Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 142 (2018) 174–181

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Conference on Arabic Computational Linguistics (ACLing 2018),
November 17-19 2018, Dubai, United Arab Emirates

Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic

Azalden Alakrot[a], Liam Murray[b], Nikola S. Nikolov[a]

[a]*Department of Computer Science and Information Systems, University of Limerick, Ireland*
[b]*School of Languages, University of Limerick, Ireland*

Abstract

Warning: this paper contains a range of words which may cause offence.
In recent years, many studies target anti-social behaviour such as offensive language and cyberbullying in online communication. Typically, these studies collect data from various reachable sources, the majority of the datasets being in English. However, to the best of our knowledge, there is no dataset collected from the YouTube platform targeting Arabic text and overall there are only a few datasets of Arabic text, collected from other social platforms for the purpose of offensive language detection. Therefore, in this paper we contribute to this field by presenting a dataset of YouTube comments in Arabic, specifically designed to be used for the detection of offensive language in a machine learning scenario. Our dataset contains a range of offensive language and flaming in the form of YouTube comments. We document the labelling process we have conducted, taking into account the difference in the Arab dialects and the diversity of perception of offensive language throughout the Arab world. Furthermore, statistical analysis of the dataset is presented, in order to make it ready for use as a training dataset for predictive modelling.

Keywords: anti-social behaviour; offensive language; harassment detection; Arabic dataset; Arabic dialects; text mining; text classification;

1. Introduction

Although the number of offensive language detection studies has increased in recent years, there are not many datasets specifically labelled for tackling this problem. Currently, and to the best of our knowledge, there are not many datasets publicly available to allow targeting the same issue in Arabic text. We found a few recent studies, one

* Corresponding author.
E-mail addresses: Azalden.Alakrot@gmail.com (Azalden Alakrot)., Liam.Murray@ul.ie (Liam Murray)., Nikola.Nikolov@ul.ie (Nikola S. Nikolov).

10.1016/j.procs.2018.10.473

of which makes two datasets available, a dataset of 1,100 manually labelled tweets as well as a dataset of 32K user comments from a popular Arabic news site, both containing data entries deemed to be inappropriate language [12]. Another study applies manual labelling of 500 Twitter accounts, with half of these 500 accounts labelled as abusive [1]. In general, the labelled datasets in these studies are relatively small. In addition, these studies predominantly use data collected from Twitter (the maximum length of Twitter posts is 140 characters), while the length of the comments on other social media platforms, such as YouTube, can be irregular in terms of number of words (e.g., on YouTube the number of words per post can exceed thousands). Therefore, an initial goal of work has been to construct a suitable corpus, different and richer than the few ones available in the research literature, which can potentially improve further the research results in detection of offensive language in online communication in Arabic.

In the design of corpora, there are essential characteristics that need to be considered such as *availability*, *representativeness*, *heterogeneity* and *balance* [14] with availability and representativeness being two crucial factors in studying offensive language detection. There are many incidents of offensive language occurring in private environments on the Internet where access is restricted, such as Facebook, which otherwise would be a good source for such data. However, there are also other sources publicly available, such as YouTube, and incidents of abuse and offensive language happen regularly on these platforms as well. Furthermore, on public platforms, victims are humiliated in front of a larger segment of people, and more people take part in the abuse compared to platforms with a higher level of privacy. Thus, such platforms are a rich source of data which is both publicly available and representative.

YouTube is a popular platform for sharing videos, which provides many activities for its users. It allows users to comment on shared videos, and these comments occasionally contain offensive language and insults. YouTube has been of special interest in research on flaming and antagonism [16], with flaming defined as posting negative comments online [10]. A study by Moor *et al.* states that hostility by insulting, swearing or using otherwise offensive language appears to be extremely common on YouTube [11]. The work of Lange presents a potential interpretation of the widespread of flaming on YouTube. It suggests that plenty of people assume that *haters* are users who do not publish videos themselves. That is, there is a category of YouTube users who tend to post comments, typically having little to do with the video they are commenting on, whilst having never to risk receiving any unpleasant criticism themselves. This opinion suggests the presence of a crowd of the YouTube users who simply enjoy offending others [10].

The minimum number of positives (i.e., profane comments) required depends on the employed data mining methodology [6]. Ideally, a dataset of this kind should represent the diversity of text present in cyberspace and also generated by a variety of people. By text diversity we mean the variety of writing styles, where the style speaks of the personal intentions of the author. It is sensible to assume that the larger the number of diverse profane comments is present in a training dataset, the more accurate offensive language and harassment detection can be made by employing the dataset for predictive modelling. Furthermore, we aimed at a balanced number of positive and negative labelled comments in the dataset, which can help for minimising the false negatives in a predictive analytics scenario.

Another challenge in constructing a dataset of this kind is the data labelling process owing to the manual labour required for it. The labelling process can vary depending on the purpose of the study. Therefore, the definition and the specific instructions for labelling should be adhered to at all times during the process.

In this paper, we present the dataset that we have collected for our experiments. We also discuss the methods of collecting and labelling the dataset, its structure as well as its suitability for offensive language and harassment detection in Arabic text.

2. Dataset Collection

According to Nalini and Sheela [13], data collection for the study of cybercrime needs to focus mainly on selecting appropriate platforms to avoid both legal and technical issues. The YouTube platform has nearly two billion users [20] and the Internet Live Stats website[1] reports that there are 70,122 YouTube videos viewed in one second. YouTube does not prevent users from publishing offensive content, and in the case when such contents is published, it takes

[1] http://www.internetlivestats.com/statistics/ [Online; accessed Jun 2018]

time to have it removed. Therefore, comments very likely contain a variety of speeches ranging from compliments to pejoratives, such comments are often available, which makes them a good source of data pertaining to cyber insulting.

Among various social media platforms, YouTube is the second-biggest social media platform with 1.8 billion Internet users, after *Facebook* with 2.2 billion as of March 2018 [7]. YouTube is localised in 88 countries and can be accessed in 76 languages [20]. It has a broad scope of users, from different age and gender groups [20]. This diversity in types of users makes the material published on YouTube represent a wide range of societal attitudes and thus is appropriate as a source for investigating the interaction between people. Similar to other social media platforms, YouTube is a place for communication between people without limits, owing to the anonymity that is allowed, which opens the door to users to speak without restrictions and *misbehave* in their interaction with others. It is common to curse and offend others, and such incidents are increasing [17, 8]. Many users, videos, and comments create a suitable environment for people to disturb and insult others through posting offensive comments in cyberspace. Table 1 contains some instances of offensive language in YouTube comments.

Table 1. Examples of instances of offensive language in YouTube comments in Arabic.

Translation	**Comment**
A whore in every sense of the word	عاهرة بكل معنى الكلمة
He is mentally sick	هذا مريض نفسي
God’s curses on the one whose face looks like a monkey’s face	احلام خرا لعنة الله على وجه اللي مثل وجه القرد
She’s a fallen woman!	هي ساقطه اشتتوقع منه
She’s a failure and artistically dead	فاشلة وميتة فنيا من زمن

The comments in Table 1 suggest the presence of harassing. Thus, they are a proper source for the dataset required for our research.

3. Sampling

For the purpose of our study, we choose to select videos based on the YouTube channel they are posted in. It can be noticed that some of the channels that are keen to increase subscriber number are posting controversial videos about celebrities. This kind of videos attracts people who like to comment on rumours and they occasionally use insults in their comments. We picked videos with the highest number of comments from the selected channels, expecting that longer discussions would contain a high number of comments with offensive language, thus helping us in increasing the number of positives. As the main target of our study is to detect comments containing offensive language, it is important to increase the number of positives (i.e. abusive comments) in the dataset and achieve balance between positives and negatives. The dataset would be imbalanced if the class of interest contains a small number of training instances (also named minority or positive class), while the rest of the most instances is the second class (also named majority or negative class) [3].

The comments collected are written by casual spoken language. Usually, in this type of online communication different people have different writing style; moreover, there are frequent changes in the manner of how people communicate, in terms of their writing. These changes and the total lack of any structural rule make the processing of this kind of content a great challenge.

The comments in our dataset were collected in July 2017, and the upload dates for these videos range from 2015 to 2017. The dataset contains the following 14 attributes: CommentID, Username, Date, Timestamp, CommentText, Likes, HasReplies, NumberOfReplies, Replies.id, Replies.user, Replies.date, Replies.timestamp, Replies.likes, Replies.commentText. In the next section we present an initial analysis of the dataset.

4. Descriptive Analysis of the Dataset

Our corpus consists of 167,549 YouTube comments posted by 84,354 users along with 87,388 replies posted by 24,039 users from 150 YouTube videos. These videos present controversial media footage of celebrities. This kind of footages provoke viewers, leading some of them to use offensive language in their comments. As we emphasised in the introduction, representativeness is an important factor. Thus, for learning more about what our dataset represents we conducted an analysis to identify whether or not the comments are written or read by people from a wide range of Arab nations. We consider the presence of people from different Arab countries in a conversation as a sign that these insults are understandable by the majority of them.

4.1. Word Frequency

As a first step, we computed the word frequency for all terms appearing in our dataset. A list of a total number of 250,382 unique words were obtained from the calculation of words frequency in the whole dataset. Then we sorted the list based on word frequency from the largest to smallest. We manually searched the first 30,000 words in the list for names of countries and nationalities and recorded their frequency. The choice of the manual search is due to misspellings that can be a reason to miss many words. The two histograms in Figures 1 and 2 illustrate the nationalities and countries, respectively. Figure 1 shows the frequency of nationalities mentioned in the first 30,000 words and Figure 2 shows the frequency of countries' names mentioned in the first 30,000 words. To give an example, Table 2 presents some comments taken randomly from the dataset and referring to the nationality of some people. These examples of comments reflect the diversity of nationalities included in our dataset and suggests that people from the majority of the Arab region understand the insults in these comments.

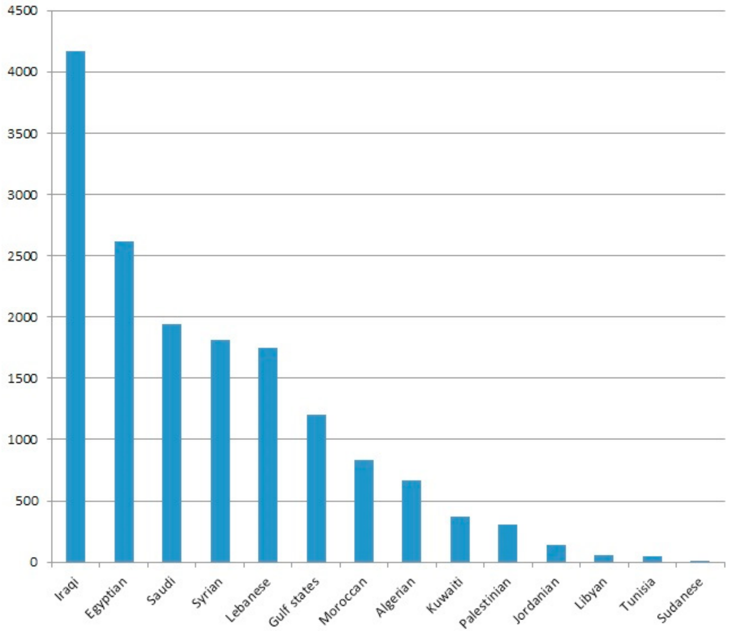

Fig. 1. Frequency of nationalities mentioned in the first 30,000 words.

4.2. Further Discovery in the Dataset

We also discovered the use of profane words from languages other than Arabic in the collected comments. These occur in the form of a single word, a phrase or whole sentences in another language. Foreign words transcribed with the Arabic alphabet and Arabic words transcribed with a non-Arabic alphabet are also present; moreover, some sentences mix languages. We discovered 475 such comments, examples of which are presented in Table 3.

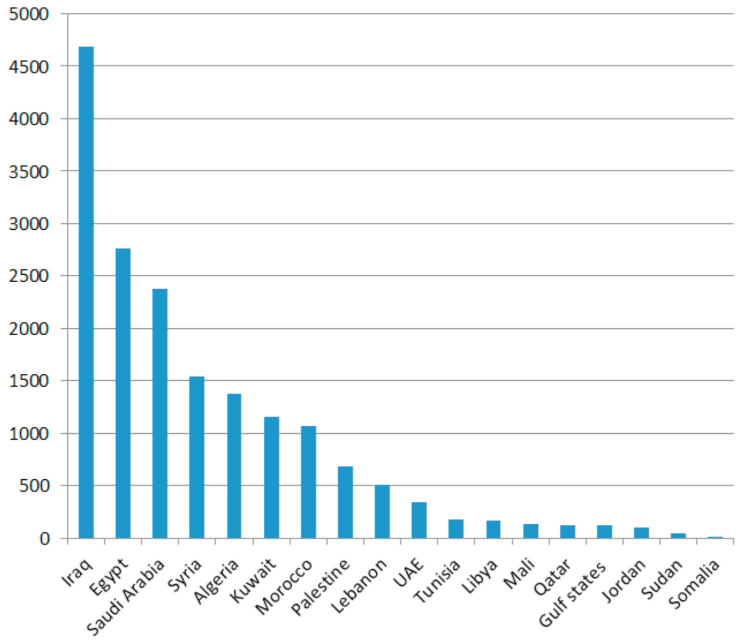

Fig. 2. Frequency of countries’ names mentioned in the first 30,000 words.

Table 2. Examples of comments made by people from different Arab regions.

Comments and Translations
بصراحه يا سما انتي شخصيه لزيزه جدا وباين عليك انك طيبه وقلبك ابيض ،اتمنى لك ان يهديك الله لما يحب ويرضى انا سودانيه Frankly, Sama, you are a very kind person, you have a good heart, I wish God guides you to what he loves, I am Sudanese.
والله العظيم احسن فنانة واحترم فنانة شفتها في حياتي تحياتي ليك حبيبتي من ليبيا I swore, you are the best and respected singer I have ever seen in my life. Greetings and my love from Libya.
بعشق العراق واهل العراق والله مصراوية I adore Iraq and the people of Iraq, I swear; I am an Egyptian.
جزائري احلام دي قلبها طيب والله العظيم لكن لا تحسب لكلامها او لتصرفه اي ان عفويتها قد تؤدي بها الى متاهات Algerian, Ahlam has a good heart and I swear by God that the problem is that she doesn’t take into account what she says or how she behaves, which means that her impulsiveness could get her into a big mess.
منو كل لكم هذ عراقي اصلا سمعت لهجته!! لهجتك لبنانية مدري سورية Who told you he’s Iraqi! I heard him, his dialect is either Lebanese or maybe Syrian.
انتي سوريه مو You’re Syrian, aren’t you?
هيه مو كويتيه ليش متحجي كويتي....مصري ع لبناني ع سوري She’s Kuwaiti, isn’t she, why she doesn’t speak Kuwaiti. . . Mix of Egyptian, Lebanese and Syrian dialects.

Table 3. Examples of non-Arabic comments and Arabic comments written in non-Arabic alphabet.

Comment	**Language**
Fu** off	English
who the phu** is she ????? pffff	English (misspelled)
vous łtes trs trs mauvaise je sais pas pourquoi comporter comme a avec ahlam moi je l'aime bien c'est une femme vivante c'est trs mchant tous les gens qui sont jaloux d'elle il faut la laisser tranquille ahlam on t'aime les Marocains	French
poquito enferma	Spanish
inti asln min ma ma7loki mn li3rab chali fomk bsman 9bal ma thadri 3la lmgharba fhamti ya nakira	Arabic in Latin alphabet transcription
Tfu mnin awdi nti mlmgrib tfuu	Arabic in Latin alphabet transcription
لا يا بتش تعرفي انقليزي دى فنانة بتاع بورنو دي بتعمل اغاني سكس مش فديوكلب	English mixed with Arabic, both in Arabic alphabet transcription

5. Annotation

Previous related studies employ a variety of strategies for labelling datasets. For example, Warner and Hirschberg manually label user comments and a corpus of websites [19]. Huang *et al.* choose to label about 13,000 messages to be positive or negative for cyberbullying detection [5]. They asked three students to perform the job, and comments with disagreement in labelling were rejected. Dadvar proposes to ask three students to label posts to be either *yes* or *no* in terms of bullying [4]. Posts on which at least two students agree are marked as positive, i.e. bullying. Reynolds *et al.* employ a dataset which includes 2,696 posts labelled with the use of Amazons Mechanical Turk service[2] [18]. Kontostathis *et al.* hired three workers also on Amazons Mechanical Turk to label their dataset [9]. Al-garadi *et al.* [2] report that the dataset labelling is done with the assistance of three people as well. We have followed the same labelling process by employing three annotators. Details about our annotators and labelling process are stated next.

Out of the whole dataset, we picked nine videos with offensive comments which also have a relatively high total number of comments, assuming that the longer the conversation is the higher amount of offensive content it contains. These nine videos contain nearly 16,000 comments with the number of words in each comment ranging from 1 to 2,338 with average of 75 words. The labelling was performed on this sample out of the whole data collection. We assigned the labelling task to three annotators from three different nationalities; one is Iraqi and the second is Egyptian, i.e. from the two nationalities most highly represented in the dataset as shown in Figures 1 and 2; moreover, they are from high-density urban areas. The third person is from Libya, a country and nationality with low representation in the dataset, and also from low-density urban area. The ages of three annotators are 44, 34 and 32, respectively. Two of them finished their third level education, one in information technologies, the other one in accounting and the third one quit university in his second year.

We asked the annotators to label offensive comments as positive and inoffensive comments as negative and leave unlabelled any comment they are not sure about. The three annotators agreed on 10,715 comments, the inter-annotator agreement is 71% of the whole sample. The number of comments with at least one disagreement is 4,335, and the number of unlabelled comments by at least one person is 848. We excluded these 848 comments. A summary of these numbers is presented in Table 4. In addition to the attributes mentioned in the last paragraph at section 3, four more attributes have been added, three of which represent the opinion of the three annotators, and the fourth attribute is the final decision about the comment, whether it is offensive or not, based on the agreement between the three annotators.

[2] https://www.mturk.com/

Table 4. Number of agreements and disagreements between annotators in the labelled dataset.

Comments on which all annotators agree	**Inter-annotator agreement**	**At least one annotator disagree**	**Comments unlabelled by at least one**
10715 comments	71%	4335 comments	848 comments

For accomplishing the construction of the dataset we adopt the two following scenarios:
Scenario 1: Label as offensive the comments on which all annotators agree, and label as inoffensive the rest.
Scenario 2: Label as offensive the comments on which at least two annotators agree, and label as inoffensive the rest.

In the first scenario the number of comments labelled as positives and negatives are 3,532 and 11,518, respectively, i.e. 23% positives. In the second scenario the number of comments labelled as positives and negatives are 5,817 and 9,233, respectively, i.e. 39% positives, as shown in Table 5. Table 5 summarises the number of positives in both scenarios for the whole sample. Moreover, the Inter-annotator agreements are calculated between each pair of annotators using kappa statistics and presented in Table 6.

Table 5. Number of positives in the two scenarios.

Scenario 1		**Scenario 2**	
labelled offensive by three annotators	Percentage of positives	labelled offensive by two annotators	Percentage of positives
3532 comments	23%	5817 comments	39%

Table 6. Inter-annotator agreement using kappa statistics.

Inter-annotator agreement between	**Kappa statistics**
Egyptian and Libyan	0.698
Iraqi and Libyan	0.579
Egyptian and Iraqi	0.512

We made our dataset publicly available at https://goo.gl/27EVbU.

6. Limitations

There are two notable limitations associated with this dataset. The first one is the use of a nine-video sample. Because of this, the results cannot be generalised to the entire Arab population. However, this sample introduces a large segment of Arab social media users from the Arab East, which has the highest percentage of Internet users in the Arab world[3]. Therefore, we believe that there is a strong confidence that the randomisation process along with the choice of highly popular videos related to celebrities known throughout the entire Arab world as well as the relatively high number of comments collected minimise the potential effect of limiting the data collection to nine videos.

Another limitation arises from the choice of annotators to label the dataset. Two of them are from nations highly represented among the authors of comments in this sample, and the third annotator is from a nation with a relatively lower representation in the dataset. We made this choice in order to ensure that comments labelled as offensive would be considered such throughout the entire Arab region. The inter-annotator agreement between the three annotators is 71%. This percentage is very reasonable, especially when we take into consideration another factor which is that different people have different perspectives on the same comment in terms of its offensiveness. As it has been pointed out, the responses of participants towards potentially offensive language varies depending on the context [15]. The annotators' views on each comments were also influenced by their age, gender and personal experiences. At the same time, we want to point out that the results of this study shows moderate to high inter-annotator agreement with

[3] https://www.itu.int/en/ITU-D/Statistics/Pages/stat/ [Online; accessed Jun 2018]

the kappa statistic between each pair of annotators being at an acceptable level (see Table 6), thus we believe we minimised the effect of the limitation associated with the choice of annotators.

7. Conclusion

In this paper we introduce a dataset of YouTube comments in Arabic together with a statistical analysis of it. We collected the data according to the principles of *availability*, *representativeness*, *heterogeneity* and *balance*, thus ensuring that it can be applied for training predictive analytics models for detection of abusive language in online communication in Arabic. Along with conversations scripted in Arabic, this dataset also includes foreign words transcribed with the Arabic alphabet as well as Arabic words transcribed with a non-Arabic alphabet. To the best of our knowledge, this is the first dataset of YouTube comments in Arabic of this kind.

In terms of representativeness and heterogeneity, the dataset portrays a real case of offensive language comments in Arabic by a wide variety of Arab YouTube users from various nationalities. In terms of balance, our dataset contains 39% positives based on the agreement of two out of three annotators from three different Arab nationalities. It is interesting to note that our annotators agree on their evaluation of the language as either offensive or inoffensive in 71% of the cases. This is something we anticipated, and we find that it confirms the representativeness and heterogeneity that we were seeking.

We conclude that this dataset is appropriate for employment as a training dataset in the context of machine learning. Future work will explore a variety of pre-processing techniques and machine learning algorithms for building accurate models for detecting offensive content in online communication in Arabic.

References

[1] Abozinadah, E.A., Mbaziira, A.V., Jones, J., 2015. Detection of abusive accounts with Arabic tweets. Int. J. Knowl. Eng.-IACSIT 1, 113–119.

[2] Al-garadi, M.A., Varathan, K.D., Ravana, S.D., 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior 63, 433–443.

[3] Ali, A., Shamsuddin, S.M., Ralescu, A.L., 2015. Classification with class imbalance problem: a review. Int. J. Advance Soft Compu. Appl 7.

[4] Dadvar, M., 2014. Experts and machines united against cyberbullying. University of Twente.

[5] Huang, Q., Singh, V.K., Atrey, P.K., 2014. Cyber bullying detection using social and textual analysis, in: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, ACM. pp. 3–6.

[6] Indurkhya, N., Damerau, F.J., 2010. Handbook of natural language processing. 2nd ed., CRC Press.

[7] Kallas, P., 2018. Top 15 Most Popular Social Networking Sites and Apps [Jun 2018]. URL: https://www.dreamgrow.com/. [Online; accessed Jun 2018].

[8] Kawate, S., Patil, K., 2017. Analysis of foul language usage in social media text conversation. International Journal of Social Media and Interactive Learning Environments 5, 227–251.

[9] Kontostathis, A., Reynolds, K., Garron, A., Edwards, L., 2013. Detecting cyberbullying: query terms and techniques, in: Proceedings of the 5th annual ACM Web science conference, ACM. pp. 195–204.

[10] Lange, P.G., 2007. Commenting on comments: Investigating responses to antagonism on YouTube, in: Society for Applied anthropology conference, pp. 163–190.

[11] Moor, P.J., Heuvelman, A., Verleur, R., 2010. Flaming on Youtube. Computers in Human Behavior 26, 1536–1546.

[12] Mubarak, H., Darwish, K., Magdy, W., 2017. Abusive language detection on Arabic social media, in: Proceedings of the First Workshop on Abusive Language Online, pp. 52–56.

[13] Nalini, K., Sheela, L.J., 2014. A survey on datamining in cyber bullying. International Journal on Recent and Innovation Trends in Computing and Communication 2, 1865–1869.

[14] Nguyen, D., Demeester, T., Trieschnigg, D., Hiemstra, D., 2012. Federated search in the wild: the combined power of over a hundred search engines, in: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM. pp. 1874–1878.

[15] Ofcom, 2015. Attitudes to potentially offensive language and gestures on TV and radio. URL: https://www.ofcom.org.uk//. [Online; accessed Jun 2018].

[16] Pihlaja, S., 2014. Antagonism on YouTube: Metaphor in online discourse. Bloomsbury Publishing.

[17] Protalinski, E., 2011. 47% of Facebook walls contain profanity. URL: https://www.zdnet.com/article/47-of-facebook-walls-contain-profanity/. [Online; accessed Jun 2018].

[18] Reynolds, K., Kontostathis, A., Edwards, L., 2011. Using machine learning to detect cyberbullying, in: Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on, IEEE. pp. 241–244.

[19] Warner, W., Hirschberg, J., 2012. Detecting hate speech on the world wide web, in: Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics. pp. 19–26.

[20] YouTube, 2005. YouTube press statistics. URL: https://www.youtube.com/yt/about/press/. [Online; accessed August 2018].