# "*Zo Grof*!": A Comprehensive Corpus for Offensive and Abusive Language in Dutch

**Ward Ruitenbeek, Victor Zwart, Robin van der Noord,**
**Zhenja Gnezdilov, Tommaso Caselli**
CLCG, University of Groningen, The Netherlands
t.caselli@rug.nl

## Abstract

This paper presents a comprehensive corpus for the study of socially unacceptable language in Dutch. The corpus extends and revise an existing resource with more data and introduces a new annotation dimension for offensive language, making it a unique resource in the Dutch language panorama. Each language phenomenon (abusive and offensive language) in the corpus has been annotated with a multi-layer annotation scheme modelling the explicitness and the target(s) of the abuse/offence in the message. We have conducted a new set of experiments with different classification algorithms on all annotation dimensions. Monolingual Pre-Trained Language Models prove as the best systems, obtaining a macro-average F1 of 0.828 for binary classification of offensive language, and 0.579 for the targets of offensive messages. Furthermore, the best system obtains a macro-average F1 of 0.637 for distinguishing between abusive and offensive messages.

## 1 Introduction

Social Media platforms have become an intrinsic part of the lives of lots of people. A phenomenon that accompanies Social Media platforms, with serious impacts on society, is the presence of socially unacceptable language. Socially unacceptable language is to be regarded as a generic umbrella term comprehending many different user-generated language phenomena such as toxic language (Karan and Šnajder, 2019; Bhat et al., 2021), offensive language (Zampieri et al., 2019c; Ranasinghe and Zampieri, 2020; Zampieri et al., 2020), abusive language (Karan and Šnajder, 2018; Caselli et al., 2020; Wiegand et al., 2021), hate speech (Waseem and Hovy, 2016a; Davidson et al., 2019; Basile et al., 2019), among others. While manually monitoring and flagging these phenomena is impossible, there has been a growing interest in the Computational Linguistics (CL) and Natural Language Processing (NLP) communities to develop automatic systems to flag messages containing these phenomena.

Besides the limitations of this type of reactive interventions, previous work (Nozza, 2021) has shown the necessity of language specific resources for these phenomena to properly train systems. This work contributes in this direction by presenting a comprehensive dataset to identify socially unacceptable language in Twitter messages in Dutch. We integrate and extend DALC v1.0 (Caselli et al., 2021) by introducing a new annotation layer for offensive language and expanding the size of the dataset from 8,156 messages to 11,292. The main contribution of this paper can be summarised as follows:

- a new release of DALC, DALC v2.0, with a) more than 3k newly annotated messages and b) annotations for the offensive language dimension;
- an extensive set of experiments to model the different annotation dimensions involved;
- an error analysis showing the limits of current models.

The annotation guidelines, the data, and the code for the reported experiments, and a data statement (Bender and Friedman, 2018) are publicly available.[1] Examples of offensive messages have been redacted to preserve privacy and explicit offensive lexical items have been obfuscated.

## 2 Offensive Language: Why and How

Offensive language is a broader language phenomenon when compared to other phenomena and behaviours (e.g., abusive language, hate speech or cyberbullying) and, most importantly, more subjective (Vidgen et al., 2019; Poletto et al., 2021). In

[1] https://github.com/tommasoc80/DALC



*Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 40 - 56

| Offensive Language (Zampieri et al., 2019a) | Abusive Language (Caselli et al., 2021) |
|---|---|
| Posts containing any form of non-acceptable language (profanity) or a targeted offence, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words. | Impolite, harsh, or hurtful language (that may contain profanities or vulgar language) that result in a debasement, harassment, threat, or aggression of an individual or a (social) group, but not necessarily of an entity, an institution, an organisations, or a concept. |

Table 1: Definitions of offensive and abusive language adopted in this work.

general, the use of offensive language is intrinsically connected to freedom of speech. However, in the context of social media interactions, the presence and use of offensive language towards other users should raise concerns because it may escalate the exchange in deeper verbal hostility (e.g., hate speech) and give rise to highly toxic, and unsafe environments (Chowdhury et al., 2020).

While we can identify and list parameters and details that help us to narrow down whether a message is abusive or not, the offensiveness of a message is only partially dependent on its content. Other variables such as the context of occurrence, the background and experience of the reader/annotator play a relevant role. Despite these difficulties, offensive language datasets have been developed in different languages (Sigurbergsson and Derczynski, 2020; Pitenis et al., 2020; Çöltekin, 2020; Chowdhury et al., 2020) and used in recent shared tasks (Zampieri et al., 2019c, 2020).

To maximise resource interoperability and foster the study of offensive language from a multilingual perspective, we adopt the definition of offensive language from Zampieri et al. (2019c). In Table 1 the full definition is reported and compared with the definition of abusive language adopted in the Dutch Abusive Language Corpus (DALC) v1.0. A key element distinguishing these two language phenomena is the level of detail used to describe them, the different emphasis on the intentions of the producers, the presence/absence of a target, and the effects on the receivers. In particular, target is an essential and compulsory element of abusive language, while it is not the case for offensive messages. On the other hand, given its more generic nature, offensive language can be identified in messages that do not contain any target. This is particularly evident in the use of profanities to express strong (positive or negative) emotions. To better clarify the difference between the two phenomena consider the following examples from DALC v2.0:

1. ER IS EEN F***** MUG EVEN GROOT ALS MIJN DUIM
   [There is a f****** as big as my thumb]

2. Elke [identity_term] is een potentiële terrorist
   [Every [identity_term] is a potential terrorist]

Example 1 instantiate an offensive message, due to the presence of a profanity. Its perception of being offensive can vary according to the context of use and the receivers of the message. At the same time, the message does not fully comply with the definition of abusive language for multiple reasons: there is not a (human) target and there is no intention to debase or harass an individual/group. Example 2, on the contrary, it is a clear case of abusive language. Here the abusive is express via a stereotype and a debasing act, and with an explicit target realised via a specific identity term. The message is abusive and also offensive.

In this work, we have maintained the multi-layer annotation approach of DALC v1.0, distinguishing between the explicitness of the message and its target. The explicitness and the target layers for the offensive dimension have been refined with subclasses along the existing annotation of abusive language. The explicitness layer distinguishes three subclasses: (i.) EXPLICIT; (ii.) IMPLICIT; and (iii.) NOT. While NOT is used to annotate not offensive messages, the difference between the EXPLICIT and IMPLICIT subclasses mainly rely on the surface forms of the message. Explicit offensive content refers to the presence of profanities or combination of words that unambiguously make the message offensive. Implicit messages are more subtle, lacking any surface markers, thus making the offence hidden (Waseem et al., 2017).
The target layer, on the other hand, extends the classes used for abusive language allowing for the absence of a target. In particular, we have four subclasses defined as follows: (i.) INDIVIDUAL, for messages that are addressed or target a specific

| Id | Text | Explicitness | Target |
|---|---|---|---|
| 1. | Dat gebeurt in het park en veel jongeren bij elkaar<br>[That happens in a park and many young people together] | NOT | NOT |
| 2. | En daar trap jij in. Echt slim<br>[And you fall for that. really smart] | IMP. | IND. |
| 3. | S*** worden ze niet gemaakt<br>[They don't get any d****] | EXP. | GRP. |
| 4. | j*** dat was wel schrikken geweest<br>[J**** that was scary] | EXP. | NOT |
| 5. | ons geld vervangen door die sh** euro<br>[ replace our money by that sh** euro ] | EXP. | OTH. |

Table 2: Examples of the annotation of the explicitness and the target layers. EXP. = EXPLICIT, IMP. = IMPLICIT; IND. = INDIVIDUAL, GRP. = GROUP, OTH. = OTHER. English translations in brackets.

person or individual (who could be named or not); (ii.) GROUP, for messages that target a group of people considered as a unity because of ethnicity, gender, political affiliation, religion, disabilities, or other common properties; (iii.) OTHER, for messages that target concepts, institutions and organisations, or non-living entities; and (iv.) NOT, for offensive messages without a target. In Table 2, we report some redacted examples from the dataset to illustrate the combination of the two layers in the annotation process.

**Data Collection and Annotation** DALC v1.0 is a corpus of 8,156 messages from Twitter in Dutch obtained by applying three different collection methods: keywords extraction, message geolocation, and seed users. We have extracted a total of 10k messages using only the keywords and seed users data from DALC v1.0, since these two sources proved to be denser and more suitable for the language phenomenon of interest. Following the settings of DALC v1.0, there is no overlap of messages concerning topic and authors between train and test distributions. Consequently, the 10k messages are equally and independently extracted from the train and test candidates - resulting in 5k messages per distribution. We divided the messages of each distribution in batches of 1k each for the annotation.

Given the highly subjective nature of offensive language, all annotations for both layers have been conducted in parallel by four annotators. [2] Annotators were asked to apply the definition of offensive language as reported in Table 1. Each offensive message was then annotated for the explicitness and the target layers.

The annotation has been conducted in two steps. In the first step, the annotators focused on all 6,267 messages that were marked as not abusive in DALC v1.0. This is a necessary curation phase in order to be compliant with the distinction between offensive and abusive language. In the second steps, we have annotated 5 additional batches for train and 1 batch for test. The final amount of annotated data is 12,251.[3]

Table 3 reports the pairwise Cohen's Kappa score for all the four annotators for the explicitness and the target layers. The agreement scores have been computed on all the annotated data. The agreement for explicitness layer ranges between a minimum of 0.330 to a maximum of 0.541, indicating a slight/substantial agreement, with a global Fleiss' Kappa of 0.430. It is worth noting that there is a variation in agreement across the annotators, with A.1 and A.3 being the strongest pair. Kappa scores slightly increase when aggregating the explicitness subclasses into a generic offensive (OFF) label. In this case, the values range between 0.358 (A.2–A.4) and 0.593 (A.1–A.3), with a Fleiss' Kappa of 0.473. The results for the annotation of the target layer are slightly worse, with the minimum agreement being a Cohen's Kappa of 0.250 (A.2–A.3) and a maximum of 0.474 (A.1–A.3). Overall Fleiss's Kappa for the target layer is 0.402.

To better understand these results, we have anal-

[2] The annotators are also authors of this paper.

[3] 15 messages from the last training batch were not annotated.

| Explicitness | A.1 | A.2 | A.3 | A.4 |
|---|---|---|---|---|
| **A.1** | – | 0.457 | 0.541 | 0.412 |
| **A.2** | – | – | 0.373 | 0.330 |
| **A.3** | – | – | – | 0.471 |

| Target | A.1 | A.2 | A.3 | A.4 |
|---|---|---|---|---|
| **A.1** | – | 0.391 | 0.474 | 0.379 |
| **A.2** | – | – | 0.304 | 0.250 |
| **A.3** | – | – | – | 0.457 |

Table 3: Inter-Annotator Agreement for the Explicitness and the Target layers - pairwise Cohen's Kappa.

ysed the pairwise confusion matrices of all the annotators.[4] For the explicitness layer, it clearly appears that the biggest source of disagreement is the offensive status of the message rather than the distinction between explicit or implicit, further supporting the claim that offensiveness is subjective. This has also an impact on the target layer: if a message is not annotated as offensive, the target annotation is ignore.

**Handling of disagreements** We adopt a majority voting for handling the disagreements and assigning final labels. In all cases where a tie is reached, the examples have been discussed collectively to reach a consensus. However, when subjectivity is an essential property of a language phenomenon, disagreements are more informative than detrimental (Aroyo et al., 2019; Basile, 2020; Leonardelli et al., 2021). In line with this vision, the final distribution contains the disaggregated annotations to promote further research on the relationship of subjectivity and annotation of natural language phenomena.

## 3 Data Overview

The annotated corpus contains 11,292 Twitter messages in Dutch, covering a time period between November 2015 and August 2020. For completeness, all messages marked as offensive and containing a target have also been further annotated for abusiveness. For abusive language, we applied the same annotation procedure used in DALC v1.0. Table 4 illustrates the distribution of the data for the abusive and offensive dimensions, and the target layers across the Train/Dev and Test distributions.

The unbalanced distribution between the negative and the positive examples for both the abusive and the offensive dimensions is part of the design strategy. While the actual distribution of these classes in social media is unknown, a distribution of 2/3 *vs*. 1/3 between negative and positive examples appears to be more realistic than a perfectly balanced dataset and in line with previous work (Basile et al., 2019; Davidson et al., 2017; Zampieri et al., 2019c, 2020).

[4]See Appendix B for details.

| Annotated Dimension | Subclass | Train | Dev | Test | Total |
|---|---|---|---|---|---|
| Abusive | EXP | 855 | 127 | 328 | 1,310 |
| | IMP | 536 | 116 | 135 | 787 |
| | NOT | 5,426 | 962 | 2,807 | 9,195 |
| Offensive | EXP | 1,407 | 230 | 584 | 2,221 |
| | IMP | 1,070 | 209 | 283 | 1,562 |
| | NOT | 4,340 | 766 | 2,403 | 7,509 |
| Target - Abusive | IND | 777 | 127 | 254 | 1,158 |
| | GRP | 470 | 87 | 158 | 715 |
| | OTH | 144 | 29 | 51 | 224 |
| Target - Offensive | IND | 1,147 | 191 | 361 | 1,699 |
| | GRP | 705 | 133 | 244 | 1,082 |
| | OTH | 489 | 93 | 157 | 739 |
| | NOT | 136 | 22 | 105 | 263 |

Table 4: DALC v2.0: Distribution of subclasses in Train, Dev, and Test splits for abusive, offensive dimensions and target layers. Target is split between target of abusive messages and target of offensive messages.

Overall, 2,097 messages have been annotated as abusive, with an increase of 208 messages when compared to DALC v1.0. On the other hand, 3,783 messages have been marked as offensive. In both dimensions, the explicit subclass represents the majority, with 62.47% of cases for the abusive dimension and 58.71% for the offensive one. The difference in the distribution of the implicit subclass is striking, with implicit offensive messages being almost the double of the abusive counterpart. A possible explanation can be found in the definitions of the two phenomena and their annotations: offensive messages have been labelled as such either because they contained a profanity, or because the annotators *subjectively perceived* them as offensive.

As for the targets, we observe that only a minority of offensive messages does not have a target (6.95%). When compared to other datasets for offensive language, the amount of messages associated with this class varies - for instance, being

the majority class in Sigurbergsson and Derczynski (2020) but not the minority in Zampieri et al. (2019b) - suggesting that there may be a dependency of this subclass on the method(s) used for collecting the data. On the other hand, differences in the realisation of the targets are more evident when focusing on the IND and GRP subclasses. Offensive messages have a balanced distribution between these two subclasses corresponding to 28.25% and 28.60% of all the targets, respectively. On the contrary, abusive messages see a majority of cases (55.22%) for the IND subclass, and relatively fewer cases for GRP (34.09%). Lastly, the OTH subclass has been selected more often (19.53%) with offensive messages than with the abusive ones (only 10.68%). This difference can be again explained in the light of the definitions of the two phenomena.

No significant difference in length has been found between abusive and offensive messages (average length abusive 28.79 words; average length offensive 28.44),[5] while this is not the case for offensive and not offensive messages (average length not offensive 21.93 words; average length offensive 28.44). [6] Similarly to DALC v1.0, significant differences in length between implicit and explicit messages appear only in the Test distribution, where implicit offensive messages have an average of 30.04 words compared to the 23.55 words of the explicit messages.

To gain better insights into the data and the differences between the two dimensions, we have extracted and compared the top-50 keywords between the Train and Test distributions by collapsing the subclass in the explicitness layer, resulting in OFFENSIVE, ABUSIVE, NOT (Table 11 in Appendix B illustrates the top-10 keywords). While we observe a lack of overlapping lexical items between Train and Test distributions, and the absence of any topic-specific lexical items, the differences between offensive and abusive language are not as neat as one would imagine. Besides the presence of some profanities or slurs, most of the keywords do not present any specific denotative or connotative markings for offensive and/or abusive language.

## 4 Experiments

We ran a set of experiments to validate the newly annotated corpus. We first focused on the identification of the offensiveness dimension (§ 4.1), and then on the target layer (§ 4.2). We also investigate the ability of systems to distinguish between offensive and abusive dimensions (§ 4.3). We tested four different architectures: a Linear SVM combining character and word n-gram TF-IDF vectors, a Bi-LSTM model initialised Coosto pre-trained word embeddings,[7] and two monolingual Transformer-based pre-trained Language Models (PTLMs), namely BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020). The two PTLMs differ with respect to their architectures (BERT *vs.* RoBERTa), the size (12GB *vs.* 39GB) and origin of the data used to generate the models (manually selected data *vs.* the Dutch section of the automatically derived OSCAR corpus (Suárez et al., 2019)). All models are trained on the Train split and evaluated against the held-out, non-overlapping Test split. The Dev split is used for tuning of the systems' (hyper)parameters. Models are compared using the macro-average F1. However, given the imbalance among the subclasses in the different layers, for each subclass, we also report Precision and Recall. For the offensiveness and the offensive target dimensions, systems are compared against a dummy classifier based on the majority class. In all experiments, a common preprocessing approach is applied. All preprocessing steps and (hyper)parameters are detailed in Appendix A for replicability.

### 4.1 Detecting Offensive Language

We have first modeled the offensiveness dimension both as a binary classification task, by collapsing the EXPLICIT and IMPLICIT subclasses into a single value, namely OFF(ENSIVE). Given the distribution of the annotated data, the task is already challenging. The second experiment setting follows the fine-grained, tripartite distinction between EXPLICIT, IMPLICIT and NOT.

Table 5 presents the results for the binary setting. All models outperform the dummy baseline, with RobBERT achieving the best results (macro-average F1 of 0.828). Interestingly, the second best system is the Bi-LSTM rather than the other PTLM, BERTje, with a macro-average F1 of 0.823. When comparing the results of these two latter models, we observe that BERTje underpeforms on the OFF label, especially for Recall. A possible ex-

[5] Statistical test: Mann-Whitney Test; $p > 0.05$
[6] Statistical test: Mann-Whitney Test; $p < 0.05$
[7] https://github.com/coosto/dutch-word-embeddings

| System | Class | Precision | Recall | Macro-F1 |
|---|---|---|---|---|
| Dummy | OFF | 0.0 | 0.0 | 0.423 |
| | NOT | 0.734 | 1.0 | |
| SVM | OFF | 0.644 | 0.513 | 0.718 |
| | NOT | 0.836 | 0.898 | |
| Bi-LSTM | OFF | $0.733_{0.015}$ | $0.749_{0.015}$ | $0.823_{0.004}$ |
| | NOT | $0.908_{0.004}$ | $0.901_{0.009}$ | |
| BERTje | OFF | $0.721_{0.010}$ | $0.693_{0.022}$ | $0.802_{0.002}$ |
| | NOT | $0.891_{0.006}$ | $0.903_{0.008}$ | |
| RobBERT | OFF | $0.756_{0.005}$ | $0.737_{0.013}$ | $\mathbf{0.828}_{0.006}$ |
| | NOT | $0.906_{0.004}$ | $0.914_{0.001}$ | |

Table 5: DALC v2.0: Offensive language, binary classification. Lower script numbers show standard deviations over 3 different runs. Best scores in bold.

planation can be found by taking into account the properties of the embedding representations of the models. The Coosto word embeddings used to initialise the Bi-LSTM have been obtained by using a large amount of messages from social media (624 million messages out of a total of 660 million texts), making them more suitable and inline with the text variety of the dataset. This may also be one of the reasons why RobBERT performs best: the data used to generate its embeddings are also from the Web, although not specifically from social media posts. To further validate the behaviour of the Bi-LSTM model, we ran a further set of experiments using random pre-trained embeddings obtained from the Dutch CoNLL17 corpus[8] (Fares et al., 2017). The embeddings are smaller than the Coosto ones (100 dimensions *vs.* 300 dimensions for Coosto), and obtained from a different data distribution. While the results[9] are lower (macro-F1 $0.799_{0.004}$), they are still competitive, with the macro-F1 falling within the standard deviation of BERTje.

All systems achieve very good results on the negative class but suffer on the positive one. This is mainly due to the lack of overlapping elements between the Train/Dev and the Test split, besides the impact of the unbalanced distribution of the data in the training data. This is particularly evident for the Recall of the OFF class of the SVM which is barely above 0.5. Finally, in absolute terms, the results of the top systems are in line with those reported for comparable datasets in other languages (Zampieri et al., 2020).

[8] http://vectors.nlpl.eu/repository/

[9] OFF Precision: $0.737_{0.058}$, OFF Recall: $0.680_{0.064}$; NOT Precision: $0.888_{0.016}$, NOT Recall: $0.908_{0.034}$

The outcome of the fine-grained experiments are detailed in Table 6. Rather than focusing only on the best systems, we have experimented with all of them to see whether the patterns observed in the binary classification remain valid.

| System | Class | Precision | Recall | Macro-F1 |
|---|---|---|---|---|
| SVM | EXP | 0.710 | 0.395 | 0.543 |
| | IMP | 0.297 | 0.212 | |
| | NOT | 0.820 | 0.936 | |
| Bi-LSTM | EXP | $0.766_{0.044}$ | $0.709_{0.058}$ | $0.658_{0.004}$ |
| | IMP | $0.423_{0.039}$ | $0.268_{0.025}$ | |
| | NOT | $0.889_{0.009}$ | $0.942_{0.014}$ | |
| BERTje | EXP | $0.762_{0.015}$ | $0.639_{0.054}$ | $0.663_{0.018}$ |
| | IMP | $0.374_{0.033}$ | $0.434_{0.032}$ | |
| | NOT | $0.887_{0.007}$ | $0.904_{0.032}$ | |
| RobBERT | EXP | $0.735_{0.007}$ | $0.724_{0.012}$ | $\mathbf{0.667}_{0.005}$ |
| | IMP | $0.370_{0.007}$ | $0.358_{0.042}$ | |
| | NOT | $0.904_{0.005}$ | $0.911_{0.02}$ | |

Table 6: DALC v2.0: Explicitness layer classification. Lower script numbers show standard deviations over 3 different runs. Best scores in bold.

The picture that emerges is slightly different. The performances on the EXP and the NOT subclasses are almost unchanged for the neural-based systems, while they dramatically drop for the EXP subclass for the SVM model. All systems struggle to distinguish the IMP subclass, with the Bi-LSTM achieving the best Precision. When compared to the binary classification, the results of the two PTLMs are closer and marginally better than the Bi-LSTM, confirming RobBERT as the best system (macro-average F1 0.667). Interestingly, BERTje has the highest Recall score for the IMP subclass.

### 4.2 Detecting the Targets

Target identification has an important role within the more general task of offensive language identification, especially because it can help to better assess the seriousness of the offence and contribute to the study of more specific phenomena such as hate speech (Waseem et al., 2017; Zampieri et al., 2019b). In particular, messages containing a target can be further annotated by distinguishing whether they express an insult or stronger forms of degradation (e.g., abusive language, or hate speech), and by refining the types of target (e.g., gender, race/ethnicity, political orientation, disabilities, among others).

In these experiments, we have assumed a perfect labelling of the messages for offensiveness.

This results in a reduced number of messages that we can use for training and testing our systems. Similarly to the offensiveness dimension, we have compared our results against a dummy classifier that always predicts the most frequent label, i.e., IND. The results are reported in Table 7.

| System | Class | Precision | Recall | Macro-F1 |
|---|---|---|---|---|
| Dummy | IND | 0.416 | 1.0 | 0.147 |
| | GRP | 0.0 | 0.0 | |
| | OTH | 0.0 | 0.0 | |
| | NOT | 0.0 | 0.0 | |
| SVM | IND | 0.587 | 0.892 | 0.467 |
| | GRP | 0.631 | 0.561 | |
| | OTH | 0.535 | 0.286 | |
| | NOT | 0.666 | 0.114 | |
| Bi-LSTM | IND | $0.605_{0.023}$ | $0.844_{0.054}$ | $0.471_{0.009}$ |
| | GRP | $0.673_{0.049}$ | $0.551_{0.065}$ | |
| | OTH | $0.466_{0.075}$ | $0.346_{0.047}$ | |
| | NOT | $0.359_{0.068}$ | $0.130_{0.033}$ | |
| BERTje | IND | $0.692_{0.005}$ | $0.863_{0.009}$ | $\mathbf{0.579}_{0.002}$ |
| | GRP | $0.685_{0.016}$ | $0.677_{0.020}$ | |
| | OTH | $0.600_{0.034}$ | $0.438_{0.025}$ | |
| | NOT | $0.501_{0.068}$ | $0.285_{0.041}$ | |
| RobBERT | IND | $0.681_{0.009}$ | $0.862_{0.011}$ | $0.567_{0.006}$ |
| | GRP | $0.701_{0.005}$ | $0.666_{0.004}$ | |
| | OTH | $0.590_{0.033}$ | $0.448_{0.022}$ | |
| | NOT | $0.441_{0.013}$ | $0.244_{0.021}$ | |

Table 7: DALC v2.0: Target layer classification. Lower script numbers show standard deviations over 3 different runs. Best scores in bold.

Given the higher number of subclasses and the reduced number of messages useful for training the systems, target identification is more challenging. All systems outperform the dummy baseline, with varying degrees of performance. The first striking result is the (relatively) close performance of the SVM and the Bi-LSTM models, with a macro F1 delta of 0.004. While the Bi-LSTM has a better performance for the IND and GRP subclasses, the SVM obtains better results on the OTH and the NOT. The PTLMs confirm as the best systems and for this task BERTje outperforms RobBERT, with a macro-average F1 of 0.579.

Similarly to the offensive dimension, the distribution of the labels in the Train split clearly has an impact on the results of the trained systems (see Table 4). Thus, it is not surprising that all systems tend to overgeneralise the IND subclass since it is the most frequent one. When analysing the confusion matrices across all systems, it appears that the most confounded class is OTH. The class tends to be wrongly assigned to the IND and the GRP subclasses.

### 4.3 Distinguishing between Offensive and Abusive Language

| System | Class | Precision | Recall | Macro-F1 |
|---|---|---|---|---|
| SVM | OFF | 0.383 | 0.170 | 0.530 |
| | ABU | 0.570 | 0.410 | |
| | NOT | 0.820 | 0.941 | |
| Bi-LSTM | OFF | $0.451_{0.014}$ | $0.231_{0.096}$ | $0.607_{0.021}$ |
| | ABU | $0.596_{0.027}$ | $0.637_{0.042}$ | |
| | NOT | $0.883_{0.014}$ | $0.941_{0.022}$ | |
| BERTje | OFF | $0.339_{0.021}$ | $0.383_{0.020}$ | $0.599_{0.009}$ |
| | ABU | $0.600_{0.024}$ | $0.495_{0.009}$ | |
| | NOT | $0.891_{0.005}$ | $0.901_{0.013}$ | |
| RobBERT | OFF | $0.384_{0.015}$ | $0.359_{0.036}$ | $\mathbf{0.637}_{0.009}$ |
| | ABU | $0.625_{0.012}$ | $0.644_{0.018}$ | |
| | NOT | $0.903_{0.005}$ | $0.907_{0.011}$ | |

Table 8: DALC v2.0: Abusive *vs*. Offensive classification. Lower script numbers show standard deviations over 3 different runs. Best scores in bold.

In this section, we present a set of experiments that challenges systems to distinguish between three categories: whether a message is offensive but not abusive (OFF; see example 1), whether a message is abusive (ABU; see example 2), and whether a message is neither (NOT). The task is framed as a multi-class classification problem rather than as a multi-label classification one. This results in a slightly different distribution of the labels, namely in Train we have 1,391 (20.51%) messages marked as ABU, 1,086 (16.01%) messages marked as OFF, and 4,304 (63.47%) messages for NOT. The test split has 463 (14.15%) ABU messages, 404 (12.35%) OFF messages, and 2,403 (73.48%) messages marked as NOT. The distribution between the ABU and OFF classes is unbalanced in favour of the ABU class.

Results for these experiments are illustrated in Table 8. As the figures show, the imbalance of the classes in the Train split affects the performance of all systems, with the results for the ABU messages being better than those labelled as OFF, but worse than those labelled as NOT. RobBERT qualifies again as the best system followed by the Bi-LSTM, and with the SVM being the worst. The results for BERTje are comparable to those obtained for the offensive experiments in the binary setting (see Table 5). Across all systems, we observe a tendency to wrongly classify OFF as NOT, and ABU as OFF. Connecting this with our analysis of the top- keywords per class indicates that the systems trained

in this way heavily rely on superficial linguistic cues rather than grasping deeper and more heavily discriminating cues. In addition to this, when focusing on the combination of the explicitness layers and the ABU and OFF classes, we observe that in the Train split the majority of ABU messages (i.e, 62.25%) are marked as EXPLICIT, while this holds only for 49.81% of the OFF messages. It thus appears that with varying degrees all systems have identified a clear shortcut in these experiments whereby messages that are marked as EXPLICIT are then more often associated with the ABU class.

## 5 Error Analysis

We have conducted an error analysis for the offensive dimension and the offensive target layer since they represents the new annotations in the dataset. The error analysis has been conducted on the Dev set using the best performing system for each dimension.

**Offensive Language** For the offensive language dimension, we have used the predictions by RobBERT in the binary settings. The system wrongly classifies 179 messages, with the majority (101 messages) being OFF messages wrongly labelled as NOT. To gain better insights, we have classified all the errors into six categories:

- **criticism**: 13.40% of the errors are due to messages expressing some form of criticism; 75% of them are OFF wrongly labelled as NOT;
- **obfuscation**: only 3.35% of OFF messages wrongly labelled are due to obfuscation or abbreviation of profanities or slurs;
- **sarcasm/irony**: 8.93% of the errors are due to presence of irony or sarcasm; the majority (62.5%) concerns errors for the OFF subclass wrongly considered as NOT;
- **world knowledge**: 13.4% of the errors could have been correctly classified by means of some form of world knowledge;
- **gold errors**: 7.82% of the errors are due to potential annotation mistakes in the gold standard data;
- **bias**: this category comprises the largest amount of errors, 48.6% of the messages. 60.91% of the errors are False Positives for the OFF subclass containing identity terms (e.g. "gay"), names of political parties or politicians, or religious terms; the remainder of the messages are False Negatives for the OFF subclass containing stereotypes or being implicitly offensive.

**Target** For targets, 127 messages are wrongly classified. When analysing the confusion matrices across all systems, it appears that the most confounded class is OTH. The class tends to be wrongly assigned to the IND and the GRP subclasses. On the contrary, the errors for the NOT subclass are limited and they seem to be due to lack of training data.

The large part of the errors (31.49%) are due to different elements such as mixture of pronouns in the message (e.g., "jij" and "ze"), presence of collective nouns, or presence of a user's placeholder (i.e., MENTION) but no direct address in the text, and even mentions of concepts. The second largest block of errors, 23.62%, is due to the presence of multiple placeholders in the message, often happening in Twitter when replying to a long conversation but not necessarily addressing all the users involved. 18.11% of the errors could have been avoided by correctly processing the verb form. Given the larger amount of classes, 15.74% of the messages present some errors in the gold data - note, however, that these messages also include the errors in the gold standard for the offensive language dimension. Finally, 11.02% of the targets could have been correctly assigned if some form of commonsense knowledge was available to the system.

## 6 Related Work

The interest for the development of datasets and systems for the detection of abusive and offensive language phenomena has seen a steep growth in recent years. Different phenomena have been investigated including racism (Waseem and Hovy, 2016b; Davidson et al., 2017, 2019), hate speech (Alfina et al., 2017; Founta et al., 2018; Mishra et al., 2018; Basile et al., 2019), toxicity[10], verbal aggression (Kumar et al., 2018), and misogyny (Frenda et al., 2018; Pamungkas et al., 2020; Guest et al., 2021).

Offensive language, as we have detailed in § 2, is a more general and subjective phenomenon than abusive language. Founta et al. (2018) provides an

[10]The Toxic Comment Classification Challenge https://bit.ly/2QuHKD6

extensive analysis of the correlations between different phenomena and decide to collapse messages labelled as abusive, offensive and aggressive into a single category, namely abusive. Early attempts to annotate offensive language have been conducted in German as part of broader evaluation on hate speech (Wiegand et al., 2018). The SemEval 2019 Task 6: OffensEval (Zampieri et al., 2019c) has set up a common reference framework for the definition and the annotation of offensive language. The follow-up edition of the task (Zampieri et al., 2020) applied the original definition and annotation approach to four additional languages other than English, namely Turkish, Danish, Arabic, Greek. This corpus complements these annotation efforts with a further compatible dataset to fill a gap in the Dutch language resource panorama and to promote the advancement of multilingual approaches.

A different direction to the development of multilingual offensive language datasets has been presented with XHATE-99 (Glavaš et al., 2020). In this case, the authors have semi-automatically translated selected messages from three English datasets into five target languages (Albanian, Croatian, German, Russian, and Turkish). By working with translations, the authors have managed to better disentangle the impact of language versus domain shift in a transfer learning setting. As a matter of fact, the language alignments have ensured that losses observed in the cross-lingual setting are solely due to language shift rather than domain.

## 7 Conclusion and Future Work

This paper has presented DALC v2.0, a corpus for detecting offensive and abusive language in social media for Dutch. The corpus is composed of 11,292 messages manually annotated and it currently represents the largest available resource for these language phenomena in Dutch. Offensive language captures a more subjective dimension when compared to abusive language. For this reason, the data have been annotated in parallel by all annotators. We have applied a multi-layered annotation scheme targeting two key dimensions: the explicitness of the message and the presence of a potential target. For both annotation layers, the final labels have been assigned by means of majority voting. However, in the release of the corpus, we also distribute the disaggregated labels for both layers.

We have conducted a series of experiments by applying different algorithms. We have obtained the best results by using two monolingual PTLMs, namely RobBERT for the offensive dimension, and BERTje for the targets. For the offensive dimension, we have found that a Bi-LSTM architecture is very competitive when compared to the PTLMs also when using non-domain specific embeddings. We have also experimented on the ability of the models to distinguish between abusive and offensive language, obtaining promising results, showing that the distinction between offensive and abusive language is a more complex task than targeting each phenomenon individually.

Our error analysis has indicated limits of the systems and of the dataset. In particular, it seems that systems heavily rely on surface cues to assign a label to the message, showing a lack of "comprehension" of the content of the message and a high sensitivity to the distribution of the data in the training split.

Future work will focus on further testing the abilities of the dataset to train robust system by applying trained models to dynamic benchmark on the line of the HateCheck approach (Röttger et al., 2021). Furthermore, given the presence of multiple compatible corpora in different languages, we plan to explore the application of multilingual systems to address this task.

## Ethical Statement

**Dual use** DALC v2.0 and all the accompanying models are exposed to risks of dual use from malevolent agents. However, by making publicly available the resource and documenting the process behind its creation and the training of the models (including their limitations and errors), we may mitigate such risks.

**Misrepresentation** As the error analysis has shown (§ 5), even the best system is far from being perfect, with a relatively high number of False Positive for the OFF subclass. We thus recommend caution before deploying such a model without any additional human supervision.

**Privacy** Collection of data from Twitter's users has been conducted in compliance with Twitter's Terms of Service. Given the large amount of users that may be involved, we could not collect informed consent from each of them. To comply with this limitations, we have made publicly available only the tweet IDs. This will protect the users' rights to delete their messages or accounts. However, re-

leasing only IDs exposes DALC to fluctuations in terms of potentially available messages, thus making replicability of experiments and comparison with future work impossible. To obviate to this limitation, we make available another version of the corpus, Full Text. This version of the corpus allows users to access to the full text message of all 11,292 tweets. The Full Text dataset is released with a dedicated licence. In this case, we make available only the text, removing any information related to the time periods or seed users. We have also anonymised all users' mentions and external URLs. The licence explicitly prevents users to actively search for the text of the messages in any form. We deem these sufficient steps to protect users' privacy and rights to do research using internet material.

## References


Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *DP@AI*IA*.

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.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan Awadallah, Paul Bennett, and Weisheng Li. 2021. Say 'YES' to positivity: Detecting toxic language in workplace communications. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2017–2029, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: the Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66, Online. Association for Computational Linguistics.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France. European Language Resources Association.

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

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Simona Frenda, Ghanem Bilal, et al. 2018. Exploration of misogyny in spanish and english tweets. In *Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150, pages 260–267. Ceur Workshop Proceedings.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Mladen Karan and Jan Šnajder. 2019. Preemptive toxic language detection in Wikipedia comments using thread-level context. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 129–134, Florence, Italy. Association for Computational Linguistics.

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.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

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*, pages 3498–3508, Marseille, France. European Language Resources Association.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016a. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016b. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

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. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019c. 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 the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

## Appendix A: Replicability

Preprocessing All experiments have been conducted with common pre-processing steps, namely:

- lowercasing of all words
- all users' mentions have been substituted with a placeholder (MENTION);
- all URLs have been substituted with a with a placeholder (URL);
- all ordinal numbers have been replaced with a placeholder (NUMBER);
- emojis have been replaced with text (e.g. 🥺 → `:pleading_face:`) using Python `emoji` package;
- hashtag symbol has been removed from hasthtags (e.g. #kadiricinadalet → kadiricinadalet);
- extra blank spaces have been replaced with a single space;
- extra blank new lines have been removed.

**Models' hyperparameters** All hyperparamters used for the experiments are reported in Table 9.

| Model | Task | Hyperparm. | Value |
|---|---|---|---|
| SVM | Offensive Off. Target | n-gram range | 1–2 |
| | | character n-gram range | 3–5 |
| | | C | 1.0 |
| Bi-LSTM | Offensive | LSTM nodes | 32 |
| | | Hidden Layers | 0 |
| | | Embeddings | Coosto Word2Vec |
| | | Embedding dim. | 300 |
| | | Recurrent dropout | 0.1 |
| | | Batch size | 32 |
| | | Loss | categorical crossentropy |
| | | Layer activation | ReLu |
| | | Output layer activation | SoftMax |
| | | Fully connected layer size | 16 |
| | | Optimizer | Adam |
| | | Max. training epochs | 100 |
| | | Early stopping patience | 3 |
| Bi-LSTM | Off. Target | LSTM nodes | 50 |
| | | Hidden Layers | 0 |
| | | Embeddings | Coosto Word2Vec |
| | | Embedding dim. | 300 |
| | | Recurrent dropout | 0.1 |
| | | Batch size | 32 |
| | | Loss | categorical crossentropy |
| | | Layer activation | ReLu |
| | | Output layer activation | SoftMax |
| | | Fully connected layer size | 64 |
| | | Optimizer | Adam |
| | | Max. training epochs | 100 |
| | | Early stopping patience | 3 |
| BERTje RobBERT | Offensive | Learning rate | 4e-5 |
| | | Training Epochs | 5 |
| | | Optimzer | AdamW |
| | | Max sequence length | 123 |
| | | Batch size | 16 |
| | | Num. warmup steps | 2 |
| BERTje | Off. Target | Learning rate | 6e-5 |
| | | Training Epochs | 5 |
| | | Max seq. length | 123 |
| | | Batch size | 16 |
| | | Num. warmup steps | 2 |
| RobBERT | Off. Target | Learning rate | 5e-5 |
| | | Training Epochs | 5 |
| | | Max seq. length | 123 |
| | | Batch size | 16 |
| | | Num. warmup steps | 2 |

Table 9: Hyperparameters for each of the models used in the experiments.

## Appendix B: Supplementary Analyses

### B.1. Data Distribution

Table 10 illustrates the distribution of the data per topic/source across the Train, Dev, and Test split, respectively.

| Split | Data Source | Messages Included |
|---|---|---|
| Train | Paris Attack | 511 |
| | Dutch Parliament Election | 464 |
| | Protests/BLM | 1,255 |
| | Seed users | 2,539 |
| | June 2018 | 1,044 |
| | May 2019 | 1,004 |
| Dev | Paris Attack | 98 |
| | Dutch Parliament Election | 84 |
| | Protests/BLM | 237 |
| | Seed users | 436 |
| | June 2018 | 182 |
| | May 2019 | 168 |
| Test | *Intoch* Sinterklass | 240 |
| | April 2017 | 1,275 |
| | September 2019 | 1,100 |
| | Seed users | 655 |

Table 10: DALC v2.0: distribution of the sources across Train, Dev, and Test.

### B.2. Pairwise Inter-Annotator Agreement

Figures 1 to 12 illustrate the pairwise confusion matrix for each pair of annotators for the offensive explicitness layer and the offensive target layer. Note: for completeness, the target layer contains an extra subclass (NOT OFF) indicating cases where one annotator has marked the message as OFFENSIVE and, consequently, he has annotated also the target while the other has consider the message as not containing any offence.

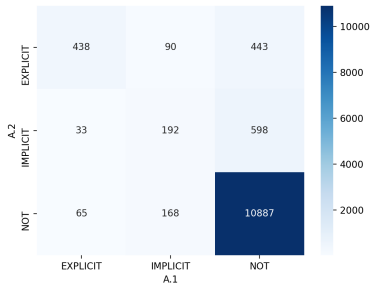


Figure 1: Explicitness Layer: A.1-A.2.

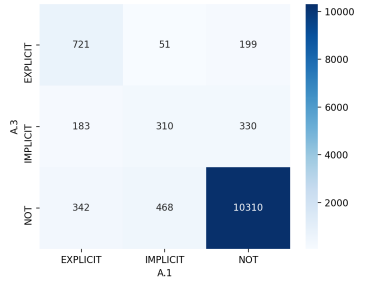


Figure 2: Explicitness Layer: A.1-A.3.

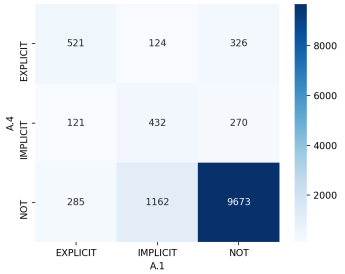


Figure 3: Explicitness Layer: A.1-A.4.

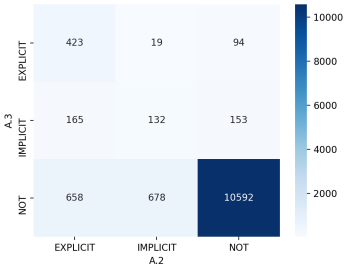


Figure 4: Explicitness Layer: A.2-A.3.

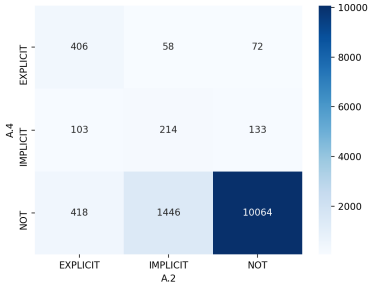


Figure 5: Explicitness Layer: A.2-A.4.

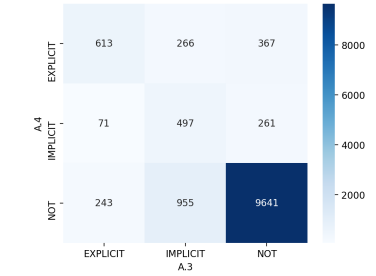


Figure 6: Explicitness Layer: A.3-A.4.

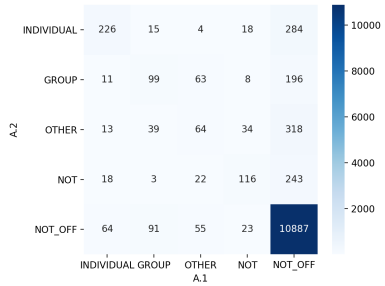


Figure 7: Target Layer: A.1-A.2.

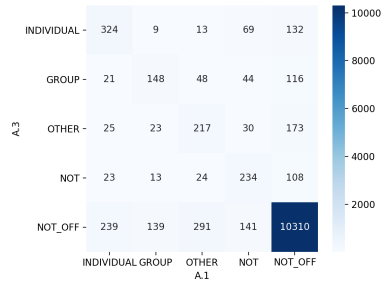


Figure 8: Target Layer: A.1-A.3.

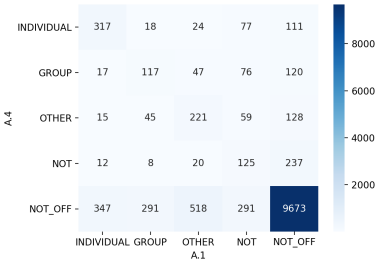


Figure 9: Target Layer: A.1-A.4.

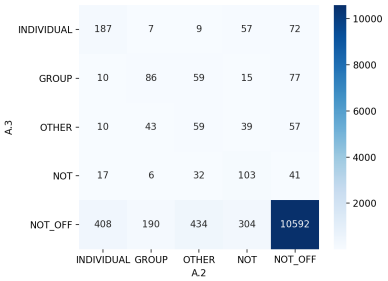


Figure 10: Target Layer: A.2-A.3.

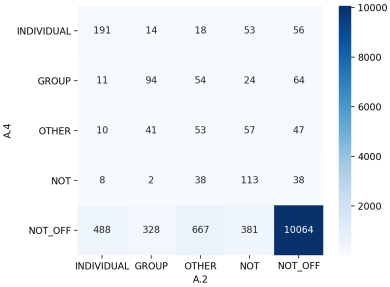


Figure 11: Target Layer: A.2-A.4.

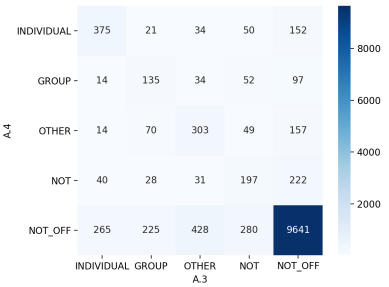


Figure 12: Target Layer: A.3-A.4.

### B.3 Keywords

Table 11 illustrates the keywords for the messages labeled as OFFENSIVE, ABUSIVE, and NOT OFFENSIVE. The keywords have been extracted using TF-IDF per language phenomenon rather than per subclass by collapsing the explicitness layers (i.e., offensive *vs.* abusive rather than abusive explicit *vs.* offensive explicit, and so forth).

### B.4 Error Analysis

Figure 13 illustrates the confusion matrix for the offensive language dimension (binary classification), while Figure 14 illustrates the confusion matrix for the target classification (offensive messages only)

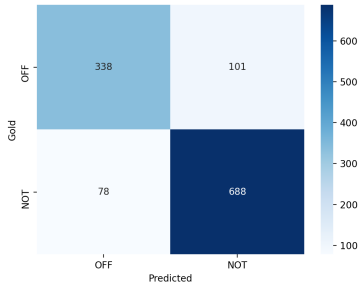


Figure 13: Confusion Matrix: Offensive Binary.

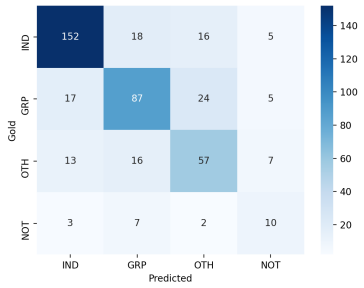


Figure 14: Confusion Matrix: Offensive Target.

| Train | | | | Test | | | |
|---|---|---|---|---|---|---|---|
| **OFF.** | **ABU.** | **NOT OFF.** | **NOT ABU.** | **OFF.** | **ABU.** | **NOT OFF.** | **NOT ABU.** |
| sod*****er | sod*****er | schaambeek | zand | onderbuikonzin | lelijk | peuzelen | amaai |
| klimaatwappie | lelijkerd | prop | klimaatwappie | 🤪 | 😃 | jonko | 🌳 |
| j**d | ontslaan | geboorteplaats | fokken | ha | arrogante | 😹 | ha |
| fari***r | kansloze | fokken | fari***r | och | 😂 | sad | 🥺 |
| lelijkerd | veenendaal | bong | bong | beesten | ma****ten | haarpijn | 💪🏻 |
| zeur | lijpo | opstandig | opstandig | 😃 | laffe | amaai | meter |
| veenendaal | huile | tier | tier | k*****stad | stap | uhhh | boekenweek geschenk |
| kansloze | flathead | webshops | vrolijk | ma****ten | k*****stad | hierzo | beesten |
| ontslaan | oogeruimd | datacenters | busje | catsuit | iek | zeldzame | mannelijkheid |
| huilie | sowieso | busje | wishlist | iek | k***** | leukkkk | geverfd |

Table 11: DALC v2.0: Top 10 keywords per target phenomenon in Train and Test. Explicitly offensive/abusive content have been masked with ∗