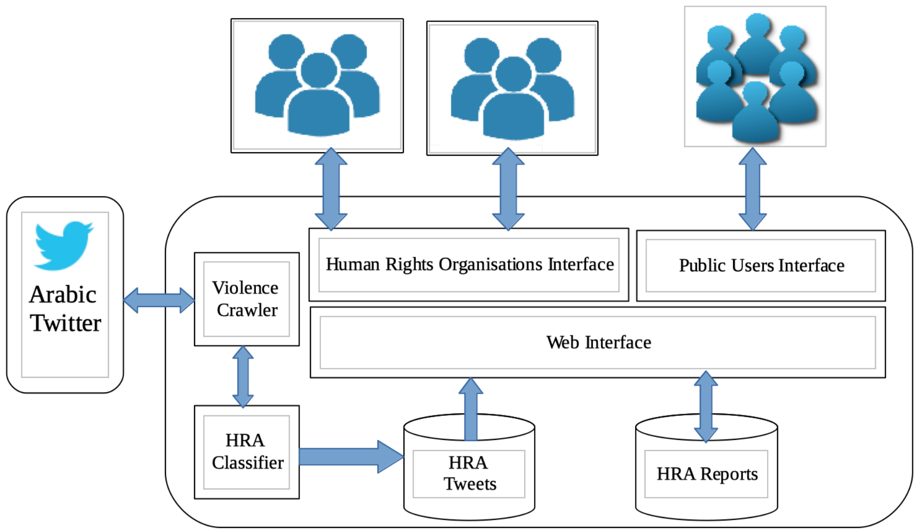

Figure 1: Ceasefire: a Framework for Reporting and Monitoring of Human Rights Abuses

Online reporting provides a secure crowdsourcing facility for victims, witnesses and activists to report human rights violation incidents. This first component can be accessed by users who intend to report their experience to a human rights organisation. This component makes the information available to local and international human rights organisations while taking care of data security and accessibility. This part of the platform was developed using Ushahidi as the backend, but with additional structural and security modifications discussed in Section 4.

The second component of our platform, based on ML-based classifiers that are applied to the output of an NLP-pipeline, is used to discover human rights violations reported in social media such as Twitter. Its purpose is to enrich the actual witness statements and reports with additional signals mined from what locals within the region are reporting, particularly from areas where human rights organisations have limited access. We will discuss this component and its underlying methodology in more detail in Section 5.

The *Ceasefire* platform was developed together with *Minority Rights Group Inter-*

national using Iraq as a case study, but extensions to other countries in the Middle East and North Africa (MENA) region are currently under development.

4. Ceasefire Deployment

We will now provide a more in-depth overview of the Ceasefire platform with reference to its first major deployment.

4.1. The Online Reporting Service

The first key component of the Ceasefire platform is an *Online Reporting Service* that allows any user–victim, witness, activist, or human rights organisation–to submit reports of human rights violation incidents. Two reporting interfaces are available, one for the general public and another one that is dedicated to human rights organisations. The data collected through the *Online Reporting Service* also paints an overall picture of the human rights situation at a specific geographical location. Figure 2 is a screenshot of the main page of Ceasefire, which shows a map of the geographical distribution of the submitted reports in Iraq categorised by the type of violation (such as physical abuse, psychological abuse, etc.).

As concluded in the previous section, we identfied several benefits of an online reporting service for the public and participating organisations. One of the main benefits for organisations is that there is no need to expose interviewers to highly dangerous environments, taking the example of Iraq, this would avoid sending anyone to Mosul while under the control of ISIS. From the point of view of the public, the service allows them to report incidents at any time and in a more confidential way than talking with a representative of an NGO, which are generally under surveillance. The feeling of reporting directly to an international human rights organisation (instead of a possibly suspicious intermediary), and the understanding that the information is treated more securely, may also make the public more confident.

The online reporting facility was developed based on the open-source platform Ushahidi 2.7, which is based on PhP and uses MySql for its backend; but several changes were necessary to the core Ushahidi engine to make it applicable in our context, such as adjusting the Arabic right-to-left view and adding a new security model.

In order to get different human rights organisations involved, custom forms were designed to fit their needs. These custom forms were designed by analysing the specific interview forms used by different organisations.

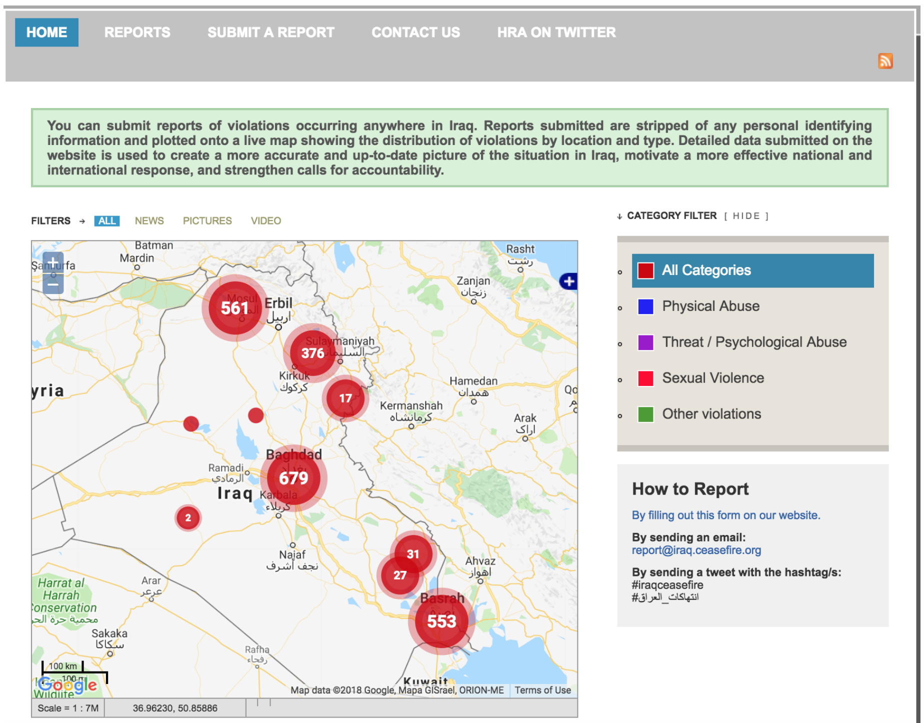

Figure 2: Ceasefire Main Page. Reports are plotted on the Map

Every participating organisation can visualize a statistical analysis over the categories of submitted reports over a specific period of time. It was a core requirement that this would be limited to reports submitted by the organisation's own users or their partner organisations. Figure 3 illustrates an example of the statistical distribution of reports over a three-month period. The categories used were developed and structured by human rights experts.

Online reporting also has some disadvantages, however. The first disadvantage is that it requires internet access, which may not be available in all areas. This problem is however being reduced all the time by the rapid spread of internet-enabled devices. The second problem is making victims aware of its existence. Media such as TV,

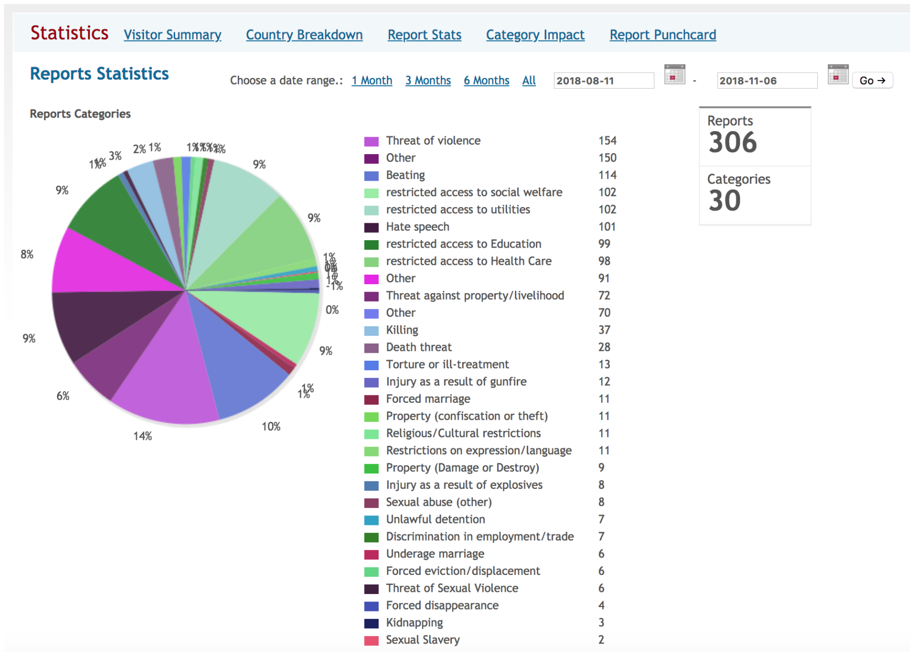

Figure 3: Ceasefire Reports Statistics. Results for a three-month Period.

radio, social media may be used to raise the public's awareness of the existence of the service. In our case study with Minority Rights Group, advertising on social media targeting some areas in Iraq made a noticeable difference on the portal visits and the number of submitted reports. A third problem is that centralising human rights abuse reporting may make it an easy target for governments which do not support such work. That may put victims and reporters at a real risk, because if the government gets access to the reports, it may make use of the information to punish the people involved, or destroy the data. Periodic backups may be a good defense for data destruction, but will not help to protect information about victims and reporters. We will now discuss how we mitigate that risk.

4.1.1. Storing information

Three protection layers are used to deal with unauthorised access, as follows:

1. **Basic user information** is saved in encrypted form.

2. **Incident details** are not automatically posted to the public-facing portal. When a user submits a new report, it is not published until a trained reviewer has anonymized all personal data, places, etc.
3. **All Report Data** are frequently pulled by another secure server, after which all personal and other critical information on the Ceasefire servers is permanently deleted. So, the Ceasefire map continues to work, and the number of reports will remain the same, the reports remain classified according to the categories used, but no identifiable information will be accessible. Also, the Ceasefire servers do not save any information about these secure servers, which pull the data before final anonymization.

4.1.2. Access control

The existing security model in Ushahidi was judged to be insufficient for the Ceasefire requirements. Therefore, new security levels and user access controls were developed.

It is not necessary for users to be registered to submit a report. But unregistered users cannot retrieve their submitted reports for editing. Also, for some partner organisations it was a requirement to register some users who would be able to keep track of their submitted reports. Once a registered user has been authenticated by the Ceasefire engine, it is the Ceasefire security model's role to determines the data the user is allowed to see or modify. Users are organised in groups where every group has its own dedicated access level.

Organisations working in the project have a hierarchical structure, and some organisations are working as partners for other organisations. Every bit of information stored on the Ceasefire platform has a security access level where the user or group who has a higher access level can get access to it. Also, users defined in the same group can get access to all reports submitted by the group. Any human rights organisation may have one or more groups to work with different access levels as defined by the Ceasefire team. Users from partner organisations can also join the organisation's groups. In the Ceasefire Iraq use case, there are different organisations working on the ground in Iraq under the Ceasefire umbrella. That model facilitates the independent operation of different partner organisations and at the same time gives the Ceasefire team access to all reports submitted by different partners. Problems with ‘elevated rights’ can contribute to unintentional data breaches, so Ceasefire enforces access controls on a ‘least privilege model’ - with new users assigned only the most basic level of data access by default.

4.2. The Social Media Classifier: Identifying Human Rights Abuse

The other major component of the framework to compile information about potential human rights violations is the automatic classifier that is applied to a continuous stream of social media feeds. Social media has become a means for people to let their opinions be known. Oftentimes, victims discharge their anger on social media even if they believe no one can or will do anything to relief their suffering. Other people do not trust human rights organisations, and prefer to make their testimony known through social media rather than via reporting to human rights organisations. This may be because they do not know the organisation, or they may find using social media easier, or they do not believe human rights organisations can make any difference.

Figure 4: A Tweet classified as a potential Human Rights Abuse (HRA).

The Ceasefire platform includes a continously running component that monitors Twitter to find tweets which mention some form of human rights abuse (we call such tweets *unintentional human rights abuse reporting*). Figure 4 shows an example of a tweet[13] that was classified as falling into that category and which will then be displayed

[13]This translates to “Mosul today turned into Hiroshima. The federal police exterminate the Mahmudien, Khazraj and Babelbead. A crime committed since the dawn of the day and continues to be committed”.

in the ”HRA on Twitter” section on the Ceasefire platform. While the public-facing portal only ever displays the latest 100 identified tweets, the human rights analyst has access to the full set as the data is saved for more in-house analytical work.

Because Twitter’s terms and conditions prevent users from keeping or redistributing the actual tweets, Ceasefire just keeps the corresponding identifiers. So, when a user navigates to the social media feeds page, the Ceasefire engine calls the Twitter API to retrieve the full tweet information. In cases where for some reason the original tweet has been deleted by the user or by Twitter, it will no longer appear on Ceasefire either. The Ceasefire framework does not keep any personal information about Twitter users either.

We will now turn from the more practical considerations to the core academic contribution. We will in particular explore the Arabic NLP processing steps applied as well as report on experiments we conducted for building a classifier identifying potential human rights violations.

5. Automatically Identifying Potential Human Rights Abuses on Twitter

Our first case study, *Ceasefire Iraq*, was focused on Iraq. Our Twitter mining method was therefore developed and tested on Arabic data. The popularity of social media in the Arab world has grown dramatically over the last decade. According to the Arab Social Media Report, there were 11.1 million Twitter users active monthly in the Arab world as of March 2017, posting on average around 27.4 million tweets per day (Salem, 2017). Social media has become a regular source of daily updated information as people share with others what they like and do not like, their political opinions, their beliefs, and also what they see. Moreover, around 52% of users are reported to share their political views on social media (Salem, 2017). Due to the dramatic problems plaguing much of the Arab world, a proportion of what people report about on social media is violence and human rights abuse. As a result, Twitter has become a common social media forum for people to share their experience.

As discussed earlier, research to detect, for example, offensive and violent content in social media, in particular with a view on cybersecurity and monitoring cyberbullying has attracted a lot of attention, e.g. (Reynolds et al., 2011; Kontostathis et al., 2010; Whittaker and Kowalski, 2015). But to the best of our knowledge there has been no research on human rights abuse discovery in Arabic text which is clearly a serious gap in the light of the earlier discussion. Unlike typical settings in other common classification tasks, as for example sentiment analysis, we are looking at under-resourced languages (Arabic in our current case study) and at non-standard categories (either binary or multi-label). Apart from contributing to the understanding of the problem, the automatic mining of information about potential human rights abuses provides an additional stream of signals that supplements detailed reports and this data actually forms an integral part of the human rights monitoring platform introduced in the previous section. We will now discuss our approach to the problem as applied to the *Ceasefire Iraq* portal.

5.1. Text Preprocessing

Preprocessing platforms for Arabic have started to become more widely available, e.g. (Althobaiti et al., 2014), however processing of social media texts remains a challenge. The first step of preprocessing carried out in our work is removing Arabic stop words and web links from the text. Secondly, a step of orthographic normalization is carried out. Because mistakes in writing Arabic letters like “Alef” and “Yaa” are common, different “Alef” forms are normalised to a single form, and the same for “Yaa” (Darwish, 2002). Finally, all numbers are replaced with one digit as a place holder, preserving the existence of numbers in the tweet text regardless the actual value.

5.2. Morphological processing

Arabic has a complex morphological structure (Al-Sughaiyer and Al-Kharashi, 2004). Various types of affixations are added to the base word to encode grammatical categories like number, gender, and tense. Masculine and feminine forms of a word differ. In Arabic, the single, plural *and double* form of the word are distinguished (double is not considered a plural in Arabic). Also, short vowels called “Diacritics” are not always written and the word with no diacritics could be interpreted as different words. The word “كتب” is a good example as it could be “كَتَبَ” (Kataba) which means “write” in the past tens or “كُتُب” (Kotob) which means “books”. The right interpretation depends on the context.

So, in addition to token features, additional morphological features are extracted to reduce the noise in the vector space. The MADAMIRA package (Pasha et al., 2014) was used to carry out morphological analyses of the text. Table 1 shows the feature vector length when using each feature and an example of the feature when using the word “المصابين” which could mean a couple of injured persons or a group of injured people. The diacritized form means a group of injured people. Both diacritized and non-diacritized are in masculine form. Lemma form means an injured person in singular masculine form. In this example both lemma and stem have the same meaning.

Feature	Description	FV length	Example
Token	The text form after preprocessing	40,692	المصابين
Diacritized	Word with most probable diacritics.	42,413	المُصابَيْنِ
Lemma	The canonical form of the word.	17,784	مُصاب
Stem	The word stem without prefix or suffix.	13,480	مُصاب

Table 1: Feature Vector (FV) lengths for different types of preprocessing

5.3. Identifying Potential Human Rights Abuses as a Classification Problem

Identifying potential Human Rights Abuses (HRA) is treated as a binary classification problem: each tweet is classified as HRA or non-HRA. Tweets are encoded as feature vectors (Salton et al., 1975). Different feature weighting schemes were tested, including Binary, TF, TF-IDF. Lexical and morphological features are extracted from the tweet text, then used to learn different models.

Two classical training methods were used to learn HRA detection using the proposed features. A Naïve Bayes classifier with binary Vector Space Model (VSM) was used as the baseline approach. A Support Vector Machine (SVM) classifier was trained with two different kernels, *linear* and *Gaussian* (Schölkopf and Smola, 2002).[14] SVMs

[14] The Scikit-Learn package (Pedregosa et al., 2011) was used to carry out our experiments.

have traditionally been demonstrated as very effective for text classification tasks. Precision, Recall, and F1 were used as commonly applied evaluation metrics.

More recently, deep learning methods have been shown to be very effective for text classification, e.g. (Miroczuk and Protasiewicz, 2018; Chen et al., 2017). So in addition to Naïve Bayes and SVM, we trained models based on those neural network models that have been shown to be most effective at text classification, namely Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) and Bi-Directional LSTM (biLSTM).

5.4. Creating a Gold Standard Dataset for Training and Testing

The Arabic Violence in Twitter (AVT)[15] dataset is a test collection created as part of the project and used in our experiments (Alhelbawy et al., 2016). AVT is a corpus of violence acts in Arabic Twitter manually annotated using crowdsourcing. It consists of 20,151 tweets covering violent acts such as killing, raping, kidnapping, terrorism, invasion, explosion, or execution, etc.

Five annotators classified every tweet into one of eight classes: Crime, Accident, Human Rights Abuse, Conflict, Crisis, Violence, Opinion, and Other. The 'Human Rights Abuse' category is defined as the tweets that mention an act that may be considered as a human rights violation according to international definitions, such as crimes committed by government, militia, or organisations against civilians. As we are just interested in Human Rights Abuse detection, only the HRA class is used and all other classes are treated as non-HRA. Table 2 shows two examples of tweets that mention violence episodes, one classified as HRA, the other as non-HRA.

Different annotators may assign different classes for the same tweet. The single label for a tweet was therefore determined using as aggregation criterion a class confidence score[16] CS, calculated as shown in Equation 1, where C_i refers to class i, K is the set of all contributors judging a certain tweet, M is the set of contributors assigning a tweet to class C_i, and TS_j is the Trust Score for a contributor j where $0 < j < k$

[15] Downloadable from : https://github.com/Alhelbawy/Arabic-Violence-Twitter

[16] https://success.figure-eight.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score

Tweet Text & Translation	Class
مُعظم الكلمات التي نتحفظ بها في صدورنا تقتل أكثر ، من تلك التي يقرؤها العالم Words that we hold in our hearts kill more, than those the world read.	non-HRA
جيش الأسد في دمشق يرتكب مجزرة مروعة راح ضحيتها عشرات الأطفال داخل مدرستهم فيديو صور The army of Assad committed a terrible massacre in Damascus, claiming the lives of dozens of children in their school video images	HRA

Table 2: Examples of HRA and non-HRA tweets from the AVT dataset.

and $0 < TS_j < 1$.

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

The aggregate class confidence score threshold is set to discard all tweets with low class confidence score. Only tweets with a confidence score above 0.45 are used in our experiments resulting in 16,292 tweets distributed over eight classes.

As we are training a classifier to detect HRA incidents, we used HRA as the main (positive) class, and all other classes, Crime, Accident, Conflict, Crisis, Violence, Opinion, and Other were aggregated into one, non-HRA class. Such setup makes the task more challenging where there are a good number of negative examples (14,424 samples) which have a high level of overlap with the positive examples (1,868 samples). 70% of the dataset is used for training and 30% for testing. Table 3 shows the resulting number of tweets used for training and testing in each class.

Class	**Train**	**Test**	**Total**	**%**
HRA	1,303	565	1,868	11.5
Non-HRA	10,101	4,323	14,424	88.5
Total	11,404	4,888	16,292	

Table 3: AVT Dataset Details

To study data separability, two clustering algorithms were used to cluster the dataset into two clusters. The first is k-means, a hard clustering algorithm (Hartigan and Wong, 1979). A soft clustering algorithm was also used, Latent Dirichlet Allocation (LDA) (Blei et al., 2003). For each instance, the topic assigned the highest probability is used as the instance class. For each of the clustering algorithms, the training data is used to assign each cluster to one class aiming at distinguishing HRA and non-HRA as representing the two clusters. Table 4 shows the results of clustering the test dataset into two clusters using LDA and k-means, respectively. The results shows a high level of overlap between HRA and non-HRA classes. A further evaluation for the clustering results was carried out by calculating homogeneity and completeness of clusters (Rosenberg and Hirschberg, 2007). For LDA, we obtain homogeneity = 0.07 and completeness = 0.04; and 0.0002 and 0.0004, respectively, for k-means. These results can be interpreted as meaning that the data does not naturally split into the classes we aim to model. The main conclusion from these results is that there is a high degree of overlap between HRA and non-HRA tweets.

Clustering	K-means		LDA	
Class	HRA	Non-HRA	HRA	Non-HRA
Cluster1	509	4,167	101	2,410
Cluster2	35	242	443	1,999

Table 4: Dataset separability analysis

5.5. Classification results

The two classical classifiers performed reasonably well at identifying HRA on Twitter. Bag of Words (BoW) was used in our experiments as feature representation. We explored different weighting scheme (Binary, TF, and TF-IDF), but TF-IDF tended to achieve overall better results, so we only report those results in this paper.

Table 5 shows the results at HRA detection using Naïve Bayes and SVM classifiers with different kernels. The baseline Naïve Bayes achieves the highest recall across all tested classifiers, but very low precision. The SVM classifiers achieved good results with both kernels. Our results show that the linear kernel outperformed the Gaussian kernel in terms of recall, but not precision.

Feature	**Naïve Bayes**			**SVM (Linear)**			**SVM (Gaussian)**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Token	25.1	94.3	40.2	65.3	61.2	63.1	85.3	50.9	63.7
Diacritized	38.5	67.2	49.1	49.8	53.1	51.4	76.6	42.9	54.9
Lemma	40.6	52.9	46.2	51.1	38.2	43.6	76.8	27.2	40.0
Stem	44.1	44.1	44.1	62.3	26.1	36.7	81.9	23.3	36.2

Table 5: HRA Classification Results (Precision / Recall / F1), confidence = 0.45, 10-fold cross validation

As discussed in Section 5.2, two sets of features were tested, some resulting in high-dimensional feature vectors, some in low-dimensional ones. Table 1 (in Section 5.2) shows the dimensions of each feature vector. We note that *Token* and *Diacritized* features result in high dimensional vectors ($> 40,000$) while using *Lemmas* or *Stems* reduces this by more than 50%. We also observe that incorporating diacritics does not improve the results over using simple tokens, indicating an increase in non-discriminating features. Furthermore, morphological analysis (i.e., as reflected by *Lemma* and *Stem*) does not appear to boost the performance in either of the SVM settings. A possible explanation can be found when analysing the misclassified samples: most of these are written in Dialectal Arabic (DA).[17] By contrast, available morphological analysers are designed to analyse Modern Standard Arabic (MSA)[18] or the Classical Arabic (CA)[19] so perform best with those varieties of Arabic. Failure to extract morphological features properly is likely to result in improper tweet representation and misclassification.

[17]The term ‘Dialectal Arabic’ is used to indicate the varieties of Arabic spoken in different regions: the Maghreb, Egypt, the Middle East, etc.

[18]Modern Standard Arabic or Fusha is the language of formal writing and speech in Arab countries and it is understandable across Arab countries.

[19]Classical Arabic is the old version of the standard Arabic used in the Quran and in the early Islamic literature from the 7th-9th centuries.

We also explored deep neural networks for the classification task at hand.[20] We applied Convolutional Neural Networks (CNN) in two different varieties, LSTMs, and bidirectional LSTMs, and we conducted the experiments as follows. Let D be a tweet with n tokens, and let t_i be the ith token in tweet D, where each $t_i \in D$ is represented by a k-dimension embedding $v_i \in \mathbb{R}^k$. Tweet document D is converted to a matrix of shape $(30 \times k)$ where every row represents a token vector of length k with k either 100 or 300. The maximal-length token sequence (of tokens in a tweet) is set to 30, and zero-padding is used if the tweet tokens are less than 30. For all models, distributed word embedding representations were presented in the input layer. *Word2Vec* was used to train word embedding vectors with 100 and 300 dimensions using a corpus of collected tweets. Because the number of examples used in training is relatively small given the number of training parameters, overfitting problems were observed. Dropout regularisation was therefore used to prevent the model from overfitting.[21]

Our basic CNN architecture consisted of three convolutional layers, each followed by a *max pooling* layer with *pool size* of 3 and at each layer 64,32,16 *filters* and *kernel size* 5,3,3 respectively. The last *max pooling* layer is fully connected to a *dense* layer of size 256. Two different *dropout* values were tested to avoid overfitting in two different CNN architectures. The first CNN architecture, referred as CNN0.2, applied *dropout* of 0.2 on the output of the first convolution layer. The second architecture, CNN0.5, applied *dropout* of 0.5 after all convolution layers which improves precision but decreases recall. Overall we observe some improvement in terms of F1 score as shown in Table 6. Obvioulsy, there is always a trade-off between precision and recall, but in our application we are mainly focussing on high F1.

Our LSTM model consists of 50 LSTM units and $dropout$ 0.2. The bi-directional LSTM is tested with the same settings where both forward and backward outputs concatenated before being passed on to the next $dense$ layer.

For all our neural network architectures, the final classification is generated by

[20]All experiments were run with Keras and Tensorflow as backend.

[21]Dropout means that a percentage of units are randomly dropped out from the neural network during training to prevent units from co-adapting too much (Srivastava et al., 2014).

		CNN0.5			CNN0.2			LSTM			biLSTM		
Output	dim	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Softmax	100	75.8	59.3	66.5	77.5	58.6	66.7	**84.4**	65.1	73.5	78.5	69.7	73.9
Softmax	300	82.1	55.2	66.0	75.6	59.1	66.3	82.5	69.2	**75.3**	81.1	64.6	71.9
Sigmoid	100	74.7	64.2	69.1	76.7	57.2	65.5	83.5	66.2	73.8	75.7	**71.5**	73.5
Sigmoid	300	80.6	55.0	65.4	73.9	57.5	64.7	81.3	66.4	73.1	78.1	68.7	73.1

Table 6: Deep Neural Network (DNN) Classification Results

either a $sigmoid$ or $softmax$ function and both functions were tested in our experiments. For reproducibility purposes we also report, that Tensorflow and numpy random number seeds are set to 123 before any experiment.

Table 6 shows the results of all deep neural network experiments. The best results were obtained with the LSTM model, using softmax and size 300 for the word embeddings. In general, using softmax activation in the output layer improves the precision. Overall, the CNN, LSTM, and bi-LSTM models using word embeddings substantially improve on the classical approaches, i.e Naïve Bayes and SVM, by almost ten percentage points.

6. Overall Impact of the Platform

The *Ceasefire Iraq* portal was originally tested as an internal deployment. The first report by a partner organisations of *Minority Rights Group* (MRG) was submitted in February 2016, hence the portal has been running for three years now. It opened to the public towards the end of 2016. We run several Facebook advert campaigns starting in April 2017 until September 2017. These were targeted at the geographic region covered by the *Ceasefire Iraq* deployment.

While the portal has become an important tool for analysts within MRG, we also note that it has become a way of monitoring the human rights situation in Iraq to the general public, therefore serving both purposes as outlined in the motivation. As shown in the Ceasefire dashboard (Figure 5), more than 2,500 reports have so far been submitted from different locations in Iraq, distributed over 32 categories of human rights abuse. These incidents are submitted by civilians as well as partner organisations and are shown on the map with details to drill down. Partner organisations are registered with Ceasefire and use the platform to submit their reports accessing and modifying their reports using the security model discussed earlier. The collected reports contributed to a number of publications by human rights organisations, including:

- Eyes on the Ground: Realizing the potential of civilian-led monitoring in armed conflict (Puttick, 2017)
- Broken Lives: Violence against Syrian refugee women and girls in the Kurdistan Region of Iraq (Ceasefire Centre for Civilian Rights and Asuda, 2018)
- A Rising Tide: Monitoring and Documenting Violence against Women in Seven Iraqi Governorates, 2014-2016 (Asuda, 2017)
- Civilian Activists under Threat in Iraq (Ceasefire Center For Civilian Rights and Minority Rights Group International, 2018).

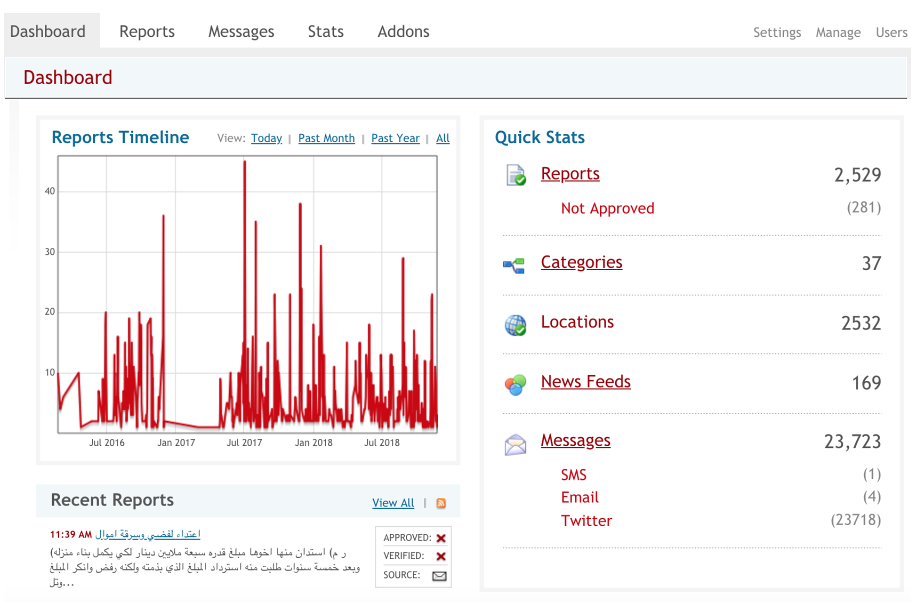

Figure 5: Ceasefire Iraq Dashboard

Ceasefire was also mentioned by the Canadian All-Party Parliamentary Group for the Prevention of Genocide and Other Crimes Against Humanity (GPG) in their report ”Leveraging New Technologies to Prevent and Monitor Genocide and Other Mass Atrocities” as one of their case studies (Canadian All-Party Parliamentary Group for the Prevention of Genocide and Other Crimes Against Humanity, 2018).

In addition to the academic evaluation we also assessed the practical usefulness of the Twitter mining tool for the analysts' work. To do this we carried out a field evaluation. A set of 200 randomly selected tweets identified as HRA by our Twitter monitor was reviewed manually by an expert. The expert confirmed 157 of them as actual reports of an HRA incident. This result, i.e. precision of 78.5%, is very close to the experimental results we obtained by evaluating our classifier on a test set. This was deemed of high enough quality to be used in the practical setting.

7. Conclusions and Future Work

In this paper we presented Ceasefire, a platform that supports grassroots-based human rights monitoring in addition to assisting human rights organisations in their work. The platform also serves as an information portal to the general public providing an insight into human rights violations and abuses within a specific geographical region. Ceasefire has been active for three years; during this period, it has proven that grassroots based monitoring is a viable alternative to the riskier strategy normally adopted by human rights organisations. Our improved security and structural organisation model incorporated in an existing open-source reporting framework helped us to convince a number of organisations to collaborate in the portal using a unified framework. In addition to manually submitted reports, NLP technology has been exploited to identify potential human rights abuse incidents from social media with an accuracy of about 80%, which is promising given the motivation to employ this technology to tap into the many signals obtained from social media by the many victims of such incidents that might not trust human rights organisations or are not aware of the existence of portals such as Ceasefire. Among the technical contributions, this work is to the best of our knowledge the first attempt to use NLP technology for human rights abuse identification from social media. Our work also suggests that deep neural network models such as LSTMs and bi-LSTMs outperform conventional text classification approaches such as SVMs which is in line with findings in other NLP areas.

The success of the Iraq use case has motivated the participating organisations to get involved in an effort to use this technology to develop new platforms to support monitoring in more countries, and we are currenty in the process of rolling the platform out to the broader Middle East and North Africa (MENA) region. The latest deployment is *Ceasefire Iran*[22] which is available in Persian and English.

In future research, we plan to continue developing more advanced models to improve the precision of HRA identification.

Acknowledgement

The authors would like to thank Miriam Puttick for her support and manual review of test cases. Also, we want to thank Minority Rights Group International and Innovate UK for funding part of this research work through a Knowledge Transfer Partnership (KTP) project between MRG and the University of Essex. The research was also in part supported by the UK Economic and Social Research Council (ESRC) through the Big Data Human Rights and Technology project (grant number ES/M010236/1).

[22] `https://www.ceasefire.org/iran/`

References

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

Al-Sughaiyer, I. A. and Al-Kharashi, I. A. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

Aldayel, H. K. and Azmi, A. M. (2016). Arabic tweets sentiment analysis–a hybrid scheme. *Journal of Information Science*, 42(6):782–797.

Alhelbawy, A., Poesio, M., and Kruschwitz, U. (2016). Towards a Corpus of Violence Acts in Arabic Social Media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1627–1631. European Language Resources Association (ELRA).

Alston, P. and Crawford, J., editors (2000). *The future of UN human rights treaty monitoring*. Cambridge University Press.

Althobaiti, M., Kruschwitz, U., and Poesio, M. (2014). AraNLP: a Java-Based Library for the Processing of Arabic Text. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik.

Asuda (2017). A Rising Tide: Monitoring and Documenting Violence against Women in Seven Iraqi Governorates.

Awad, D., Sabty, C., Elmahdy, M., and Abdennadher, S. (2018). Arabic Name Entity Recognition Using Deep Learning. In *International Conference on Statistical Language and Speech Processing*, pages 105–116. Springer.

Bailard, C. S. and Livingston, S. (2014). Crowdsourcing accountability in a nigerian election. *Journal of Information Technology & Politics*, 11(4):349–367.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Canadian All-Party Parliamentary Group for the Prevention of Genocide and Other Crimes Against Humanity (2018). Leveraging New Technologies to Prevent and Monitor Genocide and Other Mass Atrocities. Online Report.

Cano Basave, A. E., He, Y., Liu, K., and Zhao, J. (2013). A weakly supervised Bayesian model for violence detection in social media. In *Sixth International Joint Conference on Natural Language Processing: Proceedings of the Main Conference*, pages 109–117. Asian Federation of Natural Language Processing.

Ceasefire Center For Civilian Rights and Minority Rights Group International (2018). Civilian Activists under Threat in Iraq. Online Report.

Ceasefire Centre for Civilian Rights and Asuda (2018). Broken Lives: Violence against Syrian refugee women and girls in the Kurdistan Region of Iraq. Online Report.

Chandrasekharan, E., Samory, M., Srinivasan, A., and Gilbert, E. (2017). The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3175–3187, New York, NY, USA. ACM.

Chen, T., Xu, R., He, Y., and Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, 72:221 – 230.

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

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.

Datta, A., Shah, M., and Lobo, N. D. V. (2002). Person-on-person violence detection in video data. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 433–438. IEEE.

Diab, M., Habash, N., and Zitouni, I. (2018). NLP for Arabic and Related Languages. *Revue Traitement Automatique des Langues*, 58(3):9–13.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Heinzelman, J. and Meier, P. (2015). Crowdsourcing for human rights monitoring: Challenges and opportunities for information collection and verification. In *Human Rights and Ethics: Concepts, Methodologies, Tools, and Applications*, pages 409–424. IGI Global.

Hochschild, A. (1999). *King Leopold's ghost: A story of greed, terror, and heroism in colonial Africa*. Houghton Mifflin Harcourt.

Hochschild, A. (2005). *Bury the chains: the British struggle to abolish slavery*. Pan Macmillan.

Kontostathis, A., Edwards, L., and Leatherman, A. (2010). Text mining and cybercrime. *Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK*.

Lattimer, M. and Sands, P. (2018). *The Grey Zone: Civilian Protection Between Human Rights and the Laws of War*. Bloomsbury Publishing.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute.

Miroczuk, M. M. and Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106:36 – 54.

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.

Müller, V. C. and Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence*, pages 555–572. Springer.

Nievas, E. B., Suarez, O. D., García, G. B., and Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns*, pages 332–339. Springer.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *LREC*, volume 14, pages 1094–1101.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Puttick, M. (2017). Eyes on the Ground: Realizing the potential of civilian-led monitoring in armed conflict. Technical report, Ceasefire Centre for Civilian Rights, London.

Raad Al Hussein, Z. (2016). Opening Statement by Zeid Raad Al Hussein, United Nations High Commissioner for Human Rights. *33rd Session of the UN Human Rights Council*, 3.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops*

(ICMLA), 2011 10th International Conference on, volume 2, pages 241–244. IEEE.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson Education Limited.

Salem, F. (2017). The Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World. *Dubai: MBR School of Government.*, 7.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

The United Nations (1948). Universal declaration of human rights.

Whittaker, E. and Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of School Violence*, 14(1):11–29.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7.

Zirikly, A. and Diab, M. (2015). Named entity recognition for Arabic social media. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 176–185.