BEING OFFENSIVE IN DANISH

A Machine Learning Project on Danish Offensive Language

Programme: Master's Programme in Language Technology

Topic: Machine Learning for Statistical NLP

Level: Advanced

Semester and year: Autumn 2021

Examiner: Asad Sayeed

Github: https://github.com/JoFunch/Projet-Hate

Abstract

Over the recent years, behaviour and language in online fora has become a quintessential focal point in general discourse. The attention is due to the harsh tone online communities tend to produce, why several official agents of online sites have made it their endeavour to combat unnecessary harmful and offensive language use on their "private pages". This project tests two different language models for the possible solution of automatic recognition of such offensive language.

1. Introduction

While the attention and debate has somewhat recent origin, abusive language detection is not a new phenomenon. Rather, the academic field has undergone quite the shift in more recent years and has increased tremendously in topicality with the rise of global, online social media. Throughout the evolution of online social interaction, language, generally, has changed. It has become a popular trait for online, public, as well as private agents to advocate and *enforce* "proper" rhetoric for a decent, clean and productive debate, as individuals and groups suffer real, physical harm as a consequence of foul language. One of the primary issues is, however, the overwhelming user-generated content produced everyday for humans to filter, why the use of automated solutions have been engaged for a more efficient clean up.

While being a field dating back to the turn of the century, the field of online abusive language has never been united in a common front, why quite the number of previous studies are rather incomparable (Nakov et al. 2021, 3; Sigurbergsson and Derczynski 2019a, 2; Zampieri et al. 2019a, 1). This incomparability originates with the studies being preoccupied with slightly different foci and taxonomy, wherefore unique terminology and classification overlap, differentiate, and at times, directly mismatch. The overall field counts *cyber-bullying*, *aggression detection, toxic comment detection,* and more recently, *abusive language detection, offensive language detection,* and *hate speech identification*, where this paper is primarily concerned with *cyber-bulling* (CB), *offensive language detection* (OL), and *hateful speech identification* (HS).

While HS is a direct expansion of the OL-field, both studies rely heavily from the rule of definition and terminology from that of CB, which was the original field of study. In order to understand the relation between these three fields, one is better to understand the novel dataset/annotation scheme of the coordinated Abusive Language Detection Tasks, called

SemEval/OffensEval & (S)OLID; initiatives made to iteratively create offensive language detection models (respectively Multi-lingual Offensive Language Identification on Social Media and (Semi-Supervised) Offensive Language Identification). The annotation scheme is very much a mirror of the evolution of study and as well of language use online beginning with CB and ending in HS. This is to be understood as a 3 hierarchical "ladder" of categorisation that will result in several classifications of a message. Both SemEval and SOLID shares the three-step classification-"tree":

Level A/1: Offensive Language Detection:

(OFF): Inappropriate language, insults, threats.

(NOT) Neither Offensive nor profane

Level B/2: Categorisation of Offensive Language:

(TIN): Targeted insult or threat towards a group or an individual.

(UNT): Untargeted profanity or swearing.

Level C/3: Offensive Language Target Identification

(IND): Individual, explicitly or implicitly

(GRP): Hate speech targeting a group of people based on ethnicity, gender,
sexual orientation, religion, or other common characteristic
(OTH): Targets that does not fall into the previous categories, e.g., organizations, events, and issues.

The progression of the three levels is very similar to that between CB, OL and HS. Characteristic to CB is that it primarily takes place amongst young teenagers on small online fora, where single users would experience being bullied (Sigurbergsson and Derczynski 2019a, 2). CB is characterized as targeted towards one specific receiver and typically spans over a period of time (ibid, 2). Another key feature for CB, and partially OL, is the language used to facilitate insults: where CB and OL would remain on a personal and individual level, HS is defined by relying on target-memberships and membership-features on a more general plane (ibid, 2). While HS might still be rhetorically targeting an individual, the individual itself is not being the target for whom his is rather than the category in which he belons. In that sense, the hate, directly or not, involves inevitably his individual feature(s); a feature which is projected to represent membership group as a whole.

1. 2. Related works

There have been several studies in Abusive Language Detection. For a general overview and discussion see (Nakov et al. 2021).

CB was somewhat initiated, or picked up, by (Xu et al. 2012) and (Dadvar et al. 2013), who made initial detection models basing their models on respective language features. They were followed by (Van Hee, Lefever, et al. 2015, A; Van Hee, Verhoeven, et al. 2015, B) who laid some of the more descriptive work of CB.

OL was pioneered by (Zampieri et al. 2020; 2019b; 2019a) who created the classification ground-work for OL and later HS as well as one of the widely used datasets. Further readings on OL and HS depend on the language in which one wish to inquire. For the purpose of this paper, the Danish dataset and paper is consulted, see (Sigurbergsson and Derczynski 2019a).

It is also worth consulting (Rosenthal et al. 2021) and the work on SOLID as well as the various SemEval-workshop and task forces focused on each their specific social media niche.

In section 2, the materials and methodology is described. Section 3 presents results and proposes a discussion on model, data and results followed by a conclusion.

2. Materials and Methodology

2.1 Dataset "DKHATE"

The dataset used for the language detection task is called *DKHATE* (Sigurbergsson and Derczynski 2019a), and is made according to the OLID-structure described in Section 1 with three subtasks in a mixture between binary classification and multiclass-classification. The creators sought to find sources where each of the three overall categories could be accessed excessively, wherefore the Danish sentence examples originates two sources, comments from news paper *Ekstra Bladet* page on Facebook and posts made on the /Denmark and /DANMAG page of Reddit. Originally, Twitter was included too in the mix of data, but was dropped from the dataset as the usage and general quantity of such data were too limited. From Facebook, Sigurbergsson & Derczynski manually collected data as crawling proved too difficult (ibid, 3), to which they describe the data as "great […] as they have high degree of variation [in language]" (ibid). The final dataset contains 3600 user-generated input, 800 from *Ekstra Bladet's* Facebook-page comment section, and 1400 scraped from the Danish Reddit pages.

Data Source	#Comments	% of all
Facebook / Ekstra Bladet	800	22.2
Reddit r/Denmark w off term	200	5.6
Reddit r/Denmark no off term	1200	33.3
Reddit r/DANMAG w off term	32	0.9
Reddit r/DANMAG	1368	38.0

Table 1 (Sigurbergsson and Derczynski 2019b, 4)

Total	Subtask A	#	%	Subtask B	#	%	Subtask C	#	%
3600	**OFF**	441	12,25	**None**	3161	87,81	**None**	3350	93,06
	NOT	3159	87,75	**UNT**	188	5,22	**IND**	95	2,64
				TIN	239	6,64	**GRP**	120	3,33
				TTH	12	0,33	**ORG**	34	0,94
							OTH	1	0,03
Sum		**3600**	**100,00**		**3600**	**100,00**		**3600**	**100,00**

Table 2, Original weights of data/class

2.2 Model 1: LSTM

The first model introduced to the detection task is an LSTM-model. The model consists of 1) an LSTM layer, 2) a relu layer, 3) two linear layers 4) and a sigmoid function.

2.3 Model 2: CNN

The second model in all three tasks consists of a 1-dimensional convolutional layer, relu, maxpool, into a single fully connected linear layer, a secondary relu finalized by a sigmoid function.

Originally, both models were more complex, but as a result of scarce data, simpler models were used.

2.4 Pre-processing and Hyperparameters.

As the dataset is a mixture of Reddit and Facebook-posts, several types of pre-processing methods were employed for data cleaning: URLS, Hashtags, Meta-commentary, HTLM-tags, enlongings of characters in words (such as 'heeeeeeello' for a more stable label-encoding), and a Danish lemmatization library provided by Søren Lind (https://github.com/sorenlind/lemmy) was employed for all three subtasks. Additionally to the pre-processed text, a Danish sentiment analysis library was employed to score the posts between -8 and +8 being respectively negative and positive with 0 as neutral (Nielsen 2011). The library also proved able to process emoticons why these have not been filtered out. The reason for not removing stop words is discussed further below.

The pre-processed data was then transformed using the KERAS library into a sequence of integers essentially capturing both the BoW-methodology but keeping the context, rather than transforming the dataframe to a traditional representation of the vocabulary as a whole. The now-sequenced data also underwent a rescaling to floats between 0-1 for data consistency. The sequences were then fixated and padded to specific, fixed lengths of the longest length of the data. Adam was used as optimizer algorithm for both CNN- and LSTM-model, and for subtask A, Binary Cross Entropy-loss function was applied while CrossEntropyLoss employed for subtask B and C. The learning rate of the two models were for subtask A and B 0.0000001, and C 0.00001 for both models. The split ratio for both sequence-to-text and the vectorised data were similar to that of Sigurbergsson & Derczynski, namely 80/20 train/test, and finally, all three subtasks in both models were trained on 100 epochs.

The dataset was also locally augmented to fit more input of the outlying classes. Originally, the data amount of the DKHATE was in total 3600 entries with respective 441 "Off" in Subtask A, 439 total of any offensive class in B, and 250 total in C (Table 2), but as numerous trial runs of the models were performed on this data to little or no success with unstable and at times erroneous training loops, the data was augmented. The augmented data-count is also seen in the Appendix A-C matrices. In total, the dataset had augmented an increase of the "offensive" data by 2, a doubling to a total of 881 "offensive" data points also increasing subtask B and C. The poor performance initially seen was discovered to be a result of the pre-processing done. Consequently, some of the Facebook-comments were nulled by the pre-processing, simply completely empty, as they originally consisted of only hashtags or URLS. To this extend, the data was once more altered during data fitting where each subtask had individual minimum and maximum lengths of inputs dropping input lower than 4 and longer than 70, which also became the aforementioned fixed

length. Beside the difference of form between the two platforms, the length of data "invited" by the two platforms are also drastically different. As seen in Appendix 4, data span between input of 1 entry and input with 800 entries (words tokenized).

3. Results and Discussions

CNN	**ACC**	**F1**
S A	53,07977207977209	31,52364273204904
S B	40,43304843304843	40,762812872467225
S C	28,809523809523807	28,48235294117647
LSTM	**ACC**	**F1**
S A	70,59544159544159	13,240418118466898
S B	13,646723646723646	13,706793802145412
S C	07,61904761904762	07,341176470588234

Table 3 Results CNN and LSTM

Confusion matrices of the three subtasks for each model is attached to appendix A-C.

As described in the previous section, the testing involved two different model-types, an LSTM and a CNN-model. From Table 3, the CNN-model performed significantly better in subtask B and C and on the F1-score of Subtask A, where the LSTM dominated sheer accuracy. However, from the Appendix A-C matrices of all three subtasks for both models, it becomes clear that the CNN did not entirely "outperform" the LSTM as Table 4 might suggest. In Subtask A and C, the LSTM has a higher accuracy in the majority "None"-class compared to the CNN, whose primary advantage in resides with detection of outlier-classes. The CNN misses the neuter class in both A, where "Off" input is misjudged for "Not", and C, where a significant portion of "None"-inputs are mistaken for "GRP". Contrary, the CNN captures Subtask B's "TIN" and "UNT" as well as Subtask C's "GRP" rather well.

What is also evident from the performance of both models is the decreasing performance with each task. The reason for this revolves around the data, of which in DKHATE is heavily misrepresented. As seen from Table 1 and 2, the Danish offensive and hate-speech data is by no means properly represented with outlier-classes as low as Subtask B class "TTH" with 0,09% representation. This is by no means enough for viable predictions, which also explains why it for both models resulted a null-score. Furthermore, the data which *is* available, as described in Section

2, origins from two very different online platforms. This means that it simply is *too* different from each other; the textual input gathered from Facebook opposing that from Reddit has too very different formats and contextual "rules" or norms. Ultimately, the same category/class of data (from two different platforms) share few characteristics, and in combination with the extreme scarcity, the data becomes exceptionally difficult for the distributional models to understand.

In general, there is serious improvements to be made before such a system could produce reliable predictions of hate-speech, although immediate offensive language estimations are not far away. Beneficial to and in accordance with the original project made by Sigurbergsson and Derczynski, BERT-models proved solid when predicting. Evidently, local understanding of tokens improves predictions significantly, and it appears that complex meaning representations is highly beneficial to tasks at hand with data varying so much depending on its origin. Such representations would also benefit from further preprocessing, i.e. more linguistic information extracted per input. Here, additions counting more detailed sentiment analysis, in-depth hashtag understanding, and other condensing methods, such as TFIDF, could prove highly useful. Another factor beneficial to a better performance would be to separate the data-origin based on social platform and train models for input-structures "simplistic" like facebook-comments vice versa Reddit-posts. And in addition to that, if anything proven by this project, a larger data-set in general seems imperative.

Conclusion

This paper attempted to predict various sub-types of offensive language from online social media platforms. The immediate conclusions are that a more detail-oriented model is required for any solid predictions and that a larger dataset ought to be made, and that the use of simple models for small data-sets proves just only decent. It was also noted that data-format of the two places of origin of offensive textual input differs significantly from each other in form, language and structure, wherefore a relatively small dataset proves difficult to make predictions based on these.

A future improvement would take off in the dataset which must be expanded and possible separated to fit only one textual format. Furthermore, a premade, token-representation based model could prove useful in assisting with extracting meaning from the at times scarce and short texts. It would also be beneficial for the performance of the model if the data pre-processing included more linguistic and meta-linguistic information in the assistance of generalisation.

Reference List

Dadvar, Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. 'Improving Cyberbullying Detection with User Context'. In *Advances in Information Retrieval*, edited by Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, 7814:693–96. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36973-5_62.

Nakov, Preslav, Vibha Nayak, Kyle Dent, Ameya Bhatawdekar, Sheikh Muhammad Sarwar, Momchil Hardalov, Yoan Dinkov, Dimitrina Zlatkova, Guillaume Bouchard, and Isabelle Augenstein. 2021. 'Detecting Abusive Language on Online Platforms: A Critical Analysis'. *ArXiv:2103.00153 [Cs]*, February. http://arxiv.org/abs/2103.00153.

Nielsen, Finn Årup. 2011. 'A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs'. *ArXiv:1103.2903 [Cs]*, March. http://arxiv.org/abs/1103.2903.

Rosenthal, Sara, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. 'SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification'. *ArXiv:2004.14454 [Cs]*, September. http://arxiv.org/abs/2004.14454.

Sigurbergsson, Gudbjartur Ingi, and Leon Derczynski. 2019a. 'Offensive Language and Hate Speech Detection for Danish'. *ArXiv:1908.04531 [Cs]*, August. http://arxiv.org/abs/1908.04531.

———. 2019b. 'Offensive Language and Hate Speech Detection for Danish'. *ArXiv:1908.04531 [Cs]*, August. http://arxiv.org/abs/1908.04531.

Van Hee, Cynthia, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy Pauw, Walter Daelemans, and Véronique Hoste. 2015. 'Detection and Fine-Grained Classification of Cyberbullying Events'. In .

Van Hee, Cynthia, Ben Verhoeven, Els Lefever, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. 'Guidelines for the Fine-Grained Analysis of Cyberbullying, Version 1.0'. In .

Xu, Jun-Ming, 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*, 656–66. Montréal, Canada: Association for Computational Linguistics. https://aclanthology.org/N12-1084.

Zampieri, Marcos, 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*, 1415–20. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1144.

———. 2019b. 'SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)'. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 75–86. Minneapolis, Minnesota, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2010.

Zampieri, Marcos, 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)'. *ArXiv:2006.07235 [Cs]*, September. http://arxiv.org/abs/2006.07235.

Appendix A-C

CNN	Not	Off
Not	358	309
Off	82	90
LSTM	Not	Off
Not	571	96
Off	153	19
Total/ Test set 3356 / 839	2638/667/ 79,4 %	718/172 / 20,5 %

Subtask A

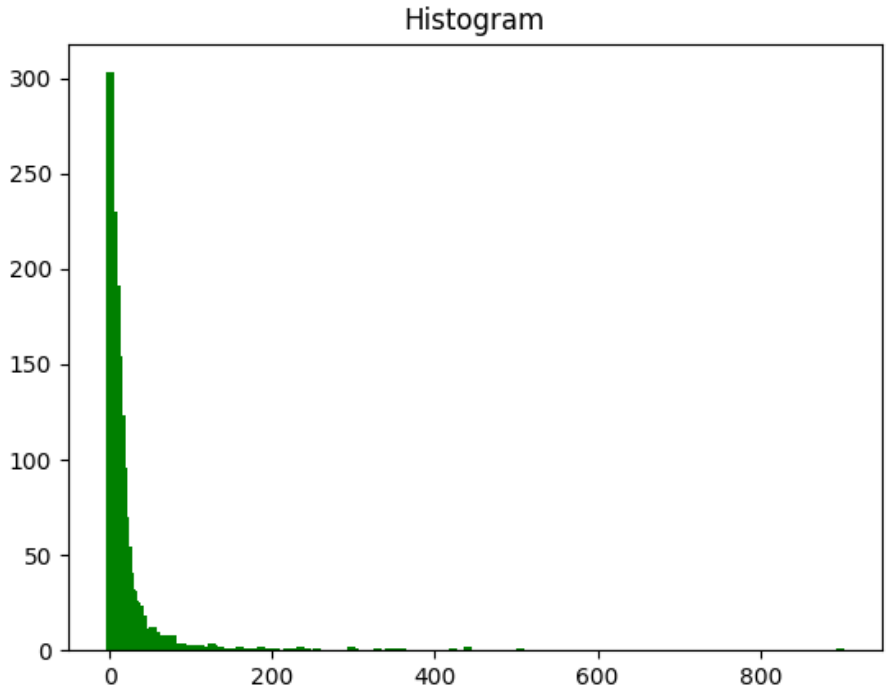
CNN	None	TIN	TTH	UNT
None	294	245	5	123
TIN	47	37	3	14
TTH	5	2	0	1
UNT	34	17	1	11
LSTM				
None	56	369	95	147
TIN	8	68	12	13
TTH	0	6	0	2
UNT	6	32	11	14
Total/test 3359/839	2638/667/79,4%	402/101/12,0 %	22/8/0,09 %	294/63/7,5%

Subtask B

CNN	GRP	IND	None	ORG
GRP	47	2	0	4
IND	30	5	3	7
None	569	37	11	111
ORG	10	2	1	1
LSTM				
GRP	15	0	20	18
IND	18	3	15	9
None	292	19	233	195
ORG	6	1	6	1
Total/Test 3359/840	208/53/6,3%	170/45/5,3 %	2923/728/86,6%	58/14/1,6%

Subtask C

Appendix 4



Original data sorted according to occurrences of lengths