DALC: the Dutch Abusive Language Corpus

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman and Malvina Nissim
CLCG, University of Groningen
{t.caselli, m.nissim}@rug.nl
{a.j.schelhaas, m.i.weultjes }@student.rug.nl
{f.a.leistra,h.f.van.der.veen}@student.rug.nl
gerbentimmerman@protonmail.com

Abstract

As socially unacceptable language become pervasive in social media platforms, the need for automatic content moderation become more pressing. This contribution introduces the Dutch Abusive Language Corpus (DALC v1.0), a new dataset with tweets manually annotated for abusive language. The resource address a gap in language resources for Dutch and adopts a multi-layer annotation scheme modeling the explicitness and the target of the abusive messages. Baselines experiments on all annotation layers have been conducted, achieving a macro F1 score of 0.748 for binary classification of the explicitness layer and .489 for target classification.

1 Introduction

The growth of online user generated content poses challenges to manual content moderation efforts (Nobata et al., 2016). In a 2016 Eurobarometer survey, 75% of people who follow or participate in online discussions have witnessed or experienced abuse, threat, or hate speech.[1] The increasing polarization of online debates and conversations, together with the amount of associated toxic and abusive behaviors, call for some form of automatic content moderation. Currently, the mainstream approach in automatic content moderation uses reactive interventions, i.e., blocking or deleting 'bad' messages (Seering et al., 2019). There is an open debate on its efficacy (Chandrasekharan et al., 2017) and on the risks of perpetrating bias and discrimination (Sap et al., 2019). Alternative, less drastic, and more interactive methods have been proposed, such as the generation of counter-narratives (Chung et al., 2019). In either case, the first step towards full or semi-automatic moderation is the detection of potentially abusive language. Such step relies on language-specific resources to train tools to distinguish the "good" messages from the harmful ones. As a contribution in this direction, we have developed the Dutch Abusive Language Corpus, or DALC v1.0, a manually annotated corpus of tweets for abusive language detection in Dutch.[2] The resource is unique in the Dutch-speaking panorama because of the approach used to collect the data, the annotation guidelines, and the final data curation.

DALC is compatible with previous work on abusive language in other languages (Waseem and Hovy, 2016a; Papegnies et al., 2017; Founta et al., 2018; Mishra et al., 2018; Davidson et al., 2019; Poletto et al., 2020) but presents innovations both with respect to the application of the label "abusive" to messages and the adoption of a multi-layered annotation to distinguish the explicitness of the abusive message and its target (Waseem et al., 2017).

Our contributions can be summarized as follows:

- the promotion of **a bottom-up approach to collect potentially abusive messages** combining multiple strategies in an attempt to minimize biases that may be introduced by developers;
- the release of a manually **annotated corpus for abusive language detection in Dutch**, DALC v1.0;
- a series of **baseline experiments** using different architectures (i.e., a dictionary based approach, a Linear SVM, a Dutch transformer-based language model) showing the complexity of the task.

[1] https://what-europe-does-for-me.eu/en/portal/2/H19

[2] The corpus, the annotation guidelines, and the baselines models are publicly available at https://github.com/tommasoc80/DALC

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 54–66

2 Related Work

Previous work on abusive language phenomena and behaviors is extensive and varied. However, limitations exist and they mainly concentrate along three dimensions: (i) definitions; (ii) data sources and collection methods; and (iii) language diversity.

The development of automatic methods for detecting forms of abusive language has been rapid and has seen a boom of definitions, labels, and phenomena being investigated, including racism (Waseem and Hovy, 2016a; Davidson et al., 2017, 2019), hate speech (Alfina et al., 2017; Founta et al., 2018; Mishra et al., 2018; Basile et al., 2019), toxicity[3] and verbal aggression (Kumar et al., 2018), misogyny (Frenda et al., 2018; Pamungkas et al., 2020; Guest et al., 2021), and offensive language (Wiegand et al., 2018; Zampieri et al., 2019a; Rosenthal et al., 2020). Variations in definitions and in annotation guidelines have given rise to isolated datasets, limiting the portability of trained systems and reuse of resources (Swamy et al., 2019; Fortuna et al., 2021). Comprehensive frameworks that integrate and harmonize the variety of definitions and investigate the interactions across the annotated phenomena are still at early stages (Poletto et al., 2020). DALC v1.0 is compatible with existing definitions of abusive language and promotes a multi-layered annotation scheme compatible with previous initiatives, with a special attention to the reusability of datasets.

Collecting good representative data for abusive language is a challenging task. The majority of existing datasets focuses on messages from social media platforms, with Twitter being the most used Vidgen and Derczynski (2021). Unlike other language phenomena, e.g., named entities, abusive language is less widespread and cannot be easily captured by means of random sampling. Schematically, we identify three major methods to collect data: namely: (i) use of communities (Tulkens et al., 2016; Del Vigna et al., 2017; Merenda et al., 2018; Kennedy et al., 2018) which targets online communities known to be more likely to have abusive behaviors; (ii) use of keywords (Waseem and Hovy, 2016b; Alfina et al., 2017; Sanguinetti et al., 2018; ElSherief et al., 2018; Founta et al., 2018), where manually compiled lists of words corresponding either to potential targets (e.g, "women", "migrants", a.o.) or profanities are employed; (iii) use of seed users (Wiegand et al., 2018; Ribeiro et al., 2018), which collects messages from users that have been identified to post abusive texts via some heuristics. Each of these methods has advantages and disadvantages. For instance, the use of keywords may create denser datasets, but at the same time risks of developing biased data are very high (Wiegand et al., 2019). Furthermore, according to the specific platform used, some of the methods cannot be reliably applied. For instance, in a platform like Twitter targeting online communities is not trivial. Recently, refinements have been proposed to address limitations of each approach. In some cases controversial posts, videos or keywords are used as proxies for communities (Hammer, 2016; Graumans et al., 2019), in other cases hybrid approaches are proposed by combining keywords and seed users (Basile et al., 2019), others exploit platform pre-filtering functionalities (Zampieri et al., 2019a). DALC v1.0 integrates different bottom-up approaches to collect data providing a first cross-fertlization attempt across two social media platforms and paying attention to minimize the introduction of biases.

Vidgen and Derczynski (2021) provides a comprehensive survey covering 63 datasets all targeting a specific abusive phenomenon/behavior. The majority of them (25 datasets) is for English, with a long tail of other languages mostly belonging to the Indo-European family, although limited in their diversity. The lack of publicly available datasets for any Sino-Tibetan, Niger-Congo, or Afro-Asiatic languages is striking.

When it comes to abusive language datasets, Dutch is less-resourced. Notable previous work has been conducted by Tulkens et al. (2016) who developed a dataset and systems for detecting racist discourse in Dutch social media. DALC v1.0 differentiates because it is a "generic" resource for abusive language where all possible types of abusive phenomena are valid. This leaves room to refinement in the proposed corpus to investigate potential sub-types of abusive phenomena and their associated linguistic devices.

3 Data Collection

DALC v1.0 is based on a sample of a large ongoing collection of Twitter messages in Dutch at the University of Groningen (Tjong Kim Sang, 2011). For its construction, rather than focusing individually on any of the mentioned approaches,

[3]The Toxic Comment Clas- sification Challenge https://bit.ly/2QuHKD6

we propose a combination of three methods that only partially overlap with previous work.

Keyword extraction The first method is based on van Rosendaal et al. (2020), where keyword collection is refined via cross-fertilization between two social media platforms, namely Reddit and Twitter. Controversial posts from the subreddit `r/thenetherlands`, the biggest Reddit community in Dutch, at specific time periods are scraped, and a list of unigram keywords is extracted using TF-IDF. The top 50 unigrams are used as search terms in the corresponding time period in Twitter. This approach avoids the introduction of bias from the developers in the compilation of lists of search term. Obtaining them from controversial posts in Reddit may lead to denser samples of data in Twitter for abusive language phenomena.

We identified 8 different time periods between 2015 and 2020. We include both periods of time that may contain "historically significant events" (e.g., the Paris Attack in November 2015; the Dutch General Election in March 2017; the *Sinterklaas intocht* in December 20218; the Black Lives Matter protests after the killing of George Floyd in August 2020) and random time periods where no major events occurred, at least to our knowledge (e.g., April 2015; June 2018; May and September 2019). This results in a total of 12,884,560 retrieved tweets.

To ease the annotation process, we have sampled the retrieved data in smaller annotations batches. From each time period, we have generated samples of 10k messages composed as follows: 5k messages are randomly sampled, while the remaining 5k (non-overlapping) messages are extracted using two Dutch lexicon of potentially offensive/hateful terms, namely HADES (Tulkens et al., 2016) and HurtLex v1.2 (Bassignana et al., 2018). The actual manual annotation is performed on randomly extracted batches of 500 messages each. Table 1 provides an overview of the number of messages extracted per time period and the amount that has been manually annotated.

Geolocation The second method is inspired by previous work showing that in the Western areas of the (north hemisphere of the) world hatred messages tend to be more frequent in geographical areas that are economically depressed and where disenfranchised communities live (Medina et al., 2018; Gerstenfeld, 2017).[4] We use data from the Dutch *Centraal Bureau voor de Statistiek* (CBS) about unemployment to proxy such communities in the Netherlands, identifying two provinces: Zuid-Holland and Groningen.[5] We develop a set of heuristics, including the use of city names in these two provinces, to randomly collect messages from these areas. This is needed since the geolocation of the users is optional and does not have a fixed format. We managed to successfully extract 356,401 messages that can be reliably assigned to one of the two provinces. Similar to the keywords method, a sample of 5k messages is extracted using the lexicons and an additional 5k randomly. Four batches of 500 instances each have been manually annotated.

Seed users The last method uses seed users. We manually compile an ad-hoc list of 67 profanities, swearwords, and slurs by extending our lexicons. We then search for messages containing any of these elements in a ten-day window in December 2018 (namely 2018-11-12 – 2018-11-22). This results in a total of 3,105,833 messages. We rank each users according to the number of messages containing at least one of the target words. We select the top 50 users as seed users. We then extract for each of the selected user a maximum of 100 messages in a different time period, namely between May and June 2020, for a total of 5k tweets. Contrary to the other two methods, we directly created batches of 500 messages each for the manual annotation.

Since we are interested in original content, all messages sampled for the manual annotation do not contain retweets.

4 Annotation and Data Curation

DALC v1.0 has been manually annotated using internally developed guidelines. The guidelines provides the annotators with a definition of abusive language that refines proposals in previous work (Papegnies et al., 2017; Founta et al., 2018; Caselli et al., 2020). In particular, abusive language is defined as:

> *impolite, harsh, or hurtful language (that may contain profanities or vulgar language) that result in a debasement, harassment, threat, or*

[4] See also `https://bit.ly/3aDqoLd`.
[5] `https://bit.ly/2RPGSt5`

Time Period	Related Event	Extracted	Annotated
12-22 November 2015	Paris Attack	631,041	1,824
07-17 March 2017	Dutch Parliament Elections	265,256	1,824
April 2017	n/a	1,769,426	2,563
12-22 November 2018	*Intoch*[Arrival] Sinterklaas	377,007	526
June 2018	n/a	1,985,337	2,514
August 2020	Protests/BLM	733,985	3,128
May 20219	n/a	4,390,695	2,504
September 2019	n/a	2,731,813	2,504

Table 1: DALC v1.0 - Keywords: overview of the data collected and annotated

aggression of an individual or a (social) group, but not necessarily of an entity, an institution, an organization, or a concept.

Notably, this definition requires that an identifiable target must be present in the message to qualify as potentially abusive. This is a necessary requirement in our definition and it also helps us to discriminate abusive language from more generic phenomena like offensive language, forms of harsh criticism, and other socially unacceptable language phenomena. We have specifically introduced harsh criticism to restrict the application of the abusive language label. Indeed expressing heavy criticisms against an institution (e.g., the E.U. Commission, or a government) may result in inappropriate and offensive language but it does not entail being abusive. Exceptions, however, hold: cases of synecdoches where an institution, an entity, or a concept are used to attack the members of a social group are considered instances of abusive language.

Following Waseem et al. (2017) and Zampieri et al. (2019a) we perform a multi-layered annotation distinguishing the levels of **explicitness** of the abusive messages and the **targets**. Explicitness combines three factors: (i) the surface evidence of the message; (ii) the assumed intentions of the user (i.e., *is the message debasing someone?*); and (iii) its effects on the receiver(s) (i.e., *can the message be perceived as debasing by a targeted individual or a community?*). While the last two factors (intentions and effects) help to identify the abusiveness nature of the message, the surface forms is essential to distinguish overtly abusive messages from more subtle forms. A distinguishing criterion, in fact, is the presence of profanities, slurs, and offensive terms. We define three values:

- **Explicit** (EXP): A message is marked as explicit if it is interpreted as potentially abusive and if it contains a profanity or a slur;
- **Implicit** (IMP): A message is marked as implicit if it is interpreted as potentially abusive but it DOES NOT contain any identifiable profanity or slur;
- **Not abusive** (NOT): A message is marked as a not abusive if it is interpreted as lacking an intention of the user to debase/harass/threat a target and there is no debasing effect on the receiver. The mere presence of a profanity does not provide sufficient ground for annotating the message as abusive.

A further differentiating criteria is that all messages where the author debases or offends him-/herself (e.g., messages that contain the first person singular or plural pronoun) are considered as not abusive

The target layer makes explicit *to whom* the message is directed. We reuse the values and definitions from Zampieri et al. (2019a). In particular, we have:

- **Individual** (IND): any message that targets a person, being it named or unnamed, or a famous person;
- **Group** (GRP): any message that targets a group of people considered as a unity because of ethnicity, gender, political affiliation, religion, disabilities, or other common properties; and
- **Other** (OTH): any abusive message that addresses an organisation, an institution, or a concept. Instances of synecdoches are marked with this value rather than with group.

The annotation has been conducted in two phases. Phase 1 (March–May 2020) has seen five annotators, all bachelor students in Information Science. The students conducted the annotation of the data as part of their bachelor thesis project. Phase 2

(November–December 2020) has been conducted by one master student in Information Science with previous experience in this task. All annotators are native speakers of Dutch. More details are reported in the Data Statement A.

During Phase 1, we validate the annotation guidelines by means of a pairwise inter-annotator agreement (IAA) study on two independent subsets of 100 messages each. The first sample is obtained using the keyword method and the second using the geolocation. For the keywords sample, Cohen's kappa is 0.572 for the explicitness and 0.670 for the target. For the geolocation sample, the kappa for explicitness is comparable (0.522) although that for target is lower (0.466). The results are comparable previous work (Caselli et al., 2020) indicating substantial agreement. Cases of disagreement have been discussed between the annotators and resolved. The data used for the IAA has been integrated in DALC v1.0. No IAA has been computed for the messages collected using seed authors. In phase 2 we further expanded the initial data annotation.

The final corpus has been manually curated by one of the authors of this paper. The data curation phase focuses on the creation of the Train, Dev, Test splits in such a way that there is no overlap for time periods and, most importantly, users. Table 2 reports an overview of the data of each split and the number of annotated messages included.

Split	Data Source	Messages Included
Train	Paris Attack	1,051
	Dutch Parliament Election	996
	Protests/BLM	1,767
	Seed users	2,060 (+58)
Dev	Paris Attack	109
	Dutch Parliament Election	90
	Protests/BLM	156
	Seed users	196 (+6)
Test	*Intoch* Sinterklass	121
	April 2017	266
	June 2018	333
	May 2019	307
	September 2019	323
	Seed users	258 (+54)

Table 2: DALC v10: distribution of the sources across Train, Dev, Test. Numbers in parentheses indicate adjustments to prevent data overlap.

Overall, DALC v1.0 contains 8,156 tweets. In each split, the abusive messages correspond roughly to 1/3 of the messages. Maintaining this balance is not a trivial task. As it appears from Table 2, the different methods we used to collect the data results in different proportions of messages. Concerning the use of keywords, the combination of controversial keywords and historically relevant events works best, i.e., returns more densely annotated batches for the positive class, than the use of controversial keywords in random time periods. The geolocation method has been excluded due to the extremely low number of messages belonging to the positive class. Furthermore, a closer inspection revealed that these messages could be easily aggregated by their authors. We thus merge them with the seed users. Indeed, seed users results as the most successful method. Out of 5,000 messages collected, we managed to annotate and keep 2,520 of them. Excluding the merged users from the geolocation data, the Train/Dev split contains 38 unique users with an average of 54 messages each. On the other hand, the Test set contains 11 unique users and 23 messages each on average. To avoid any possibility of data overlap, we check that no message retrieved using the keyword method in one data split (e.g. Train) belongs to a seed users in a different data split (e.g., Test). For instance, we have found that 8 messages from the Paris Attack source have the same seed users of the test split. Only 118 messages were involved in these adjustments. In Table 2 we have marked these changes by showing the additional messages in parenthesis next to the seed users rows.

Table 3 shows DALC v1.0's label distribution per split. Overall, 1,879 messages have been annotated as containing forms of abusive language. The majority of them, 65.40%, has been classified as explicit. When focusing on the Train and Test splits, the most remarkable difference concerns the number of abusive messages labeled as implicit: 38.25% vs. 28.10%, respectively. As for the targets, the majority is realized by IND (55.18%) followed by GRP (34.64%) and OTH (10.69%). Interestingly, the distributions of the target is comparable to that of other datasets in other languages such as OLID (Zampieri et al., 2019a).

The average length of a message in DALC v1.0 is 25.94 words. Tokenization has been done by using the Dutch tokenizer available in SpaCy (Honnibal et al., 2020). In general, abusive messages are significantly[6] longer than the non abusive ones, with an average of 27.58 words compared to 22.77. While the differences between explicit and implicit

[6]Statistical test: Mann-Whitney Test; $p < 0.05$

Split	Explicitness			Target		
	EXP	IMP	NOT	IND	GRP	OTH
Train	699	443	4,564	634	399	109
Dev	72	38	439	62	33	15
Test	458	179	1,264	341	219	77
Total	1,229	660	6,267	1,037	651	201

Table 3: DALC v1.0: Distribution of Train, Dev, and Test splits for explicitness and target.

messages are basically non existent in the Train split, we observe significantly[7] longer implicit messages in the test data, with an average of 27.99 words against the 24.16 of the explicit ones. Standard deviations suggest that the length of the messages is skewed both in training and test for the three classes, with values ranging between 16.23 (EXPLICIT) and 13.71 (NOT) in Train, and 15.57 (IMPLICIT) and 14.03 (NOT) in Test.

We further investigate the composition of the DALC v1.0 by analysing the top 50 keywords per class between the Train and Test distributions by applying a TF-IDF approach. Table 4 illustrates a sample of the extracted keywords. As expected, clear instances of profanities and slurs appear in the EXP class. The IMP class does not present surface cues linked to specific lexical items. Actually, without knowing the class label and simply comparing the keywords, it is impossible to distinguish the IMP messages from those labeled as NOT. A further take-away of the keyword analysis is the lack of prevalence of any topic specific items (Wiegand et al., 2019). This, however, does not necessarily means that DALC v1.0 does not contain biases: indeed, the messages are not equally distributed across the time periods and seed users. On the other hand, our inspection of keywords has shown the lack of topic-specific keywords across the three classes.

We complete our analysis by exploring the similarities and differences between Train and Test splits. We investigate these aspects by means of two metrics: the Jensen-Shannon (J-S) divergence and the Out-of-Vocabulary rate (OOV). The J-S divergence assesses the similarity between two probability distributions, q and r. On the other hand, the OOV rate helps in assessing the differences between the Train and Test splits as it highlights the percentage of unknown tokens. We obtain a J-S score of 73% and an OOV rate of 64.6%. This means that while the Train and Test distributions are quite similar to each other, the gap in terms of lexical items between the two is quite large. This supports the validity of our data curation approach where overlap between Training and Test split is not allowed.

[7]Statistical test: Mann-Whitney Test; $p < 0.05$

5 Baselines

We present a set of baseline experiments that accompany the release of DALC v1.0 for the two annotation layers. For the explicitness layer, we first experiment a simplified setting by framing the problem as a binary classification task. In this setting the distinction between EXP and IMP labels is collapsed into a new unique value for all abusive messages (i.e., ABU). The follow-up experiment, on the other hand, maintains the fine-grained distinction in the three classes (i.e., EXP *vs.* IMP *vs.* NOT).

For the target layer no simplification of the labels is possible since each oh them identified a specific referent. Thus, target experiments preserve the original three labels (i.e., IND *vs.* GRP *vs.* OTH).

In all experiments we adopt a common preprocessing of the data. All user mentions and links to external web pages are replaced with dedicated placeholders symbols, respectively USER and URL. Emojis are replaced with their corresponding text using the `emoji` package. Hashtags symbols have been removed but we have not split hashtags composed by multiple words in separate tokens.

The models are trained on the Train split and evaluated on the held out Test set. The Dev split is used for parameter tuning. As illustrated in Table 3, the distributions of the labels in the classes for both annotation layers is unbalanced. We thus evaluate and compare our models using the macro-average F1. Furthermore, we report Precision and Recall for each class. In each annotation layer, we compare the models to a majority class baseline (MFC).

Abusive *vs.* Not Abusive This binary setting allows to test the classification abilities of different architectures in a simplified setting. It also provides evidence of the complexity of the task given the lack of overlap across time periods and seed users between Train and Test.

We experimented with three models. The first is a dictionary-based approach. The approach is very simple: given a reference dictionary of profanities,

Train			Test		
EXP	**IMP**	**NOT**	**EXP**	**IMP**	**NOT**
sod*****er	kansloze	schaambeek	spoort	ha	fashion******sisters
lelijkerd	huilie	onderbuikonzin	huilie	stap	boekenkast
ontslaan	nakijken	maradonny	arrogante	iek	hierzo
ha	lijk	jood	mal****	schaapskleren	tuu
st*****	slimste	haarpijn	la**e	aantonen	kúnnen
sowieso	dissel	geboorteplaats	blind	fuhrer	ouuuuttttttt
f***head	stem	huurauto	k*****stad	trapt	och
paras*****	binnenlaten	spinnend	gebruik	dommie	penny
k*t	jaily	leukkkk	k****r	verhaal	nieuwjaar
uitgemergelde	gestraft	afloopt	gebruikte	rollen	supermooi

Table 4: DALC v1.0: Top 10 keywords per class in Train and Test. Explicitly offensive/abusive content have been masked with ∗

abusive terms, slurs in Dutch, if any message contains one or more of the terms in the dictionary, then it is labeled as abusive (i.e., ABU). We have created a new lexicon of 847 potentially abusive term by refining the original Dutch entries in HurtLex v1.2 (Bassignana et al., 2018) and integrating the list with 256 culturally specific terms. In particular, most of the new entries concerned names of diseases (e.g., *kanker* [cancer]) that in Dutch are commonly used to debase or harass people. Each term has also been classified as belonging to one of two macro-categories, namely "negative stereotypes" (representing 45.1% of the entries) and "hate words and slurs beyond stereotypes" (including the remaining 54.9% of the entries). The list has not been extended with additional terms from the EXP messages in the Train split of DALC v1.0.

The second model is a Linear Support Vector Machine (SVM) model. We used the available implementation in `scikit-learn` (Pedregosa et al., 2011). Each message is represented by a TF-IDF vector combining word and character ngrams. We run a grid search to find the best ngram combination and parameter tuning. The final configuration uses bigrams and character ngrams in the range 3–5, a C values of 1.0, and removal of stopwords.

The last model is based on a monolingual Dutch pre-trained language model, BERTje (de Vries et al., 2019), available through the `Hugging Face` transformers library.[8] The model is fine tuned for five epochs, with a standard learning rate of 2e-5, AdamW optimizer (with `eps` equals to 1e-8), and batch size of 32.

The results of the experiments are reported in Table 5. All models outperform the MFC baseline,

System	Class	Precision	Recall	Macro-F1
MFC	ABU	0	0	0.399
	NOT	0.664	1.0	
Dictionary	ABU	0.716	0.433	0.685
	NOT	0.761	0.913	
SVM	ABU	0.858	0.323	0.655
	NOT	0.740	0.973	
BERTje	ABU	0.850	0.500	**0.748**
	NOT	0.791	0.955	

Table 5: DALC v1.0: Binary classification. Best scores in bold.

however, the task proves to be challenging. BERTje obtains by far the best results with a macro F1 of 0.748. Quite surprisingly, the Dictionary model has more competitive results than the SVM. The gap in scores can be explained by the large OOV rate between Train and Test split. SVMs usually are very competitive models but one of their shortcoming is the heavy dependence on a shared vocabulary between training and test distributions. A further element of attention is the low Recall that all models have for the positive class. While this behavior is expected due to the unbalanced distributions of the classes, we claim that this is an additional cue with respect to the data distribution of DALC v1.0.

To further confirm this intuition, we ran an additional set of experiments on a different data split. We maintained exactly the same amount of messages and distribution in the classes. On the other hand, we did allow for overlap across time periods and seed users. The OOV rate between Train and Test splits drops to 55.21%. At the same time, by re-running the experiments with the same settings for all models, the Dictionary model is the weakest, with a macro F1 of 0.680. On the other hand, the Linear SVM achieves competitive results when

[8] https://huggingface.co/GroNLP/bert-base-dutch-cased

compared to BERTje (macro F1 of 0.749 *vs.* 0.786, respectively).

Explicit *vs.* Implicit For the fine-grained classification, we compare only two architectures, the linear SVM and BERTje. As already stated, this is a more challenging setting namely due to a combination of factors such as the number of classes, the data distributions, and the class imbalance. The grid search for the SVM confirmed the same settings as for the binary experiment. We re-used the same settings for BERTje. Table 6 summarizes the results.

System	Class	Precision	Recall	Macro-F1
SVM	EXP	0.805	0.270	0.433
	IMP	0.461	0.033	
	NOT	0.719	0.986	
BERTje	EXP	0.759	0.447	**0.561**
	IMP	0.373	0.189	
	NOT	0.790	0.962	

Table 6: DALC v1.0: Explicitness classification. Best scores in bold.

BERTje is again the model achieving the best results, with a macro F1 of 0.561. Both models, however, struggle to correctly classify the IMP messages correctly. Observing the distribution of the errors for this class, both models tend to be misclassify the IMP messages as NOT, further confirming the observations from the keyword analysis. The increased granularity of the classes has a negative impact on the performance of the SVM also for the EXP messages. While Precision is comparable to the binary setting, the system largely suffers in Recall. This is not the case for BERTje, where Precision and Recall for the EXP and NOT classes are in line with the results of the binary setting. On the other hand, the results for the IMP classes are encouraging, although far from being satisfying.

Target Classification Models for this task are trained to distinguish among the three target classes: individuals (IND), group(s) (GRP), and other (OTH). For this experiment the amount of training data is smaller since only abusive messages have been used. We experimented with two models' architectures only: a Linear SVM and BERTje. The grid search for the SVM results in the same settings of for the explicitness layer. When it comes to BERTje, we apply the same settings: fine tuninig for five epochs, standard learning rate of 2e-5, AdamW optimizer (with `eps` equals to 1e-8), and batch size of 32. Results are reported in Table 7.

System	Class	Precision	Recall	Macro-F1
MFC	IND	0.535	1.00	0.232
	GRP	0	0	
	OTH	0	0	
SVM	IND	0.693	0.897	0.492
	GRP	0.698	0.602	
	OTH	0.285	0.026	
BERTje	IND	0.745	0.841	**0.498**
	GRP	0.634	0.730	
	OTH	1.0	0.012	

Table 7: DALC v1.0: Target classification. Best scores in bold.

Both models clearly outperform the MFC baseline. However, the gap between the two is very small differently than for the explicitness layer. Both models struggle with the OTH class. The lower amount of training examples for this class (only 109) is a factor the impact the performance. However, this class is also less homogeneous than the others. It contains different types of targets such as institutions, events, and entities that do not fit in the other two classes. When focusing on the results for the IND and OTH classes, it seems that models suffer less when compared to the explicitness layer. This suggest that there may be a reduced variation in the expressions of the targets. Finally, the results are in line with previous work on target detection in English (Zampieri et al., 2019b).

6 Conclusions and Future Work

This paper introduces DALC v1.0, the first "generic" resource for abusive language detection in Dutch. DALC v1.0 contains more thn 8k Twitter messages manually labeled using a multi-layer annotation scheme targeting the explicitness of the message and the targets. A further peculiarity of the dataset is the complete lack of overlap for time periods and users between Train and Test splits, making the task more challenging.

The combination of multiple data collection strategies aims at promoting new bottom-up approaches less prone to additional biases in the data other than those from the manual labeling.

DALC v1.0 adopts a definition of abusive language and an annotation philosophy compatible with previous work, paying attention to promote interoperability across language resources, languages, and abusive language phenomena.

The baseline experiments and systems that have been developed further indicate the challenges of this dataset. The best results are obtained with a fine tuned transformer-based pre-trained language model, BERTje. Fine-grained distinction for the explicitness layer is particularly difficult for implicitly abusive messages. Furthermore, target classification is a challenging task, with overall macro-F1 below 0.50.

Future work will focus on an in-depth investigation of the errors to identify easy and complex cases.

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.

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.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

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.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

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*, volume 11.

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)*, pages 86–95.

Daria Denti and Alessandra Faggian. 2019. In *5th International Conference on Hate Studies*, Spokane, USA. non-archival. [link].

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

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.

Phyllis B Gerstenfeld. 2017. *Hate crimes: Causes, controls, and controversies*. Sage Publications.

Leon Graumans, Roy David, and Tommaso Caselli. 2019. Twitter-based polarised embeddings for abusive language detection. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE.

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.

Hugo Lewi Hammer. 2016. Automatic detection of hateful comments in online discussion. In *International Conference on Industrial Networks and Intelligent Systems*, pages 164–173. Springer.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, and et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech.

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.

Richard M Medina, Emily Nicolosi, Simon Brewer, and Andrew M Linke. 2018. Geographies of organized hate in america: a regional analysis. *Annals of the American Association of Geographers*, 108(4):1006–1021.

Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. Source-driven representations for hate speech detection. In *CLiC-it*.

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.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

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.

Etienne Papegnies, Vincent Labatut, Richard Dufour, and Georges Linarès. 2017. Detection of abusive messages in an on-line community. In *CORIA*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

Juliet van Rosendaal, Tommaso Caselli, and Malvina Nissim. 2020. Lower bias, higher density abusive language datasets: A recipe. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 14–19, Marseille, France. European Language Resources Association (ELRA).

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. 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 Language Resources Association (ELRA).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950.

Erik Tjong Kim Sang. 2011. Het gebruik van twitter voor taalkundig onderzoek. *TABU: Bulletin voor Taalwetenschap*, 39(1/2):62–72.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. The automated detection of racist discourse in dutch social media. *Computational Linguistics in the Netherlands Journal*, 6:3–20.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15.

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*.

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.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. 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 NAACL*.

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.

A Data Statement

Data set name: Dutch Abusive Language Corpus (DALC) v1.0
Data will be released to the public in compliance with GDPR and Twitter's Terms of Service.

A. CURATION RATIONALE The corpus is composed by tweets in Dutch extracted using different strategies and covering different time windows.

- **Keywords**: we have used a cross-platform approach to identify relevant keywords and reduce bias that may be introduced in manual selection of the data. We first identified a time window in Reddit, extracted all posts that received a controversial label. We then identified keywords (unigram) and retained the top 50 keywords per time window. We then used the keywords to extract tweets in corresponding periods. For each time period, we selected a sample 5,000 messages using two dictionaries containing know profanities in Dutch. An additional 5,000 messages are randomly selected. The messages are then re-shuffled and annotated.

- **Geolocation**: following Denti and Faggian (2019) that show the existence of a correlation between hateful messages and disenfranchised and economic poor areas, we selected two geo-graphical areas (Zuid-Holland and Groningen) that according to a 2015 study by the Ducth Buraeu of Statistics (CBS) have the highest unemployement rates of the country. We collected 706,044 tweets posted by users whose location was set to the two target areas. The amount of messages was further filtered by removing noise (i.e., messages containing URLs), dropping to 356,401 tweets. Similarly to the keywords approach, we further filtered 2,500 messages using one profanity dictionary and collected an additional 2,500 randomly.

- **Authors**: we looked for seed users, i.e., users that are likely to post/use abusive language in their tweets. We created an ad-hoc list of 67 profanities, swearwords, and slurs and then searched for messages containing any of these elements in a ten-day window in December 2018 (namely 2018-11-12 – 2018-11-22), corresponding to a moment of heated debate in the country about Zwarte Piet. We collected an initial amount of 3,105,833 tweets. We

then selected as seed users the top 15, i.e., the top 15 users who most frequently use in their messages any of the 67 keywords. For each of them we further collected a maximum of 100 tweets randomly, summing up to a total of 1390 tweets

- **Dictionaries used**: HADES (Tulkens et al., 2016); HurtLex v1.2 (Bassignana et al., 2018)

Time periods (DD-MM-YYYY):

1 12-11-2015/22-11-2015 (November 2015 Paris attacks);

2 07-03-2017/17-03-2017 (2017 Dutch general election);

3 12-11-2018/22-11-2018 (Intocht Sinterklaas 2018);

4 2020-08 (protests in solidarity with the Black Lives Matter movement);

5 2015-04;

6 2018-06

7 2019-05

8 2019-09

B. LANGUAGE VARIETY/VARIETIES
BCP-47 language tag: `nl`
Language variety description: Netherlands and Belgium (Vlaams)

C. SPEAKER DEMOGRAPHIC N/A

D. ANNOTATOR DEMOGRAPHIC

- **Annotator #1**: Age: 21; Gender: female; Race/ethnicity: caucasian; Native language: Dutch; Socioeconomic status:n/a Training in linguistics/other relevant discipline: BA in Information science
- **Annotator #2**: Age: 21; Gender: male; Race/ethnicity: caucasian; Native language: Dutch; Socioeconomic status:n/a Training in linguistics/other relevant discipline: BA in Information science
- **Annotator #3**: Age: 21; Gender: male; Race/ethnicity: caucasian; Native language: Dutch; Socioeconomic status:n/a Training in linguistics/other relevant discipline: BA in Information science
- **Annotator #4**: Age: 21; Gender: male; Race/ethnicity: caucasian; Native language: Dutch; Socioeconomic status:n/a Training in linguistics/other relevant discipline: BA in Information science
- **Annotator #5**: Age: 23; Gender: male; Race/ethnicity: caucasian; Native language: Dutch; Socioeconomic status:n/a Training in linguistics/other relevant discipline: BA in Information science
- **Annotator #6**: Age: 24; Gender: male; Race/ethnicity: caucasian; Native language: Dutch; Socioeconomic status:n/a Training in linguistics/other relevant discipline: MA in Information science

E. SPEECH SITUATION N/A

F. TEXT CHARACTERISTICS Twitter messages; short messages of max. 280 characters; the original messages may contain multimedia materials, external URL links, and mentions of other users. For all experiments, URLs and users' mentions have been anonymized. Time period of collection illustrated in §A **Curation Rationale**.

G. RECORDING QUALITY N/A

Data Statements are from the University of Washington. Contact: datastatements@uw.edu. The markdown Data Statement we used is from June 4th, 2020. The Data Statement template is based on worksheets distributed at the 2020 LREC workshop on Data Statements, by Emily M. Bender, Batya Friedman, and Angelina McMillan-Major.

B Ethical considerations

Dual use DALC v1.0 and the accompanying models are exposed to risks of dual use from malevolent agents. However, we think that by making publicly available the resource, documenting the process behind its creation and the models, we may mitigate such risks.

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, releasing 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, DALC Full Text. This version of the corpus allows users to access to the full text message of all 8,156 tweets. The DALC Full Text dataset is released with a BY-NC 4.0 licence. In this case, we make available only the text, removing any information related to the time periods or seed users. We have also anonymized all users' mentions and external URLs. The CC licence is extended with further restrictions explicitly preventing 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.