ROFF - A Romanian Twitter Dataset for Offensive Language

Mihai Manolescu
University of Tübingen
mihai.manolescu@student.uni-tuebingen.de

Çağrı Çöltekin
University of Tübingen
ccoltekin@sfs.uni-tuebingen.de

Abstract

This paper describes the annotation process of an offensive language data set for Romanian on social media. To facilitate comparable multi-lingual research on offensive language, the annotation guidelines follow some of the recent annotation efforts for other languages. The final corpus contains 5000 micro-blogging posts annotated by a large number of volunteer annotators. The inter-annotator agreement and the initial automatic discrimination results we present are in line with earlier annotation efforts.

1 Introduction

The use of words to hurt others is a curious aspect of natural languages. Besides the scientific curiosity, however, certain forms of offensive language can be harmful to individuals, may have discriminatory, toxifying effects on the society, and can be against law. An annotated corpus of offensive language is, hence, a valuable resource for understanding, identifying and preventing offensive content, particularly in online communication. This paper describes such an annotated corpus of Romanian on social media.

The study of offensive content online goes back to the earlier days of the Internet (Lea et al., 1992; Kayany, 1998). However, identifying various forms of offensive or abusive language online, such as hate speech or verbal harassment, has attracted considerable recent interest. The increased interest is evidenced by a number of recent shared tasks (Kumar et al., 2018; Wiegand et al., 2018; Zampieri et al., 2019b; Basile et al., 2019) to the extent that OffensEval 2020 shared task in SemeEval evaluation campaign (Zampieri et al., 2020) received submissions from 145 teams (out of 528 registered teams).

An important ingredient of these studies is an annotated corpus. Recent annotations efforts with the aim of detecting offensive language typically focus on a particular form of offensive content, such as hate speech (Agarwal and Sureka, 2015; Davidson et al., 2017; Del Vigna et al., 2017; ElSherief et al., 2018; Gao and Huang, 2017; Gitari et al., 2015; Sanguinetti et al., 2018a; Waseem, 2016, just to name a few), or cyberbullying (Xu et al., 2012; Dadvar et al., 2013; Dinakar et al., 2012; Nitta et al., 2013; Van Hee et al., 2015). However, it is often difficult to study a particular form of offensive language in isolation (Malmasi and Zampieri, 2018; Vidgen and Derczynski, 2020). Fore example, a classifier trained on a data set that annotates only instances of hate speech may detect other forms of offensive language as hate speech. As a result, some of the recent studies annotate multiple forms of offensive language together (Wiegand et al., 2018; Struß et al., 2019; Zampieri et al., 2019a). In particular, the annotation scheme for the OLID data set of English tweets by Zampieri et al. (2019a) has been used for annotating a variety of languages, including Arabic (Mubarak et al., 2020), Danish, (Sigurbergsson and Derczynski, 2020), Greek, (Pitenis et al., 2020), and Turkish (Çöltekin, 2020). Although not identical, the labels used in GermEval shared tasks (Wiegand et al., 2018; Struß et al., 2019) are also similar to the label set used by these data sets.

In this study we use a similar label set and follow similar guidelines for annotation of Romanian tweets. Following earlier annotation efforts for offensive language, we try to maximize the number of offensive tweets by obtaining part of the tweets to annotate using a list of offensive words, but also include a larger number of randomly selected tweets. The annotations were performed by volunteers.

For the remainder of this paper, we first describe the data set and the annotation process, present evaluation of the data and initial experiments with identifying offensive language and types of it for

Proceedings of Recent Advances in Natural Language Processing, pages 895–900
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_102

the present data, and conclude with a brief discussion and outlook.

2 Dataset

This dataset, to our knowledge, is the first of its kind for Romanian. It includes tweets from a wide range of topics between the second and third week of March in 2020. Since COVID-19 has been one of the main topics of the year, the data in this dataset consists largely of opinions about the virus, but also includes political slur and day to day tweets. In the following subsections, we provide an overview of how the data was collected and the annotation guidelines and process.

2.1 Data collection

The tweets for this dataset were collected using the Twitter API. We use two strategies to collect tweets. Since most of the offensive language annotation effort goes into finding few offensive languages posts among many non-offensive ones, annotating a completely random tweets is a big undertaking. As a result, most annotation studies collect the data based on queries that maximize the number of offensive posts. We followed the same practice, and a list of offensive words or abbreviations in Romanian such as *dracu* 'hell', *măta* 'your mother' or *pulă* 'dick' were used to collect approximately 1000 tweets. The number of offensive words used for this part of the task is 25. We complemented this list with tweets that were randomly sampled from the Twitter stream by querying 400 most frequent words based on the Romanian section of the Leipzig corpora (Quasthoff et al., 2014). This is a method used frequently, since the collected tweets can be seen as day to day life situations, without going to deep into a specific area, where multiple specific words are mentioned. We collected a total number of 15 000 full length tweets, from which we removed duplicates and promoted tweets. We also filtered the tweets based on the following criteria.

- Retweets were filtered out even if the original tweet was not part of the data set.
- Tweets with less than ten characters were filtered out.
- Any links inside of a tweet were removed.

After filtering, our data includes 820 tweets obtained through the query using offensive words, and we complemented the tweets to 5000 in total from the randomly obtained tweets.

2.2 Set of Labels

We follow the annotation scheme suggested by Zampieri et al. (2019a), and create a three-level annotation system. At first the annotator needs to decide whether a tweet is offensive or non-offensive. Following this decision, if a tweet is marked as offensive, it needs to be decided if it is targeted or non-targeted. The final level of the process is to decide the target of an offensive targeted tweet. This category is split into three different target classes, *individual*, *group* and *other*. A *group* is defined by being part of an entity such as race, ethnicity, political interests or gender. If the target of the offense is an individual, or multiple individuals that do not fit into the group definition above, then the target needs to be marked as *individual*. Cases that do not fit any of the two categories, for example an event or an organization, are being labeled as *other*. Together with the full data set, the annotation guidelines are published at `https://github.com/guzimanis/ROFF`. Although the relation is not necessarily a one-to-one relation, the offensive expressions that target a group are likely candidates for hate speech, while offensive statements that target individuals are likely to correlate with instances of cyberbullying or harassment. This set of labels does not cover every possible aspect of offensive language, For example, it may be interesting to annotate aspects of offensive language, like the aggressiveness (Basile et al., 2019) or the strength of the offense (Sanguinetti et al., 2018b). However, for simplicity and compatibility with a wider set of earlier annotation projects, our choice is limited to the label set defined above.[1]

2.3 Annotation Process and Data Analysis

The annotators were recruited from the author's contacts. All 33 annotators are native speakers of Romanian and have at least a high school degree or higher. Annotators volunteered for this project and did not get any benefits. The age of the annotators ranges between 18 and 55 years and the experience of using Twitter as a platform is between zero and 2 to 3 times weekly. It is worth mentioning that some annotators live outside Romania and can therefore have a different perception of

[1]The task involves a fair degree of subjectiveness. Different people have different interpretations of what is offensive. The additional aspects of the offensive language also tend to include more subjectivity, and lower inter-annotator agreement as reported in these studies.

offensiveness of a tweet. The annotators received clear instructions on how to perform the annotation and were asked to annotate 50 example tweets, before being handed the proper data. The initial annotations are reviewed by the authors and the annotation guidelines were revised to resolve some of the potential ambiguities in the initial guidelines that resulted in high disagreement.

All 5000 tweets in our sample were annotated by at least one annotator, and 2100 of the tweets received annotations from two annotators. We report the inter-annotator agreement on these 2100 doubly-annotated tweets. The agreement for the first level annotation (is the tweet offensive or not?) is 86.70 % (Cohen's Kappa $\kappa = 0.52$), for the second level (is the tweet targeted or not?) 58.30 % ($\kappa = 0.17$) and for the third level (who is the target of the tweet?) 43.20 % ($\kappa = 0.20$). Cohen's Kappa shows that there is a moderate agreement between the annotators when looking at a tweet and deciding if it is offensive or not. Presumably due to subjectivity of the task, and differences in annotation instructions, the reported annotator agreement on similar tasks vary considerably. As a reference, the study that forms the basis of our annotation scheme (Zampieri et al., 2019a) report a raw agreement rate of 60 %. The agreement on further levels of annotation is lower. The metrics and scores reported in earlier literature vary considerably. However, the low agreement is a known problem for this task (Vidgen and Derczynski, 2020).

The difficulty of this task can be observed in (1). The annotators disagreed on this example. Since most of the sentences that include foul language are due to the subjective judgment of each annotator, it is open to argument whether examples like these are offensive or not.

(1) *@sopeprotector lol am si eu ham din asta e al dracu de tampit tho il port cateodata*

@sopeprotector lol I have a harness from this too, it's damn stupid, I wear it sometimes

The second level of annotation is to decide if a tweet is targeted or not. For example, (2) was marked as targeted by one annotator but untargeted by the other. There is no clear answer that can be applied to each sentence, hence, there are many discrepancies between the annotators on this level.

(2) *@FIFAMobiledaily in pula cu satelitul ca trebuie sa ma duc la munca*

@FIFAMobiledaily In the cock with the satellite that I have to go to work

The most complicated part of the annotation was to decide which target the tweet has. Here the annotators had different answers in many cases, hence there the highest disagreement between the annotators is for this annotation level. Most of the time it is difficult to recognize the target directly, since it is not always directly mentioned in the tweet. As can be seen in (3), 'those' and 'TV guys' are possible targets of the tweet.

(3) *Ce plm au ăstia de la TV impotriva ăstora care vând chestii mai rare?*

What the fuck do these TV guys have against those who sell weirder stuff?

One of the annotators decided that this tweet is targeted towards an individual and the other chose the label *OTH*, since it is not clear enough which target it is. This confusion can be explained by the lack of context for these tweets. Since each tweet is being handled individually, the annotators often do not have a certain feeling for the context of the tweet. In general, some tweets can be seen as offensive without context, others need the context to be correctly annotated.

For doubly-annotated texts, the conflicts were resolved by the authors. Resolution action favored the offensive label in most cases. The final data set consists of 5000 tweets, from which 924 were labeled as offensive (18.48 %) and 4076 tweets as non-offensive. The detailed label distribution is presented in Table 1. All except one of the 820 tweets obtained by querying offensive words were

label	count	percent
non-offensive	4076	81.52
offensive	924	18.48
not targeted	185	20.02
targeted	739	79.98
group	176	23.82
individual	413	55.89
other	150	20.30

Table 1: The distribution of the labels in the final data set. The percentages represents percentage of the label within its parent category (e.g., 23.82 % of the *targeted* tweets have target 'group').

annotated as offensive, while only 3.47 % of the random tweets got an offensive label in the final annotation.[2]

3 Classification Experiments

In this section, we report initial classification results on the data set. Since we use a 'closed experiment' setup without any external information or data augmentation, the results can be improved. However, we believe the classification scores presented here can serve as a strong baseline.

Following OffensEval shared tasks (Zampieri et al., 2019b, 2020), we present results of three separate tasks. The first task, *task 1*, is identifying whether a given post is offensive or not. In *task 2*, we aim to identify whether an offensive tweet is targeted or not. Finally, *task 3* is about classifying the target of a targeted offensive tweet.

For all tasks, we use an LSTM-based classifier. The model uses words (tokenized using regular expressions) as input. The model first embeds the words in an embedding space of 64 dimensions. The embeddings are initialized randomly, and only learned during training. The embeddings are passed to a single left-to-right LSTM layer with 64 units and followed by a dropout rate of 0.10. The representation built by the LSTM layer at the final time step is passed to a fully-connected classification layer with a sigmoid or softmax activation depending on the task. Model was trained on 50 epochs with the Adam optimizer and cross-entropy loss function. The implementation uses Python Keras library (Chollet et al., 2015).

The results for all three tasks were presented in Table 2. The scores are similar to the earlier results obtained for other languages on similar data sets. For example the best macro-averaged F1 score reported in the OffensEval 2019 shared task for English are 82.90, 75.50 and 66.00 for tasks 1, 2 and 3 respectively.

Not surprisingly, the model's performance on detecting offensive language is better than detecting whether an offensive text targeted or not, which, in turn, is better than the 3-way classification of targeted texts to respective target groups. A reason for the low score on target classification is due to the fact that the label OTH is not clearly defined, since it can be used for various targets, e.g., events or organizations, and can therefore be often confused with other target groups.

[2]This means that to obtain the same amount of offensive tweets, we would need to annotate more than 25 000 randomly sampled tweets.

Task	Precision	Recall	F1 Score
Task 1	0.91	0.88	0.87
Task 2	0.89	0.88	0.86
Task 3	0.49	0.48	0.47

Table 2: Results of all three classification tasks. For compatibility with earlier reports in the literature, all scores are macro averaged.

4 Conclusion & Future Work

We presented a manually annotated corpus of Romanian offensive language on Twitter. Our annotation scheme is compatible with a number of recent offensive language annotation projects for other languages. As a result, the corpus is suitable for multi-lingual and cross-lingual research on analysis or detection of offensive language.

Overall, the inter-annotator agreement and initial machine learning experiments on the data yield similar results with earlier studies on offensive language. Although similar to the earlier studies, the offensive language detection results can definitely be improved using external resources, such as by using word embeddings, or by fine-tuning large pre-trained language models on this data. Furthermore, the uniform annotation scheme with other languages may allow cross-lingual transfer through translation and/or use of cross-lingual, shared word or sentence representations.

Although the data set created in this study comparable to many of the other data sets presented in the field (see Vidgen and Derczynski (2020) for a recent review), an obvious direction for future research is to increase the size and diversity of the data.

Most earlier data sets are either annotated by experts, or through crowd sourcing. In this study, we chose to rely on volunteers. Even though this is a common practice for finding participants for experiments in many fields, it is not common in annotation projects. Even though we did not use any gamification or other means to attract annotators, we received annotations from over 30 annotators. The present method can be applied where a crowd annotation is possible, and potentially result in a higher quality in comparison to the crowd sourcing. An interesting direction for future research is to

compare the quality of the data obtained through different methods of recruiting annotators.

Acknowledgements

We would like to thank every annotator who participated in this project and invested some of their time.

References

Swati Agarwal and Ashish Sureka. 2015. Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442. Springer.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 6174–6184. ELRA.

François Chollet et al. 2015. Keras. https://keras.io.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on 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 *Eleventh international AAAI conference on web and social media*, pages 512–515.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Joseph M Kayany. 1998. Contexts of uninhibited online behavior: Flaming in social newsgroups on Usenet. *Journal of the American Society for Information Science*, 49(12):1135–1141.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Martin Lea, Tim O'Shea, Pat Fung, and Russell Spears. 1992. *'Flaming'in computer-mediated communication: Observations, explanations, implications.*, pages 89–112. Harvester Wheatsheaf.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on Twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. 2013. Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 579–586, Nagoya, Japan. Asian Federation of Natural Language Processing.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Uwe Quasthoff, Dirk Goldhahn, and Thomas Eckart. 2014. *Building Large Resources for Text Mining: The Leipzig Corpora Collection*, Theory and Applications of Natural Language Processing, pages 3–24. Springer.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018a. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018b. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Julia Maria Struß, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 352–363, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop at KONVENS 2018*, pages 1–10.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of SemEval*.