A Hierarchical Annotation of Offensive Posts in Social Media: The Offensive Language Identification Dataset

Marcos Zampieri[1], Shervin Malmasi[2], Preslav Nakov[3], Sara Rosenthal[4]
Noura Farra[5], Ritesh Kumar[6]
[1]University of Wolverhampton, UK, [2]Harvard Medical School, USA
[3]Qatar Computing Research Institute, HBKU, Qatar, [4]IBM Research, USA
[5]Columbia University, USA, [6]Bhim Rao Ambedkar University, India
m.zampieri@wlv.ac.uk

Abstract

Offensive content has become pervasive in social media, drawing attention to research in the application of computational methods for identifying potentially offensive messages. Previous work in this area did not consider the problem as a whole, but rather focused on detecting specific types of offensive content, e.g. hate speech, profanity, cyberbulling, or cyber-aggression. In this paper we consider several different kinds of offensive content using a single annotation scheme. In particular, we propose to model the task hierarchically, identifying the type and the target of offensive messages in social media. We present OLID, the new Offensive Language Identification Dataset compiled specifically for this purpose. The dataset contains tweets annotated with a fine-grained three-layer annotation scheme and will be made publicly available to the research community.

1 Introduction

Offensive content has become pervasive in social media and a reason of concern for government organizations, online communities, and social media platforms. One of the most common strategies to tackle the problem is to train systems capable of recognizing offensive content, which then can be deleted or set aside for human moderation.

In the last few years, there have been several studies published on the application of computational methods to deal with this problem. Most prior work focuses on a different aspect of offensive language such as abusive language (Nobata et al., 2016; Mubarak et al., 2017), (cyber-)aggression (Kumar et al., 2018), (cyber-)bullying (Xu et al., 2012; Dadvar et al., 2013), toxic comments[1], hate speech (Kwok and Wang, 2013; Djuric et al., 2015; Burnap and Williams, 2015; Davidson et al., 2017; Malmasi and Zampieri, 2018), and offensive language (Wiegand et al., 2018). Prior work has focused on these aspects of offensive language in Twitter (Xu et al., 2012; Burnap and Williams, 2015; Davidson et al., 2017; Wiegand et al., 2018), Wikipedia comments[1], and Facebook posts (Kumar et al., 2018).

Recently, Waseem et. al. (2017) acknowledged the similarities among previous work and discussed the need for a typology that differentiates between whether the (abusive) language is directed towards a specific individual or entity or towards a generalized group and whether the abusive content is explicit or implicit. Wiegand et al. (2018) followed this trend as well on German tweets. In their evaluation, they have a task to detect offensive vs not offensive tweets and a second task for distinguishing between the offensive tweets as profanity, insult, or abuse. However, no prior work has explored the target of the offensive language, which is important in many scenarios, e.g., when studying hate speech with respect to a specific target. Therefore, we expand on these ideas by proposing a a hierarchical three-level annotation model that encompasses:

A: Offensive Language Detection
B: Categorization of Offensive Language
C: Offensive Language Target Identification

We use this annotation model to create a new large publicly available dataset of English tweets.[2] We provided this dataset to participants of the SemEval 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)[3] shared task (Zampieri et al., 2019). In OffensEval, each annotation level is an independent sub-task.

[1] https://bit.ly/2FhLMVz

[2] Downloadable at https://scholar.harvard.edu/malmasi/olid

[3] https://competitions.codalab.org/competitions/20011

2 Related Work

Different abusive and offense language identification sub-tasks have been explored in the past few years including aggression identification, bullying detection, hate speech, toxic comments, and offensive language.

Aggression identification: The TRAC shared task on Aggression Identification (Kumar et al., 2018) provided participants with a dataset containing 15,000 annotated Facebook posts and comments in English and Hindi for training and validation. For testing, two different sets, one from Facebook and one from Twitter were provided. Systems were trained to discriminate between three classes: non-aggressive, covertly aggressive, and overtly aggressive. 130 teams signed up for this shared task evidencing the interest of the community in the topic. The best performing systems in this competition used deep learning approaches based on convolutional neural networks (CNN), recurrent neural networks (biLSTM), and LSTMs (Aroyehun and Gelbukh, 2018; Majumder et al., 2018). The results of the TRAC competition motivated us to experiment with the deep learning approaches presented in this paper.

Bullying detection: Several studies have been published on bullying detection. One of them is the one by Xu et al. (2012) which apply sentiment analysis to detect bullying in tweets. Xu et al. (2012) use topic models to to identify relevant topics in bullying. Another related study is the one by Dadvar et al. (2013) which use user-related features such as the frequency of profanity in previous messages to improve bullying detection.

Hate speech identification: Hate speech identification is perhaps the most widespread abusive language detection sub-task. There have been several studies published on this sub-task such as Kwok and Wang (2013) and Djuric et al. (2015) who build a binary classifier to distinguish between 'clean' comments and comments containing hate speech and profanity. It has also been modeled as a binary classification task (Burnap and Williams, 2015) using cyber hate in Twitter data. More recently, Davidson et al. Davidson et al. (2017) presented the hate speech detection dataset containing over 24,000 English tweets labeled as non offensive, hate speech, and profanity. The dataset has been used in several studies such as Malmasi and Zampieri (2018) which discusses the role of profanity in hate speech detection.

Offensive language: The GermEval[4] (Wiegand et al., 2018) shared task focused on Offensive language identification in German tweets. A dataset of over 8,500 annotated tweets was provided to participants who could choose to participate in two sub-tasks. Sub-task 1 was a course-grained binary classification task in which systems were trained to discriminate between offensive and non-offensive tweets. In sub-task 2 the organizers broke down the offensive class into three classes: profanity, insult, and abuse. In addition to using English tweets, our task differs from this because we have three levels in our hierarchy as well as different labels in level B.

Toxic comments: The Toxic Comment Classification Challenge was an open competition at Kaggle which provided participants with comments from Wikipedia labeled in six classes: toxic, severe toxic, obscene, threat, insult, identity hate. The dataset continues to be used outside the competition (Georgakopoulos et al., 2018) and it has been used by one of the participants of the aforementioned TRAC shared task as additional training material (Fortuna et al., 2018).

3 Modelling Offensive Content Hierarchically

In the OLID dataset, we use a hierarchical annotation model split into three levels to distinguish between whether language is offensive or not (A), and type (B) and target (C) of the offensive post.

Each level is described in more detail in the following subsections and examples are shown in Table 1.

3.1 Level A: Offensive language Detection

Level A discriminates between offensive (OFF) and non-offensive (NOT) tweets.

Not Offensive (NOT): Posts that do not contain offense or profanity;
Offensive (OFF): We label a post as offensive if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This category includes insults, threats, and posts containing profane language or swear words.

[4] https://projects.fzai.h-da.de/iggsa/

Tweet	A	B	C
@USER He is so generous with his offers.	NOT	—	—
IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	—
@USER Fuk this fat cock sucker	OFF	TIN	IND
@USER Figures! What is wrong with these idiots? Thank God for @USER	OFF	TIN	GRP

Table 1: Several tweets from the dataset, with their labels for each level of the annotation model.

3.2 Level B: Categorization of Offensive Language

Level B categorizes the type of offense and two labels are used: targeted (TIN) and untargeted (INT) insults and threats.

Targeted Insult (TIN): Posts which contain an insult/threat to an individual, group, or others (see next layer);
Untargeted (UNT): Posts containing non-targeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

3.3 Level C: Offensive Language Target Identification

Level C categorizes the targets of insults and threats as individual (IND), group (GRP), and other (OTH).

Individual (IND): Posts targeting an individual. It can be a a famous person, a named individual or an unnamed participant in the conversation. Insults and threats targeted at individuals are often defined as cyberbulling.
Group (GRP): The target of these offensive posts is a group of people considered as a unity due to the same ethnicity, gender or sexual orientation, political affiliation, religious belief, or other common characteristic. Many of the insults and threats targeted at a group correspond to what is commonly understood as hate speech.
Other (OTH) The target of these offensive posts does not belong to any of the previous two categories (e.g. an organization, a situation, an event, or an issue).

4 Data Collection

The data included in OLID has been collected from Twitter. We retrieved the data using the Twitter API by searching for 10 keywords and constructions that are often included in offensive messages, such as 'she is' or 'to:BreitBartNews'[5]. We carried out a first round of trial annotation of 300 instances with six experts. The goal of the trial annotation was to 1) evaluate the proposed tagset; 2) evaluate the data retrieval method; and 3) create a gold standard with instances that could be used as test questions in the training and test setting annotation which was carried out using crowdsourcing. The breakdown of keywords and their offensive content in the trial data of 300 tweets is shown in Table 2.

Keyword	Offensive %
-filter:safe	57.1
medical marijuana	0.0
she is	35.7
he is	15.0
you are	23.5
to:BreitBartNews	26.1
to:NewYorker	8.3
gun control	14.3
they are	5.9

Table 2: The keywords from the trial annotation and the percentage of offensive tweets for each keyword.

One of the best offensive keywords was tweets that were flagged as not being safe by the Twitter 'safe' filter (the '-' indicates 'not safe'). The vast majority of content on Twitter is not offensive so we tried different strategies to keep a reasonable number of tweets in the offensive class amounting to around 30% of the dataset including excluding some keywords that were not high in offensive content such as 'they are' and 'to:NewYorker'. Although 'he is' is lower in offensive content we kept it as a keyword to avoid gender bias. In addition to the keywords in the trial set, we searched for more political keywords which tend to be higher in offensive content, and sampled our dataset such that 50% of the the tweets come from political keywords and 50% come from non-political keywords. In addition to the keywords 'gun control', and 'to:BreitbartNews', political keywords used to collect these tweets are 'MAGA', 'antifa', 'conservative' and 'liberal'. We computed Fliess' $kappa$

[5] *to* is a special Twitter word indicating that the tweet was written directly to a specific account (e.g., BreitBartNews).

on the trial set for the five annotators on 21 of the tweets. $kappa$ is .83 for Layer A (OFF vs NOT) indicating high agreement. As to normalization and anonymization, no user metadata or Twitter IDs have been stored, and URLs and Twitter mentions have been substituted to placeholders .

We follow prior work in related areas (Burnap and Williams (2015); Davidson et al. (2017)) and annotate our data using crowdsourcing. We used the crowdsourcing platform Figure Eight[6] and we ensure data quality by: 1) we only received annotations from individuals who were experienced in the platform; and 2) we used test questions to discard annotations of individuals who did not reach a certain threshold. Each instance in the dataset was annotated by multiple annotators and inter-annotator agreement has been calculated. We first acquired two annotations for each instance. In case of 100% agreement, we considered these as acceptable annotations, and in case of disagreement, we requested more annotations until the agreement was above 66%. The breakdown of the data into training and testing for the labels from each level is shown in Table 3.

A	B	C	Train	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	—	524	27	551
NOT	—	—	8,840	620	9,460
All			13,240	860	14,100

Table 3: Distribution of label combinations in the data.

5 Baseline Scores

We present baseline scores for each individual level using the test set. We report Precision (P), Recall (R), and F1 for each baseline on all classes along with weighted averages and Macro-F1 score. The results for the level A classes are presented in Table 4, for level B in Table 5, and for level C in Table 6.

In parallel to the OffensEval competition, we are assessing the performance of offensive language identification system on OLID using our own systems which rely on traditional and deep learning methods.

6 Conclusion

This paper presents OLID, a new dataset with annotation of type and target of offensive language. The dataset contains 14,100 tweets and is released freely to the research community. The dataset was used in the OffensEval: Identifying and Categorizing Offensive Language in Social Media (SemEval 2019 - Task 6) shared task.

To the best of our knowledge, this is the first dataset to contain annotation of type and target of offenses in social media and it opens several new avenues for research in this area. We present simple baseline scores on all the classes in each of the three levels. To assessment the performance of offensive language detection system we are analyzing the results of the OffensEval shared task and carrying out experiments using our own systems which rely on SVMs and neural networks as described in more detail in the next section.

In the future, we would like to make a cross-corpus comparison of OLID and datasets annotated for similar tasks such as aggression identification (Kumar et al., 2018) and hate speech detection (Davidson et al., 2017). This comparison is, however, far from trivial as the annotation of OLID is different.

Acknowledgments

The research presented in this paper was partially supported by an ERAS fellowship awarded by the University of Wolverhampton.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression Ddetection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *Advances in Information Retrieval*, pages 693–696. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

[6] https://www.figure-eight.com/

	NOT			OFF			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**Macro-F1**
All NOT	-	0.00	0.00	0.72	1.00	0.84	0.52	0.72	0.60	0.42
All OFF	0.28	1.00	0.44	-	0.00	0.00	0.08	0.28	0.12	0.22

Table 4: Results for offensive language detection (Level A). We report Precision (P), Recall (R), and F1 for each baseline on all classes (NOT, OFF), and weighted averages. Macro-F1 is also listed.

	TIN			UNT			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
All TIN	0.89	1.00	0.94	-	0.00	0.00	0.79	0.89	0.83	0.47
All UNT	-	0.00	0.00	0.11	1.00	0.20	0.01	0.11	0.02	0.10

Table 5: Results for offensive language categorization (level B). We report Precision (P), Recall (R), and F1 for each baseline on all classes (TIN, UNT), and weighted averages. Macro-F1 is also listed.

	GRP			IND			OTH			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
All GRP	0.37	1.00	0.54	-	0.00	0.00	-	0.00	0.00	0.13	0.37	0.20	0.18
All IND	-	0.00	0.00	0.47	1.00	0.64	-	0.00	0.00	0.22	0.47	0.30	0.21
All OTH	-	0.00	0.00	-	0.00	0.00	0.16	1.00	0.28	0.03	0.16	0.05	0.09

Table 6: Baselines for offense target identification (level C). We report Precision (P), Recall (R), and F1 for each baseline on all classes (GRP, IND, OTH), and weighted averages. Macro-F1 is also listed.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of WWW*.

Paula Fortuna, José Ferreira, Luiz Pires, Guilherme Routar, and Sérgio Nunes. 2018. Merging Datasets for Aggressive Text Identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 128–139.

Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. 2018. Convolutional Neural Networks for Toxic Comment Classification. *arXiv preprint arXiv:1802.09957*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets Against Blacks. In *Proceedings of AAAI*.

Prasenjit Majumder, Thomas Mandl, et al. 2018. Filtering Aggression from the Multilingual Social Media Feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 199–207.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1 – 16.

Hamdy Mubarak, Darwish Kareem, and Magdy Walid. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of ALW*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of WWW*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *arXiv preprint arXiv:1705.09899*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval*.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from Bullying Traces in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.