Online Social Networks and Media 30 (2022) 100210

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media

Wenjie Yin *, Arkaitz Zubiaga

School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London, E1 4NS, United Kingdom

ARTICLE INFO

Keywords:
Hate speech
Abusive language
Social media
Text classification

ABSTRACT

While social media offers freedom of self-expression, abusive language carry significant negative social impact. Driven by the importance of the issue, research in the automated detection of abusive language has witnessed growth and improvement. However, these detection models display a reliance on strongly indicative keywords, such as slurs and profanity. This means that they can falsely (1a) miss abuse without such keywords or (1b) flag non-abuse with such keywords, and that (2) they perform poorly on unseen data. Despite the recognition of these problems, gaps and inconsistencies remain in the literature. In this study, we analyse the impact of keywords from dataset construction to model behaviour in detail, with a focus on how models make mistakes on (1a) and (1b), and how (1a) and (1b) interact with (2). Through the analysis, we provide suggestions for future research to address all three problems.

1. Introduction

While social media provides a platform for all users to freely express themselves, cases of **offensive language** are not rare and can severely impact user experience and even the civility of a community [4]. When such offence is intentional or targeted, it is further considered **abuse** [5]. **Hate speech**, which is speech that directly attacks or promotes hate towards a group or an individual member based on their actual or perceived aspects of identity, such as ethnicity, religion, and sexual orientation [6–9], is a sub-category of abuse that is identity-oriented [10] and particularly harmful for silencing marginalised groups.[1]

The problems caused by the posting of abusive and offensive language by social media users has increased the need both for analysing the phenomenon [11–13] and for developing automated means for moderation of such content [14]. While methods for automated detection of abusive language are improving, the detection models share shortcomings that limit their practicality, due to their reliance on prominent lexical features, i.e. indicative keywords such as slurs and profanity. The effect of this reliance is two-fold. On one hand, they struggle with implicit expressions without such features and non-abusive speech with such features; on the other, the models have limited generalisability, i.e. models trained on one abusive language dataset do not perform well on other, unseen datasets [15], where prominent features can be different. The two problems go hand-in-hand.

In this paper, we focus particularly on how the presence of keywords categorised as profanity or slurs can impact the tendency of models to label content as abusive, offensive or normal. To study this, we assess how decisions made in the creation of these datasets, including data sampling and annotation, can have an impact on the keyword and class distributions in the datasets, which in turn impacts model behaviour, both with detection and generalisation across datasets.

1.1. Problem statement and research questions

Abusive language detection methods are often built upon lexicon-based sampling methods, causing them to struggle with indirect forms of expressions [16], and are easily misled by prominent words [17].

Challenges posed by keywords are two-fold: (1) non-abusive use of profanity and slurs, (2) implicit abuse without slurs or profanity. Implicit abuse is the most commonly mentioned cause of false negatives in error analysis [18–21]. On the other hand, even state-of-the-art models, such as Perspective API,[2] tend to flag non-abusive use of slurs [10,22], which is common in certain language styles, such as African American English dialect [23]. These two challenges hurt the practicality of applying models as real-world moderation tools [24]: whilst cases of (1) may leave nonetheless-harmful content unaffected by moderation, cases of (2) may amplify the harm against minority groups instead of mitigating such harm as intended [25].

* Corresponding author.
E-mail address: w.yin@qmul.ac.uk (W. Yin).

[1] For a more elaborate comparison between similar concepts, see [1–3].

[2] https://www.perspectiveapi.com/.

https://doi.org/10.1016/j.osnem.2022.100210
Received 12 August 2021; Received in revised form 7 April 2022; Accepted 16 April 2022
Available online 23 May 2022

In addition to this, limited generalisability, demonstrated by the model performance drop when applied on unseen datasets, severely hurts the practical value of these automated detection models [15,17,26–29]. Models that suffer from this problem range from simpler classical machine learning models to recent state-of-the-art neural models [22,30,31]. What dataset factors contribute to better generalisability is thus an important and pressing issue.

Despite the recognition of the problems, there still exist considerable gaps in the understanding of them. On the relationship between keywords and model behaviour, most findings have been high-level, and the mechanism of the effect remains unclear. Dataset comparisons in generalisation studies mainly compared whole datasets and adopted a binary classification task, which oversimplify the problem and create inconsistent observations. Furthermore, the link between these two important and related issues is missing, such as the difference in, and factors that contribute to, generalisation on implicit and explicit abuse, as well as instances with misleading keywords that are not abusive.

Research questions. In this study, we aim to address the above-mentioned challenges brought by keywords (profanity and slurs), whilst distinguishing offensive and abusive (including hate speech).

In particular, we ask:

- **RQ1:** How do approaches in dataset construction lead to different patterns of slurs and profanity presence in the dataset?
- **RQ2:** How does such keyword presence in turn affect the detection model behaviour?
- **RQ3:** How does this effect differ when it comes to model generalisation?

To answer these questions, we make the following contributions:

- We analyse the presence of keywords and the association between keywords and classes, comparing them across three commonly-studied datasets, and relating back to dataset construction in the analysis.
- We then investigate the effect of such keyword presence on model behaviour, using a well-studied and strong model as a representative example.
- We perform in- and cross-dataset experiments to compare the effects on detection and generalisation.

As well as for dataset creation (involving collection, sampling and annotation) and for building automated detection models, our findings have important implications to enable large-scale analyses of behavioural and linguistic patterns linked to abusive and offensive language that incorporate nuanced abusive or non-abusive examples where the presence – or lack thereof – of keywords entails the opposite of the expected meaning.

2. Background

2.1. Implicit expressions and non-abusive use of slurs and profanity

Implicit expressions of abusive language and non-abusive use of slurs and profanity are the two sides of the same coin. As abusive language is associated with slurs and profanity, both its absence and presence can be misleading to the detection model — when the speech is abusive and not so respectively.

Both challenges have been recognised in abusive language detection research, but usually separately and rather on a high level.

On one side, implicit abuse can be expressed through stereotypes, sarcasm, irony, humour, and metaphor [32–34]. It has been proposed that abusive language should be systematically classified into explicit and implicit [35]. Several subsequent studies have identified nuanced, implicit expression as a particularly important challenge in abusive language detection for future research to address [24,27,36]. Addressing implicit abuse is especially necessary for model explainability [33]. Manifested in the model behaviour, such implicit expressions are the most commonly mentioned cause of false negatives [21,28,37].

The definition of explicitness and implicitness in the context of abusive language detection has been usually based on whether keywords – slurs or profanity – are present [5,35,38], although this definition is not equal to their linguistic or social definitions, as implicitness and explicitness are highly subjective notions [39].

The other side of the same coin – non-abusive keyword use – is equally important for abusive language detection. Profanity can be used for stylistic purposes or emphasis [40]; some slurs have been reclaimed by targeted groups [10], and is common in African American English (AAE) dialect [23]. A model that falsely flags these instances as abuse could discriminate against minority groups that the model is intended to protect [25]. Indeed, non-abusive slur and profanity use is a common cause of false positives [36]; simply by adding the f-word can make positive statements carry a high "toxicity" score [17].

Despite its importance in the abusive language detection task, it was only integrated into the taxonomy of abuse-surrounding phenomena recently [10].

Annotated implicit abuse and non-hateful slur use are very limited [5,10]. Motivated by the lack of suitable data, recent studies have attempted to surface implicit abuse in unlabelled data, using initial samples with keywords [41] or through identifying "influencial" training instances [16]. By providing more training instances with implicit abuse, both studies' approaches benefited model performance.

However, when it comes to the actual *effect* of keywords, findings have been limited and largely high-level. In the error analysis of classification studies, it is generally considered that using keyword search and biased sampling contribute to a dataset containing more explicit expressions [5,26–28], which can lead to a higher recall of the positive class [42] but overall lower F1 [43] in binary classification. Many questions still remain: how annotation plays a role in the process, how the resulting keyword presence in the data impact the detection of implicit and explicit expressions differently, how implicitness interact with other factors, etc.

2.2. Generalisability in abusive language detection

Generalisation refers to how well a machine learning model performs on previously unseen data, which assesses its ability of capturing the real relationship between features and expected outputs [44].

Recently, the generalisability of abusive language detection models have received increasing attention. Cross-dataset testing – evaluating models on a different dataset than the one(s) they were trained on – has revealed that model performance is severely over-estimated when evaluated only on the set-aside "test set" of the same dataset. A recent study has summarised the impact of model and dataset factors on cross-dataset performance [29]. Across different cross-dataset studies, the macro-averaged F1 scores most commonly drop by between 20 and 40 points when the model is applied on a different dataset [15,17,26–28]. Models that suffer from a significant performance drop when applied cross-dataset range from classical machine learning models to the state-of-the-art neural models [22,30,31].

Nonetheless, fine-tuning large language models that have been pre-trained extensively, with BERT [31] as a representative example, seems to be relatively more generalisable [28,45].

Studies on factors that affect generalisation have mostly compared entire datasets. Similar datasets generalise better to each other, with the similarity attributed to search terms for sampling [27], topics [45,46], and class label definitions [28,45]. When it comes to what produces high-quality generalisable data in general, wider coverage of abuse phenomena, including topics, is believed to be beneficial [46,47]. A more broadly defined positive class is also perceived as more generalisable to a narrowly defined one than the other way round in binary classification [28,46]. Surprisingly, the effect of the data size is rather limited, compared to other factors [28,46]. Inconsistency exists

in the literature, when it comes to the effect of class proportions: while some observed that a larger proportion of the positive class makes a dataset more generalisable [27,48], others found the opposite [26] or could not confirm its effect [28]. One explanation is that the balance between true positive and true negative is not reflected in the overall performance [46].

Findings on the relationship between the presence of keywords and generalisation are limited to a few isolated observations. Some hold that containing more explicit expressions makes a dataset more generalisable [26]; others observed that most non-offensive instances with keywords in [7] were mislabelled as offensive by a model trained on [49], and attributed this to the high frequency of keywords in the former [27]. Thus, to fill in the gaps, systematic investigations on this topic is needed.

2.3. Offensive language vs. abuse

There exists a key difference between offensive and abusive language: abusive language has a strong component of intentionality; the definition of offensiveness has more emphasis on lexical content and the receiver's emotional response. Hate speech, with a strong intention to "direct attack or promote hate", thus falls under abusive language. Experts can distinguish abusive and offensive, both conceptually and in practice during annotation [5]. However, both are used as umbrella terms for harmful content in the context of automatic detection studies, and these two terms are often confused, especially by crowd annotators. In a large-scale crowd-annotated dataset [8], the annotations for "abusive" and "offensive" were so similar that the two class labels were combined in the end.

This distinction carries significant practical value. On offensive language, a purely lexicon-based detection model can achieve competitive performance [50], while abuse is captured by lexical features less [5]. Thus, distinguishing abuse and offensive language can reveal more insights into implicit expressions.

In summary, despite their importance, inconsistency and unanswered questions still largely remain, in both the challenges posed by keywords and model generalisability. Furthermore, existing studies on these two topics have always worked with binary classification without distinguishing offensive language and abuse, limiting the practical value.

Thus, our study addresses the gap in the literature by providing an all-round and in-depth analysis of the challenge posed by keywords — unifying the two sides of the same coin and following the entire chain of effect: from sampling and annotation to the data, then finally to model behaviour. By extending the above analysis from in-dataset detection to cross-dataset generalisation, we offer a new perspective of looking at model generalisation. We include but distinguish both abuse and offensive language, enabling insights into fine-grained model behaviour and better understanding of implicitness and keyword use. As a result, we clarify confusions in the interacting factors seen in previous studies.

3. Materials

3.1. Definitions

Table 1 summarises the main concepts used in this study. We consider three main types of nature of speech — abuse, offensive, and normal, separating abuse and offensive with intentionality (Section 2.3) and including hate speech as a special case of identity-oriented abuse [10].

Following the definitions of previous studies [5,35,38], whether an instance is implicit or explicit is then dependent on the presence of keywords, which can be any slur or profanity: if an instance of speech, whilst being abuse or offensive, contains more than one keyword, it is explicit. If keywords are present without the whole instance of speech being offensive or abusive, we consider it non-abusive keyword use, similar to [10].

Our study focuses on all types of indirect expressions: implicit abuse, implicit offensive language, and non-abusive use of keywords.

Table 1
Definition of implicit, explicit, non-abusive use in relation to the nature of speech.

	Abuse (incl. Hate speech)	Offensive	Normal
With keyword	Explicit abuse	Explicit offensive language	*Non-abusive use of keywords*
Without keyword	Implicit abuse	Implicit offensive language	/

Table 2
Statistics of the three datasets used in our study.

	Abuse	Offensive	Normal	Total
AbusEval	2927	1713	9460	14 100
Davidson	1430	19 190	4163	24 783
Founta	4965	27 150	67 881	99 996

3.2. Resources

We make use of two types of resources for our research: (1) a set of abusive language datasets labelled as abuse, offensive or normal, and (2) collections of keywords that enable us to distinguish, within the datasets, the cases that make explicit use of these keywords. By using the keywords to find matches, we break down the datasets into two subsets: "**Any keyword**" and "**No keyword**". When the "**No keyword**" subset overlaps with either the abuse or offensive label, we deem these **implicit abuse** and **implicit offensive language**, respectively.

3.2.1. Datasets

We chose to use three multi-class datasets: *AbuseEval* [5,51], *Founta* [8] and *Founta* [8]. As opposed to the vast majority of existing datasets providing binary labels (abuse vs. not), these three datasets were selected for enabling distinction of the three categories of our interest: (1) abuse, which subsumes hate speech, (2) offensive, and (3) normal.

The original class labels needed adapting slightly to enable comparative analysis across datasets by mapping them into the above three classes:

- *AbuseEval*. "Abuse" were used as-is. Instances that fall under "offensive" but not "abuse" were used as "offensive".
- *Davidson*. Original classes ("hate", "offensive (but not hate)", "neither") were directly mapped into ("abuse", "offensive", "normal").
- *Founta*. "Spam" and "normal" were combined into "normal". We made the decision based on the other two datasets — both *Davidson* and *AbuseEval*'s "normal" classes contain instances that would be considered "spam" in *Founta*.[3] "Hate" was mapped into "abuse". "abusive" was renamed to "offensive". We made this decision because their definition of "abusive" does not mention any intentionality and is hardly indistinguishable from that of "offensive"; the annotators could not distinguish them, either.

The class labels after mapping are distributed as shown in Table 2.

There are two main differences to notice in the datasets. *Founta* is a few times larger than the other two, and *Davidson* close to twice the size of *AbuseEval*. All three datasets are imbalanced, but in different ways: The majority class is Normal in *Founta* and *AbuseEval* as in most other abusive language datasets, but Offensive in *Davidson*. The smallest class is Abuse for both *Davidson* and *Founta*, but Offensive for *AbuseEval*.

[3] Such as "I added a video to a USER playlist URL ...", " Charlie Sheen engaged to porn star URL ... ".

Table 3
The three datasets broken down by class labels and whether having keywords. The chi-squared statistics shows the dependency between keyword presence and class labels. All p-values < 0.001.

	Abuse	Offensive	Normal	Total	χ^2
AbuseEval					
Any keyword	764	707	*583*	2054	1831.459
No keyword	2163	1006	8877	12 046	
Davidson					
Any keyword	1408	19 163	*3992*	24 563	619.159
No keyword	22	27	171	220	
Founta					
Any keyword	2673	23 755	*5517*	31 945	57 343.064
No keyword	2292	3395	62 364	68 051	

3.2.2. Keywords

We gathered widely used sets of keywords that can be categorised as either slurs or profanity. We use the Hatebase[4] lexicon to cover 1532 slurs, and the No Swearing[5] lexicon to cover 298 profanity words, after excluding from the latter those that are also considered slurs. In the case of the 1532 Hatebase slurs, we also preserve information on what attribute of the victim (topic) the slur is targeting at, e.g. ethnicity or religion — the latter enables an analysis by topic in Section 4.2.

4. The source of the problem: slurs and profanity in the data, and their association to offensiveness and abusiveness

We first analyse the **Any keyword** and **No keyword** subsets to understand (1) how implicit abuse and offensive language are manifested in the datasets, and (2) to assess the presence of keywords in the normal class.

We show that certain datasets are (a) more **keyword-intensive** – containing keywords more often overall–, and certain datasets are (b) more **keyword-dependent** –the association between keyword presence and class labels is stronger. It is important to note that one does not necessarily determine the other. We then break the keyword presence down by possible topics — about ethnicity, gender, . . . or just general swearing.

4.1. Keyword presence and its association to abusiveness and offensiveness

We show in Table 3 the breakdown of the three datasets by class label and keyword presence. Overall, regardless of the nature of speech, the overwhelming majority of *Davidson* posts contain keywords, making it the most keyword-intensive, in contrast to less than half for the other two datasets.

Having at least one keyword means that an instance is far more likely to be offensive or abusive than innocent. A chi-squared (χ^2) test of dependence further confirms this. A dataset being more keyword-intensive overall does not mean that the class labels in that dataset are more keyword-dependent, comparing the ratio of instances with keywords and the chi-squared values.

The proportion of implicit abuse and implicit offensive language (red, underlined), and non-abusive use of keywords (green, italic) in a dataset depends on both overall keyword-intensity and class keyword-dependency.

4.2. A closer look into the topics

The topics of the keywords present – the overall presence in the whole dataset and the relative presence across different types of speech – are shown in Fig. 1.

Regardless of the nature of speech and the specific dataset, general profanity is the most common type of keyword, followed by slurs related to ethnicity and gender. At the same time, *Davidson* stands out from the other two datasets — general profanity appears much more frequently than any slurs in both *Founta* and *AbuseEval*, while, in *Davidson*, gender- and ethnicity-related slurs are also very common, with the former even more so than general profanity. For any keyword topic, the frequency is the highest in *Davidson*, the most keyword-intensive dataset.

Certain topics are associated with certain class labels. One might expect ethnicity-related slurs to be associated with abuse and especially hate speech in the datasets, as they have been with racism historically. Interestingly, only *Founta* displayed such tendency unequivocally: under the "abuse" class, which is formed by annotated hate speech, ethnicity-related slurs are about five times more likely to appear than the other two classes. In clear contrast, such slurs are also commonly found under the "normal" class in *Davidson*, meaning they are used in innocent settings. Nonetheless, they are still much more common in abusive than offensive language. The difference is much smaller in *AbuseEval*, with these slurs taking up only a small fraction in all classes. The pattern across classes for gender-related slurs is much more consistent across the three datasets. They seldom appear in the normal cases; they are more common in offensive language that is not considered abuse or hate speech. General profanity follows a similar trend, except the frequency is much higher across all classes. Some types of slurs cover a noticeable proportion only in *Davidson*: Sexual orientation-related slurs are fairly common in hate speech, but very rare in non-hate offensive language and normal speech; nationality- and class- related ones are most common in normal speech.

All in all, the different patterns across topics show how different slurs and profanity are used and perceived and echo with the overall keyword presence and its association with offensive and abusive speech (Section 4.1).

4.3. Discussions: from dataset construction to model training

Overall presence of keywords and the association between keywords and class labels can be traced back to how the datasets were built, and will have effects on any classification model trained on the datasets. In what follows we link our analysis with the sampling and annotation strategies, as well as discuss the expected impact on abusive language detection models.

4.3.1. Sampling

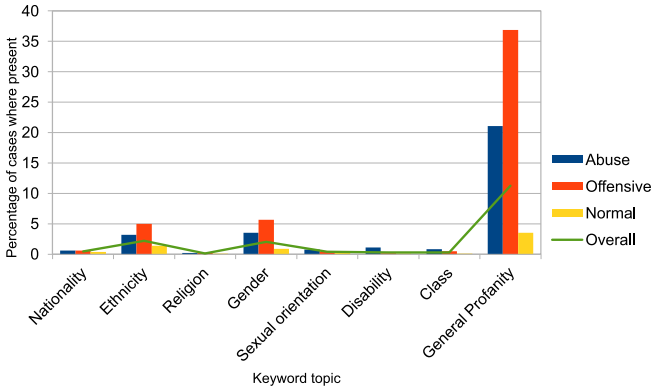
Sampling – how the initial data is gathered before annotation – affects how keyword-intensive the dataset is and the distribution of the topics, and reflects the domain of the dataset by showing how the data construction is motivated.

Offensive language generally represent less than 3% of social media content [8,51]. Thus, all datasets apply some process, such as text search, to increase the proportion of offensive and abusive content.

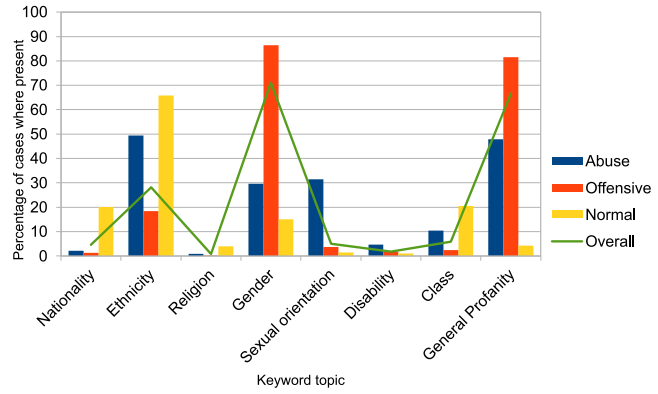
There are two key components which contribute to how biased towards keywords the sampling process is: the approach and the criteria. Sampling approaches in offensive language datasets can be divided into boosted random sampling and biased sampling [26]. While the former approach applies the criteria after drawing an initial random sample, the latter draws a biased sample with the criteria. The criteria can be compared in terms of whether inherently offensive terms are used for search.

Biased sampling was applied on *Davidson*. An initial sample was drawn through lexicon search with Hatebase, and timelines from users identified in the lexicon search were also included. This very focused approach resulted in the dataset having the most intensity and broad coverage of slurs and profanity overall. On *Founta*, Hatebase was similarly used for lexicon search, in addition to No Swearing. Negative sentiment was also part of the criteria. Despite the direct search of slurs

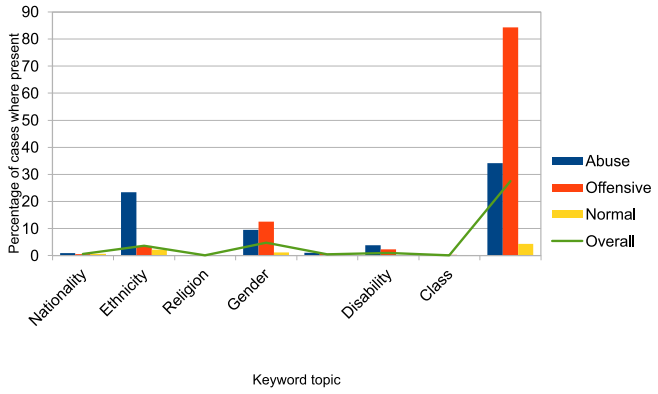
[4] https://www.hatebase.org/.
[5] https://www.noswearing.com/dictionary/.



(a) AbuseEval



(b) Davidson



(c) Founta

Fig. 1. Keyword topics in the three datasets, shown as the percentage of instances in the whole dataset or under each class that contain any keyword of that topic. Note that the y-axis is different to prioritise easier comparison across topics in each dataset.

and profanity, a boosted random sampling approach was taken. As a result, it contains such keywords much less frequently than *Davidson*, which in turn makes the dataset a lot less keyword-intensive, with a lot of clean cases without any keyword or offensive language. *AbuseEval* is a re-annotated version of the *OLID* dataset [49], and therefore inherits the sampling criteria of the latter. The text search criteria applied in this case is a lot less direct, such as "you are", "she is", "MAGA", "gun control". They also looked for replies to extreme-right news and posts that get filtered out by safe search. These less direct criteria reflect that the dataset aims to cover a broader spectrum of and less extreme types of offensive language. The presence of slurs and profanity keywords in the resulting dataset is thus much less frequent than the other two hate speech datasets, making it the least keyword-intensive among the three. Nonetheless, interestingly, the topic distribution is very similar to that of *Founta* (Fig. 1).

What is more, sampling through text search has an unintended effect on class distributions. It fulfils the intended purpose to boost the proportion of offensive or abusive posts, indeed. As shown in Table 3, all three datasets contain much more offensive or abusive posts than social media generally, although using boosted random sampling instead of

biased sampling (*Founta*) or using terms not inherently offensive for biased sampling (*AbuseEval*) makes the effect milder. On the other hand, sampling through text search reduces the actual proportion of actually abusive posts within the offensive ones. This is because abuse is less likely to contain keywords, as shown in all three datasets, and as pointed out by [5] that abuse depends less on lexical features. Comparing the three datasets, the more keyword-intensive a dataset is, the smaller the "abuse" class is compared to "offensive", with *Davidson* being the most so and *AbuseEval* being the least.

4.3.2. Annotation

General profanity can be used for stylistic purposes or emphasis [40]; some slurs have been reclaimed by targeted groups [10], and is common in African American English (AAE) dialect [23]. It is thus necessary to distinguish keyword use and actual offensiveness or abusiveness.

Inconsistencies in definitions across datasets challenge model generalisability [52]. Datasets also vary by how specific the guidelines are, such as a more detailed explanation of the definition and clarifications on edge cases. Then, annotators differ by whether they were crowdsourced or experts. Generally, expert-annotated data are considered of higher quality [53] and produce better-performing classification models [54].

Among the three datasets, *Founta* annotators were allowed the most freedom as the annotation instructions and taxonomy were not fully-fledged from the beginning, but were progressively completed as annotations were collected. The annotations were completely crowdsourced; annotators first carried out an exploratory round of annotations, based on very brief definitions of a range of possibly overlapping concepts, after which the final class labels were then defined. This resulted in annotations heavily based on keywords (Table 3). Thus, although also focusing on the specific abuse type of hate speech during annotation like *Davidson*, the annotated "hate speech" label in the original dataset is expected to cover a wider range of phenomena, including some instances that contain slurs but are not hate speech. For example, in *Founta*, "This what happens when you separate yo self from nias who don ' t eat they food cold. You FLOURISH..." was labelled as Hateful, while the n-word was used in a reclaimed manner. In comparison, similar usage of the n-word in *Davidson* such as "I ain ' t never seen a bitch so obsessed with they nigga (emoji) I ' m obsessed with mine (emoji)" was consistently labelled as Offensive.

The other two datasets had defined the class labels before the annotation process. Although also completely crowd-annotated, the annotators of *Davidson* received more intensive instructions: a paragraph explaining in detail along the definition of each class label; they were also explicitly asked to not base their judgements solely on the words in isolation. This specific instruction is expected to make its "abuse" class the most specific among the three datasets, focusing on hate speech and with annotators' bias towards keywords reduced. It is reflected in the data by having the least inter-dependency between keywords and class labels among the three datasets. *AbuseEval* has both crowd and expert annotations: in two separate studies, experts annotated the abusiveness [5] of originally crowd-annotated offensive posts [49]. The definition of offensiveness is loose and broad, covering "any form of non-acceptable language (profanity) or a targeted offense". Although carrying a broader definition of abuse without focusing on a subtype like the other two datasets, abusiveness was annotated with clearer instructions in the form of a decision-tree. This resulted in an overall moderately keyword-dependent dataset.

On the other hand, common trends are found when it comes to the keyword topics and perceived hatefulness in the two hate speech datasets (*Davidson*, *Founta*). While instances labelled as offensive contain gender-related slurs much more often than abuse, those labelled abuse much more frequently contain ethnicity-related slurs. There are two possible explanations to this: Indeed, the authors [7] noticed that annotators tended to perceive racism and homophobia as hate speech, but sexism as only offensive, consistent with the findings of an earlier dataset study [54]. This likely reflects how the western societies, where the research is focused, perceive different types of abuse. For instance, European countries most commonly centre their legal definition of hate speech around race, followed by religion and sexual orientation.[6]

To summarise, crowd annotators, who reflect the general public, display a tendency to rely on keywords, while experts rely on keywords less. Offensive and abusive are seen as fundamentally different by experts, but are confused by crowd annotators. Instructions for crowd workers to actively consider the context of words used, as with instruction to consider the dialect of the speaker [23], reduce biases induced by slurs. Some topics are perceived more abusive/hateful than others.

4.3.3. Implications on classification model training

Based on the analysis above, we can expect the data to have effects on classification model training, including both utilising pre-trained models and fine-tuning to the task.

First of all, a model trained on an imbalanced dataset is expected to display a tendency to predict the majority class the best. Then, keywords' overall presence and association with the class labels would make classification difficult. A pre-trained model would already have encoded offensive meanings of slurs and profanity. Thus, keyword-intensive data, such as *Davidson*, can mislead a pre-trained model even before exposing it to keyword-label combinations. During the fine-tuning stage, the association of keywords with offensive and abusive class labels would be further integrated into the model, with the effect being the strongest in the more keyword-dependent datasets, such as *Founta*.

Furthermore, models would struggle to generalise to unseen datasets, although generalisation is expected to better in dataset pairs which are more similar. In terms of keywords, *Founta* and *AbuseEval* are more similar, both being more keyword-dependent, less keyword-intensive, and having similar keyword topic distributions. In terms of class labels, *Davidson* and *Founta* both focusing on a specific type of abuse, hate speech. Nonetheless, the guidelines of the latter are broader, making it relatively similar to *AbuseEval*, where both "abusive" and "offensive" are umbrella terms. Considering both factors, it is thus expected that generalisation between *Davidson* and *AbuseEval* would be the most challenging.

5. Challenges reflected in the classification model: when the words present do not match the underlying meaning

In the last section, we analysed the presence and association of keywords in the datasets, related them back to dataset building, and hypothesised what impact they would have on model training. In this section, we assess such impact in detail through both in- and cross-dataset experiments. We also extend these discussions to a novel generalisation scenario where two sources of out-of-domain data are combined for training.

The model we use for this assessment, BERT [31], is commonly used in abusive language detection and achieves strong results in shared tasks (with in-dataset evaluation) [55,56] and generalisation studies (with cross-dataset evaluation) [28,45]. Thus, although we only consider one model design, the results are expected to reveal common issues in most if not all abusive language detection models.

The classification model we use is BERT-base-uncased with transformer layers and a subsequent pooling layer all initialised with the pre-trained weights obtained from Huggingface.[7] After the pooling layer, a fully-connected layer, randomly initialised, maps the pooled representation to a class prediction through a Softmax function.

[6] https://www.legal-project.org/issues/european-hate-speech-laws.
[7] http://huggingface.co/transformers.

Fig. 2. Confusion matrices of BERT predicted vs. true class labels, normalised by the true values, organised by keyword presence and datasets. True abusive and offensive instances in the "no keyword" subset are implicit expressions (red boxes); True normal instances in the "any keyword" subset are non-abusive use of keywords (green boxes). Models are trained and evaluated on the same datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
BERT performance in macro-averaged F1 scores. In-dataset training and evaluation is **bolded**; the best cross-dataset performance is underlined. "Two out-of-domain datasets": When the training sets of two datasets apart from the evaluation dataset are combined for training.

Training/Evaluation	AbuseEval	Davidson	Founta
AbuseEval	**0.634**	0.466	0.582
Davidson	0.477	**0.752**	0.633
Founta	0.544	0.582	**0.738**
Two out-of-domain datasets	0.571	0.577	0.640

We settled on a learning rate of $1e^{-5}$, maximum sequence length of 70 after hyperparameter experiments on the validation set, similar to the settings of a previous study [27]. We saved model checkpoints every 1000 steps and performed early stopping after 4000 steps of no improvement over validation macro-F1 with a maximum budget of 20 000 steps. The rest of hyperparameters were kept as default. Performance metrics reported are all means computed from the 8 models.

The mean macro-averaged F1 scores are shown in Table 4. For the remainder of this section, we focus on performance metrics of specific class labels in relation to keyword presence.

5.1. The impact of keywords on in-dataset classification

We first discuss results on in-dataset settings, i.e. were different subsets of the same dataset are used for training and testing. Fig. 2 shows, for each class label, how having or not having keywords impacts what the classification model predicts. The red boxes highlight cases of implicit abuse and offensive language, whereas the green boxes highlight cases with non-abusive use of keywords. The three datasets have a lot in common, when it comes to what mistakes the model would make. Moreover, these common patterns have a strong connection to the dependency between keyword presence and class labels shown in Table 3. We discuss these results next, first focusing on implicit and explicit expressions, and then on non-abusive posts with keywords present.

5.1.1. Implicit vs. explicit expressions

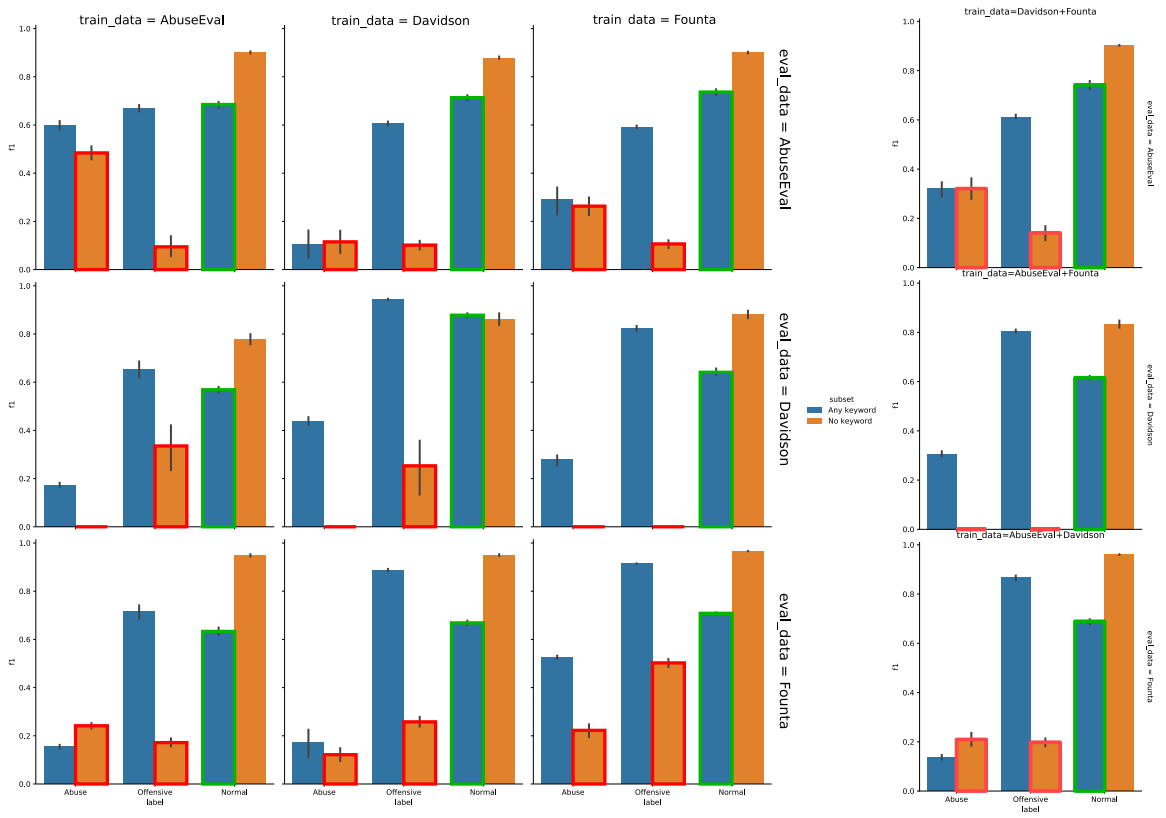
On all three datasets, without keywords, the strong tendency to predict anything as harmless does not differ much across the three datasets, although, the datasets vary by how frequently instances contain keywords (intensity) and how strong the association between having keywords and the abuse and offensive labels is (dependency). It is likely that the effect of these two factors offset each other – recall that *Davidson* is the least keyword-dependent, which is expected to be an advantage, but it is also the most keyword-intensive, limiting the available implicit instances to learn from.

At the same time, the relative performance on the implicit abuse and offensive classes depends on the imbalanced class ratios in the datasets. For both *Davidson* and *Founta*, where there are many more offensive than abuse instances, the model struggled more with implicit abuse than implicit offensive language. In contrast, in *AbuseEval*, where there are many more abuse than offensive instances, the proportion of correctly classified implicit abuse is much larger.

In clear contrast to the implicit case, explicit abuse and offensive language are much easier to detect. Nonetheless, the model tends to misclassify explicit abuse as offensive language. This problem is shared by all datasets regardless of the proportion of abuse. It is linked to the proportion of explicit instances comparing the abuse and offensive classes: in all three datasets, offensive language contains keywords much more frequently than abuse. The severity of the problem depends on how big this frequency contrast is. On *Davidson*, offensive language is almost 10 times more likely to be explicit than abuse, the model made this mistake most frequently, followed by *Founta* and then *AbuseEval*.

5.1.2. Non-abusive use of keywords

On instances with non-abusive use of keywords (intersection of Any keyword & Normal), as expected, the model displays a tendency to



(a) Single training dataset (b) Double training dataset

Fig. 3. BERT model's performance, in- and cross-dataset evaluation on three datasets, on instances with or without keywords, by class labels. Showing 95% confidence intervals generated through 1000 iterations of bootstrapping. Bars surrounded by a red box refer to cases of implicit abuse and implicit offensive language, whereas those surrounded by a green box refer to non-abusive use of keywords. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

falsely flag innocent normal speech as offensive or abuse, although this tendency is not as strong as the effect of implicitness on identifying abuse.

Generally, the model tends to mistake these misleading instances as offensive rather than abuse. Offensive language being explicit much more frequently than abuse underlies this phenomenon.

How keyword-dependent the dataset is has a clear effect on how common the model falsely flags non-abusive keyword use as offensive or abuse. On the least keyword-dependent *Davidson*, where there are a lot of instances with non-abusive keyword use, such mistakes are much rarer than in the other two datasets, and the recall of the normal class is similar with or without keywords. The model makes such mistakes more frequently on the most keyword-dependent *Founta*, followed by *AbuseEval*, even though both datasets have smaller offensive ratios than *Davidson*.

In summary, the in-dataset performances reveal that the dependency between class labels and keyword presence is the biggest factor underlying the main challenges in classification models trained and evaluated end-to-end.

5.2. Factors in cross-dataset generalisation

Fig. 3 breaks down how the model generalises to a different dataset on each class label with or without keywords, compared to in-dataset evaluation (top left, middle, bottom right). Red boxes highlight performances on implicit abuse and implicit offensive language, and green boxes highlight performances on normal, non-abusive posts with keywords.

5.2.1. Generalisation difficulty vs. detection difficulty

In the scenario of cross-dataset evaluation, the model also struggles with implicit expressions and non-abusive keyword use, although the extents differ. When the model is evaluated on an unseen dataset, it is additionally faced with the difference in class label definitions, manifested in the gap between in- and cross- performances.

Consistently, normal speech without any keywords is the easiest for the model to generalise on across datasets. Generalisation on explicit offensive language is noticeably more difficult than in-dataset evaluation. When it comes to implicit expressions, offensive language or abuse, the model struggles to both detect and generalise. Abuse carries additional challenges: the difficulty to detect and generalise posed by implicitness is even more severe than the offensive case; even the explicit cases are hard to detect and generalise on.

5.2.2. Generalisation and similarity factors between datasets

In Section 4.3.3, we discussed how sampling and annotation approaches caused datasets to have varying degrees of keyword-intensity and keyword-dependency, as well as varying specificity of "abuse" definitions. Table 5 summarises how the three datasets compare on these factors. As hypothesised, generalisation depends on dataset similarity on these factors, reflected in macro-averaged F1 scores shown in Table 4.

Breaking down generalisation by the types of instances (Fig. 3), we see that, while keywords affect the relative performance on implicit and explicit expressions, the coverage of the "abuse" class limits performance on this class.

Having more instances of a certain type of speech is related to better generalisation for that specific type for each dataset, such as

Table 5
Comparison of dataset characteristics contributing to generalisation. Summarised from Section 4.3.3.

	Keyword intensity	Keyword dependency	Abuse definition coverage
AbuseEval	+	++	+ + ++
Davidson	+ + ++	+	+
Founta	++	+ + ++	++

explicit offensive language in *Davidson*. Similarity in definitions is linked to better generalisation between dataset pairs, as seen between *Davidson* and *Founta*. Nonetheless, the higher keyword-dependency in *Founta* specifically degrades implicit offensive language. There is a one-way generalisation advantage on implicit offensive language from less keyword-dependent data to higher ones — on implicit offensive language, the *Davidson*-trained model also outperformed the *AbuseEval* one.

5.2.3. Generalisation from heterogeneous training data

Previous research showed that heterogeneous augmentation (augmenting in-domain training data with out-of-domain training data) can be detrimental [57]; 1-to-1 cross-dataset generalisation largely depends on similarity in keyword distribution and abuse class definition, reflected in implicit and abuse class performances. Here, we experiment with heterogeneous training data without in-domain data (combining two datasets that do not include the evaluation dataset) to connect these two observations.

Looking at the highest level of macro-averaged F1 scores (Table 4), combining two sources of out-of-domain data mostly results in an improvement over generalisation from only one out-of-domain training dataset, except for when both training datasets generalise poorly individually — on *Davidson*, where the gaps between cross-dataset and in-dataset performance is the largest.

Subset performance (Fig. 3(b)) shows that the improvement is on the most difficult cases for generalisation — more so on the abuse class than on implicit expressions. While an abuse class with a narrow definition generalises poorly to a broader one in one-to-one scenarios, when two narrow definitions are combined, the improvement is evident — comparing one-to-one and two-to-one generalisation on *AbuseEval* as the evaluation dataset. If one dataset is significantly larger (*Founta*), relative subset performances would resemble the one-to-one generalisation from the larger dataset.

5.3. Discussions: hypothesis verification and additional insights from model behaviour

In Section 4.3, we discussed how sampling and annotation approaches affect key characteristics of a dataset: keyword-intensity, keyword-dependency, and class proportions. We also hypothesised how each characteristic would manifest in detection and generalisation of instances of different types. Here, through the model behaviour expressed in in- and cross-dataset experiments, we complete the hypothesis verification and discuss additional insights: the relevant importance of the characteristics, and how they interact with each other. Through these analyses, we provide suggestions on dataset selection and construction, in order for better detection and generalisation of the misleading expressions — implicit expressions and non-abusive keyword use.

5.3.1. The relative effect of and interaction between keywords and class labels

Firstly, we hypothesised that the models would have the biggest difficulty when detecting instances with implicit expressions of abuse or offensive language and non-abusive keyword use, as keyword use is associated with the offensive and abuse classes. Secondly, higher keyword-intensity and keyword-dependency were both expected to worsen these two challenges. We additionally hypothesised that the models will perform the best on the respective majority classes of the training datasets, as a general pattern of machine learning models.

Indeed, models show common struggles with the two challenging types of instances, confirming our first hypothesis: the models tend to mistake implicit abuse and offensive language as normal speech, and normal speech with non-abusive keyword use as offensive or abusive. However, the effects are not equal. The models are much more likely to miss the implicit expressions in abuse or offensive language than to falsely flag normal speech with keywords as offensive or abuse. This means that, in detecting abuse and offensive language, the absence of strong indicative lexical features has a stronger effect on causing false negatives than the presence of them on causing false positives.

As expected, being highly keyword-dependent is always detrimental. The model learns a stronger association between keywords and offensive classes, hindering the classification of both implicit expressions and non-abusive keyword use, as shown in the model behaviour when training and evaluated on *Founta*.

By contrast, being highly keyword-intensive can be a double-edged sword, under the influence of keyword-dependency. On one hand, it limits the total instances without keywords available for training. As a result, the model performs poorly on the implicit expressions of abuse and offensive language, as shown in the results for *Davidson*, despite it being the least keyword-dependent dataset. Furthermore, the model can even mislabel normal speech without keywords as offensive, which the other two models seldom do on the other two datasets. On the other hand, it benefits the detection of non-abusive keyword use and explicit offensive language. If and only if, with suitable annotation instructions, the dataset's keyword-dependency is low, containing more keywords can mean having more instances with non-abusive keyword use, which facilitates the classification of such instances. The *Davidson* model displays such a clear advantage over the other two, which do not differ a lot on this aspect. In binary classification, explicitness benefited the detection of the positive class [42]. Here, however, having more explicit expressions only benefits the detection of explicit offensive language — but not explicit abuse.

The reason why explicitness does not benefit the detection of abuse in the same way as offensive language lies in the fact that the latter is always more likely to be explicit. As a result, all three models struggle with the "abuse" class the most, and in similar ways. When there are keywords, explicit abuse is often mistaken with explicit offensive language, even on *AbuseEval*, where there is more abuse than offensive language. When there are no keywords, implicit abuse is most commonly mistaken as harmless, normal speech, as is implicit offensive language.

The effect of class labels is mainly through the coverage of the abuse class, rather than majority vs. minority. This coverage is reflected in the dataset as the relative proportions between abuse and offensive language. The lower the ratio of abuse, the more likely the model is to mistake explicit abuse as offensive language. This is seen across the three datasets. Furthermore, on the extremely challenging instances with implicit expressions, the smaller class is misclassified even more: on *Founta* and *Davidson*, implicit abuse was misclassified relatively more often than implicit offensive language, while it was the opposite for *AbuseEval*. This offers an explanation to the inconsistency in the literature on the effect of the positive class ratio in binary classification: the proportion of actual abuse is likely a moderating factor.

5.3.2. Similarities and differences between generalisation and in-domain detection

In terms of model generalisability, our hypothesis was mainly based on the overall similarity between datasets: generalisation between the two most different datasets overall would be the most difficult. This similarity consisted of three key dataset characteristics that are direct products of sampling and annotation approaches: keyword intensity, keyword dependency, and the abuse class coverage.

This hypothesis is verified through cross-dataset macro-averaged F1 scores, but a breakdown of the performance by keywords and classes produced far more insights, on the connection between detection and generalisation and on the separate effects and relative importance of the three characteristics.

Instances that are harder to detect are also mostly harder for the model to generalise on, but there are exceptions. The difficulty of detecting implicit expressions and abuse seen in the in-dataset scenario is magnified when it comes to cross-dataset generalisation. However, generalisation on explicit offensive language, which is a lot easier to detect, is similarly difficult to that on instances with non-abusive keyword use. This means that while all datasets show an unequivocal, strong association between keywords and the offensive class, this association is slightly different across datasets. On the other hand, while non-abusive use of keywords misleads all models, the way in which it does so has limited variability.

As in the in-dataset case, keyword-dependency is always detrimental for generalisation, which mainly affects the implicit expressions. When keyword intensity is also low, as in *AbuseEval*, it means there is more training instances for implicit expressions, facilitating generalisation on such instances. Even in situations where both training and evaluation datasets are highly keyword-intensive, as in *Founta* and *Davidson*, the less keyword-dependent one, *Davidson*-trained model, generalises better to *Founta* on the implicit instances than the other way round.

The benefit of keyword-intensity for generalisation on non-abusive keyword use and explicit offensive language is similar to that in the in-dataset scenario. Specifically, generalisation on these instances is the best between the more keyword-intensive *Davidson* and *Founta*, which were sampled in similar ways. On these instances, they also generalise better to *AbuseEval* than the other way round. This benefit of a wider coverage of keywords outweighs the similarity in keyword topic distributions. Previous research found that wider coverage of phenomena improves generalisability [46,47]; this finding adds keywords to the type of such phenomena whose coverage can benefit generalisability — limited to instances that actually contain keywords.

This comes at a cost on the implicit expressions. Depending on the proportion of such instances, this demerit can be hidden by the smoke screen of a good performance on the explicit ones, when looking at the overall macro-averaged F1 scores. Because *Davidson* and *Founta* are both heavy in explicit instances, the overall F1 scores are dominated by the fact that they generalise better to each other on the explicit instances, while the advantage on the implicit ones displayed by *AbuseEval* is not reflected in the overall F1 scores. Only looking at the macro-averaged F1 is also the reason why previous research concluded that having more explicit expressions carries generalisability benefit [26], while in reality, such benefit is limited to the dominating explicit instances.

The abuse class again limits the above effect, similar to the in-dataset scenario. Previous work in binary classification suggested that a positive class with a broad definition is more generalisable [28,46]. This is mostly true in our results, but when the narrow ones are sufficiently similar, they generalise reasonably well to each other.

The above effects translate to the scenario when heterogeneous out-of-domain data is combined for training, with the additional benefit of better generalisation compared to single source. Having multiple data sources likely served as regularisation against overfitting to one dataset. This benefit is nonetheless limited by the pair-wise dataset similarity and one-to-one generalisation, and is specific for cross-dataset generalisation, as augmenting in-domain training data with out-of-domain training data can be detrimental [57]. Combining multiple sources helps alleviate the difficulty of generalising from narrow to broader definitions of abuse.

5.3.3. Suggestions on dataset construction and application

In Section 4.3, we discussed how annotation and sampling contribute to dataset characteristics. In this section so far, we saw that these characteristics of a dataset in turn affect model performance on different types of instances.

Drawing from these discussions, attention should be paid on both sampling and annotation when constructing a dataset, in order for better, generalisable model performance on the most challenging instances, i.e. implicit expressions and non-abusive keyword use.

To start with, the initial sample is better to be not drawn with biased sampling directly using slurs and profanity. Otherwise, such a filtering criterion increases the frequency of slurs and profanity, which in turn impairs model performance on the implicit expressions. Furthermore, it reduces the instances actually containing abuse in the sample, so this is especially important when studying abuse as a separate phenomena from offensive language.

If words and phrases are to be used to boost the ratio of the abusive class, using those that are not inherently offensive or abusive [49] can reduce the direct link between keywords and class labels. Furthermore, boosted random sampling (applying the criteria after drawing an initial random sample) results in less biased data than biased sampling (drawing a biased sample with the criteria) [26]. Instead of keyword-based biased sampling, There are a few alternatives which can reduce the bias towards slurs and profanities. One possible approach is to draw samples based on communities, such as forums which are banned due to hateful discussions [10,58]. Another alternative is to use semi-supervised learning, such as the SOLID [59] dataset, labelled with confidence scores of models trained on *OLID* [49]. By using a range of models with different inductive biases and the means and standard deviations of the confidence scores, the semi-supervised labels are not overfitted to any particular model. These methods that do not rely on keyword filters removes one significant source of potential bias. Nonetheless, these models are not free from additional risks: some narrow communities may have very specific language styles which deviate from the mainstream or require substantial contextual knowledge in order to understand certain discussions; biases in the seed dataset may be transferred to the semi-supervised data. These approaches can also be combined to reduce the influence of a small number of major biases.

Yet, it is worth keeping in mind that the challenge brought by keywords also comes from the nature of abusive language, rather than the sampling method — offensive language is always more likely to contain keywords than abuse. Thus, no matter how sophisticated the sampling process is, the ratio of implicit and explicit instances and author and topic distributions should be carefully studied, controlled, and reported.

Annotation guidelines should be as specific as possible. A prerequisite is to have detailed definitions. Confusions between perceived offensiveness and the specific abuse under investigation can be reduced by explicitly listing out common mistakes in the guidelines [20]. Biases against language styles should be controlled by explicitly asking annotators to rely on words less [7] and reminding them of factors in dialects [23]. If high levels of details appear overwhelming, a decision-tree-like guideline [5] can simplify the decision process.

The above applies to general dataset construction. Additionally, depending on the specific application scenario, there are other factors to consider.

The target domain can be rich in slurs and/or profanity, such as a community of marginalised groups that frequently use reclaimed slurs. In these communities, false positives caused by slurs carry serious negative social impact — further oppressing marginalised groups. Thus, data for training an abusive language detection model need a balance between good coverage of such keyword use and implicit expressions. In such a scenario, sampling directly through slurs and/or profanity can be actually beneficial, provided that extra care is taken to make sure there are enough implicit instances.

The definition of abuse should also be tailored to knowledge about the target domain. Having a wide coverage of all kinds of abuse in the training data is not as useful as having an accurate representation of what abuse looks like in the evaluation data. Thus, if there is enough knowledge about the abuse present in the target domain, the definition when building or choosing the training data should be as specific as possible, such as "hate speech". Otherwise, when there is much unknown about the target domain, using broadly defined abuse for training is a safe choice.

6. Conclusions

By looking at the presence of keywords categorised as profanity or slurs in datasets, in this paper we study the tendency of abusive language detection models to label content as abusive, offensive or normal. Investigating with three widely-used abusive language datasets where this 3-way classification is possible, we assess the implications of decisions made in the dataset construction stage in the development of abusive language detection models. We break our analysis into two main parts. First, we analyse the prominence of profanity and slurs in the different datasets, focusing both on keyword intensity (presence of keywords in the dataset) and keyword dependency (association of keywords with classes). And second, we assess the impact of these dataset patterns on the ability of models to detect abusive language, both in in- and cross-dataset scenarios, looking at the ability to generalise across datasets. We focus on two challenging cases in detail: (1) implicit cases of abusive and offensive language where keywords are not present, and (2) non-abusive use of keywords.

We defined three research questions for our study, which we answer next.

RQ1: How do approaches in dataset construction lead to different patterns of slurs and profanity presence in the dataset?

In sampling, factors that make the process more biased towards keywords are: biased rather than boosted random sampling, text search through slurs and profanity rather than terms that are not inherently offensive. A sampling process more biased towards keywords makes a dataset contain more keywords, i.e. more keyword-intensive. Less biased processes lead to similar topic distributions of the keywords.

In annotation, factors that make the process more biased towards keywords are: crowd-sourced rather than expert annotators, brief rather than detailed instructions. An annotation process more biased towards keywords makes the association between keywords and offensive and abuse class labels stronger, i.e. more keyword-dependent. There are also common patterns in how abusive keywords of different topics are considered.

Additionally, both sampling and annotation affect class distributions. Sampling biased towards keywords increase offensive instances but decreases the ratio of abuse compared to offensive. A broader definition of abuse in annotation leads to wider coverage of phenomena.

RQ2: How does such keyword presence in turn affect the detection model behaviour?

The role of keywords is two-folds: their association with class labels make them act as useful lexical features, which also means that instances that go against this association lack training instances. Thus, the most challenging instances for classification are the ones with fewer lexical features and instances available in the training data: implicit expressions and abuse of any kind, followed by non-abusive keyword use.

Comparing datasets, varying keyword intensity and dependency affects performance through affecting the number of instances. Keyword dependency impairs the detection of implicit expressions and non-abusive keyword use. Keyword intensity decreases performance on implicit expressions, but improves that on explicit ones. Provided with low keyword dependency, it can also benefit the classification of non-abusive keyword use.

All models make similar mistakes when it comes to abuse, mistaking it either as offensive or normal speech depending on whether keywords are present, although the coverage of the abuse class in the training data has some influence.

Depending on the makeup of the evaluation dataset, some of these effects may not be reflected in the overall model performance.

RQ3: How does this effect differ when it comes to model generalisation?

The above effects of keywords apply to both in-dataset detection and cross-dataset generalisation.

Generalisation introduces extra challenge: the effects of keywords are more prominent, especially on implicit expressions and abuse; generalisation on the abuse class largely depends on the definition difference between datasets.

Combining out-of-domain data is likely to be beneficial for generalisation compared to training on single-source out-of-domain data, mainly by addressing the challenge of abuse definition differences.

Based on the above research questions and answers, we have also provided suggestions on dataset construction.

Limitations and future work. Our work is not free from limitations, which also open up new directions for future research. We had kept the experiment conditions consistent across datasets and taken a careful hypothesis verification approach to explain the results. Our findings were also in line with available relevant research. Nonetheless, even though more rigorous hyperparameter tuning could potentially lead to marginally better performance metrics, there are random peripheral factors which would have fluctuated the results. The arguably most important limitation is the definition of implicit and explicit based on slurs and profanity. We used lexica to define explicitness, for which the motivation was two-fold: (1) it operationalises explicitness without inconsistency and subjectivity which would have otherwise been introduced by manual annotation, by biases in machine learning models or by semi-supervised approaches such as using the Perspective API; (2) our use of lexica is consistent with previous literature, putting our findings in context [5,35,38]. However, implicitness and explicitness are highly subjective notions [39]. Thus, the operational definition we use may not match the implicitness and explicitness perceived by humans in ambiguous cases, such as the use of slurs that also have a non-hateful meaning, or expressions through violent words such as "hate" or "kill". There are also instances where humans would not agree on the nature of the speech, such as "... are dumb". Thus, in reality, gold standards of what constitutes explicit and implicit hate are not always possible. It is important that the readers bear in mind that our definition of explicitness is a widely used approximation of something whose ground truth does not exist, and focus on the qualitative differences in the performance metrics rather than taking them at face value.

We compared three real, widely used multi-class datasets, which enabled our findings to be applicable to studies that have used these datasets, but also extensible to other scenarios with datasets with similar characteristics. From the perspective of experimental study, confounders, such as the size of data, can be better controlled through subsampling [43]. Nonetheless, we expect data size to have had limited influence on our findings, as previous research in abusive language detection showed that its effect is limited compared to other factors [28,46].

Because of our focus on a more fine-grained three-way classification task as opposed to the widely used binary formulation of abusive language detection, and keeping a consistent definition of implicitness across datasets to enable cross-dataset experiments, we could not make full use of some very relevant human annotations. We encourage future research to empirically study human-annotated implicitness in the original *AbuseEval* [5] and datasets on distinguishing abusive and non-abusive use of swearing [60] and implied statements of implicit hate [61]. Some binary datasets also have hierarchical sub-categories, such as targeted or untargeted [51,62], group-directed or individual directed

[51,63]. In principle, a similar cross-dataset empirical analysis can be applied to these hierarchical labels where categories match. When sub-categories do not perfectly align across datasets, cross-dataset experiments can be enabled by formulating overlapping sub-categories as a binary task, such as "Sexual Harassment & Threats of Violence" in [63] and "Sexual Violence" in [64].

Finally, our findings highlight the challenge of implicit expressions –compared to explicit ones–, and abuse –compared to offensive language–, which are not always reflected in the overall performance metrics. However, these are the instances that carry the most practical implications. We thus provide suggestions on sampling and annotation for future dataset construction. Additionally, future research that build abusive language detection models to optimise, carefully investigate model performance, or motivate model designs considering also the most challenging instances.

CRediT authorship contribution statement

Wenjie Yin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Arkaitz Zubiaga:** Conceptualization, Validation, Writing – review & editing, Visualization, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Comput. Surv. 51 (4) (2018) 1–30, http://dx.doi.org/10.1145/3232676, URL http://dl.acm.org/citation.cfm?doid=3236632.3232676.

[2] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Lang. Resour. Eval. (2020) http://dx.doi.org/10.1007/s10579-020-09502-8.

[3] M. Banko, B. MacKeen, L. Ray, A unified taxonomy of harmful content, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 125–137, http://dx.doi.org/10.18653/v1/2020.alw-1.16, URL https://www.aclweb.org/anthology/2020.alw-1.16.

[4] C. Nobata, J.R. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, B.Y. Zhao (Eds.), Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, ACM, 2016, pp. 145–153, http://dx.doi.org/10.1145/2872427.2883062.

[5] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, M. Granitzer, 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, European Language Resources Association, Marseille, France, 2020, pp. 6193–6202, URL https://www.aclweb.org/anthology/2020.lrec-1.760.

[6] Z. Waseem, D. Hovy, Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93, http://dx.doi.org/10.18653/v1/N16-2013, URL https://www.aclweb.org/anthology/N16-2013.

[7] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, 2017, arXiv:1703.04009 [cs]. arXiv:1703.04009. URL http://arxiv.org/abs/1703.04009.

[8] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of Twitter abusive behavior, in: Proceedings of ICWSM, AAAI Press, 2018.

[9] S. Sharma, S. Agrawal, M. Shrivastava, Degree based classification of harmful speech using Twitter data, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 106–112, URL https://www.aclweb.org/anthology/W18-4413.

[10] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, R. Tromble, Introducing CAD: the contextual abuse dataset, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2289–2303.

[11] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, Proc. ACM Hum.-Comput. Interact. 1 (CSCW) (2017) 1–22.

[12] U. Kursuncu, M. Gaur, C. Castillo, A. Alambo, K. Thirunarayan, V. Shalin, D. Achilov, I.B. Arpinar, A. Sheth, Modeling islamist extremist communications on social media using contextual dimensions: religion, ideology, and hate, Proc. ACM Hum.-Comput. Interact. 3 (CSCW) (2019) 1–22.

[13] B. Mathew, A. Illendula, P. Saha, S. Sarkar, P. Goyal, A. Mukherjee, Hate begets hate: A temporal study of hate speech, Proc. ACM Hum.-Comput. Interact. 4 (CSCW2) (2020) 1–24.

[14] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10, http://dx.doi.org/10.18653/v1/W17-1101, URL https://www.aclweb.org/anthology/W17-1101.

[15] A. Arango, J. Pérez, B. Poblete, Hate speech detection is not as easy as you may think: A closer look at model validation (extended version), Inf. Syst. (2020) 101584, http://dx.doi.org/10.1016/j.is.2020.101584, URL http://www.sciencedirect.com/science/article/pii/S0306437920300715.

[16] X. Han, Y. Tsvetkov, Fortifying toxic speech detectors against veiled toxicity, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7732–7739, http://dx.doi.org/10.18653/v1/2020.emnlp-main.622, URL https://aclanthology.org/2020.emnlp-main.622.

[17] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, N. Asokan, All you need is" Love" evading hate speech detection, in: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, 2018, pp. 2–12.

[18] Z. Zhang, L. Luo, Hate speech detection: A solved problem? The challenging case of long tail on Twitter, 2018, arXiv:1803.03662 [cs]. arXiv:1803.03662. URL http://arxiv.org/abs/1803.03662.

[19] J. Qian, M. ElSherief, E. Belding, W.Y. Wang, Leveraging intra-user and inter-user representation learning for automated hate speech detection, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 118–123, http://dx.doi.org/10.18653/v1/N18-2019, URL https://www.aclweb.org/anthology/N18-2019.

[20] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F.M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63, http://dx.doi.org/10.18653/v1/S19-2007, URL https://www.aclweb.org/anthology/S19-2007.

[21] M. Mozafari, R. Farahbakhsh, N. Crespi, A BERT-based transfer learning approach for hate speech detection in online social media, in: H. Cherifi, S. Gaito, J.F. Mendes, E. Moro, L.M. Rocha (Eds.), Complex Networks and their Applications VIII, in: Studies in Computational Intelligence, Springer International Publishing, Cham, 2020, pp. 928–940, http://dx.doi.org/10.1007/978-3-030-36687-2_77.

[22] P. Badjatiya, M. Gupta, V. Varma, Stereotypical bias removal for hate speech detection task using knowledge-based generalizations, in: L. Liu, R.W. White, A. Mantrach, F. Silvestri, J.J. McAuley, R. Baeza-Yates, L. Zia (Eds.), The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 49–59, http://dx.doi.org/10.1145/3308558.3313504.

[23] M. Sap, D. Card, S. Gabriel, Y. Choi, N.A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678, http://dx.doi.org/10.18653/v1/P19-1163, URL https://www.aclweb.org/anthology/P19-1163.

[24] N. Duarte, E. Llanso, A. Loup, Mixed messages? The limits of automated social media content analysis, in: Conference on Fairness, Accountability and Transparency, 2018, p. 106, URL http://proceedings.mlr.press/v81/duarte18a.html.

[25] T. Davidson, D. Bhattacharya, I. Weber, Racial bias in hate speech and abusive language detection datasets, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 25–35, http://dx.doi.org/10.18653/v1/W19-3504, URL https://www.aclweb.org/anthology/W19-3504.

[26] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 602–608, http://dx.doi.org/10.18653/v1/N19-1060, URL https://www.aclweb.org/anthology/N19-1060.

[27] S.D. Swamy, A. Jamatia, B. Gambäck, Studying generalisability across abusive language detection datasets, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 940–950, http://dx.doi.org/10.18653/v1/K19-1088, URL https://www.aclweb.org/anthology/K19-1088.

[28] P. Fortuna, J. Soler-Company, L. Wanner, How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Inf. Process. Manage. 58 (3) (2021) 102524, http://dx.doi.org/10.1016/j.ipm.2021.102524, URL https://www.sciencedirect.com/science/article/pii/S0306457321000339.

[29] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ Comput. Sci. 7 (2021) e598.

[30] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on Twitter using a convolution-GRU based deep neural network, in: A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 745–760, http://dx.doi.org/10.1007/978-3-319-93417-4_48.

[31] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, CoRR abs/1810.04805. arXiv:1810.04805. URL http://arxiv.org/abs/1810.04805.

[32] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N.A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5477–5490, http://dx.doi.org/10.18653/v1/2020.acl-main.486, URL https://www.aclweb.org/anthology/2020.acl-main.486.

[33] P. Mishra, H. Yannakoudakis, E. Shutova, Tackling online abuse: A survey of automated abuse detection methods, 2019, arXiv:1908.06024 [cs]. arXiv:1908.06024. URL http://arxiv.org/abs/1908.06024.

[34] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, H. Margetts, Challenges and frontiers in abusive content detection, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 80–93, http://dx.doi.org/10.18653/v1/W19-3509, URL https://www.aclweb.org/anthology/W19-3509.

[35] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 78–84, http://dx.doi.org/10.18653/v1/W17-3012, URL https://www.aclweb.org/anthology/W17-3012.

[36] B. van Aken, J. Risch, R. Krestel, A. Löser, Challenges for toxic comment classification: An in-depth error analysis, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 33–42, http://dx.doi.org/10.18653/v1/W18-5105, URL https://www.aclweb.org/anthology/W18-5105.

[37] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PLOS ONE 14 (8) (2019) e0221152, http://dx.doi.org/10.1371/journal.pone.0221152, Publisher: Public Library of Science. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221152.

[38] M. Wiegand, J. Ruppenhofer, E. Eder, 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, Association for Computational Linguistics, Online, 2021, pp. 576–587, http://dx.doi.org/10.18653/v1/2021.naacl-main.48, URL https://aclanthology.org/2021.naacl-main.48.

[39] B. Vidgen, S. Hale, E. Guest, H. Margetts, D. Broniatowski, Z. Waseem, A. Botelho, M. Hall, R. Tromble, Detecting East Asian prejudice on social media, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 162–172, http://dx.doi.org/10.18653/v1/2020.alw-1.19, URL https://www.aclweb.org/anthology/2020.alw-1.19.

[40] S. Malmasi, M. Zampieri, Detecting hate speech in social media, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 467–472, http://dx.doi.org/10.26615/978-954-452-049-6_062.

[41] L. Gao, A. Kuppersmith, R. Huang, Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 774–782, URL https://www.aclweb.org/anthology/I17-1078.

[42] L. Graumas, R. David, T. Caselli, Twitter-based polarised embeddings for abusive language detection, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 1–7, http://dx.doi.org/10.1109/ACIIW.2019.8925049.

[43] D. Razo, S. Kübler, Investigating sampling bias in abusive language detection, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 70–78, http://dx.doi.org/10.18653/v1/2020.alw-1.9, URL https://www.aclweb.org/anthology/2020.alw-1.9.

[44] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.

[45] E.W. Pamungkas, V. Basile, V. Patti, Misogyny detection in Twitter: a multilingual and cross-domain study, Inf. Process. Manage. 57 (6) (2020) 102360, http://dx.doi.org/10.1016/j.ipm.2020.102360, URL http://www.sciencedirect.com/science/article/pii/S0306457320308554.

[46] I. Nejadgholi, S. Kiritchenko, On cross-dataset generalization in automatic detection of online abuse, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 173–183, http://dx.doi.org/10.18653/v1/2020.alw-1.20, URL https://www.aclweb.org/anthology/2020.alw-1.20.

[47] E.W. Pamungkas, V. Patti, Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 363–370, http://dx.doi.org/10.18653/v1/P19-2051, URL https://www.aclweb.org/anthology/P19-2051.

[48] M. Karan, J. Šnajder, Cross-domain detection of abusive language online, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 132–137, http://dx.doi.org/10.18653/v1/W18-5117, URL https://www.aclweb.org/anthology/W18-5117.

[49] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1415–1420, http://dx.doi.org/10.18653/v1/N19-1144, URL https://www.aclweb.org/anthology/N19-1144.

[50] T. Pedersen, Duluth at SemEval-2019 task 6: Lexical approaches to identify and categorize offensive tweets, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 593–599.

[51] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86, http://dx.doi.org/10.18653/v1/S19-2010, URL https://www.aclweb.org/anthology/S19-2010.

[52] P. Fortuna, J. Soler, L. Wanner, Toxic, hateful, offensive or abusive? What are we really classifying? an empirical analysis of hate speech datasets, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6786–6794, URL https://www.aclweb.org/anthology/2020.lrec-1.838.

[53] B. Vidgen, L. Derczynski, Directions in abusive language training data: Garbage in, garbage out, 2020, arXiv:2004.01670 [cs]. arXiv:2004.01670. URL http://arxiv.org/abs/2004.01670.

[54] Z. Waseem, Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter, in: Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas, 2016, pp. 138–142, http://dx.doi.org/10.18653/v1/W16-5618, URL https://www.aclweb.org/anthology/W16-5618.

[55] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1425–1447, URL https://www.aclweb.org/anthology/2020.semeval-1.188.

[56] E. Fersini, D. Nozza, P. Rosso, AMI @ EVALITA2020: Automatic misogyny identification, in: Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), 2020, p. 8.

[57] G. Glavaš, M. Karan, I. Vulić, XHate-999: Analyzing and detecting abusive language across domains and languages, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6350–6365, http://dx.doi.org/10.18653/v1/2020.coling-main.559, URL https://www.aclweb.org/anthology/2020.coling-main.559.

[58] O. de Gibert, N. Perez, A. García-Pablos, M. Cuadros, Hate speech dataset from a white supremacy forum, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 11–20, http://dx.doi.org/10.18653/v1/W18-5102, URL https://www.aclweb.org/anthology/W18-5102.

[59] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, P. Nakov, SOLID: A large-scale semi-supervised dataset for offensive language identification, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 915–928.

[60] E.W. Pamungkas, V. Basile, V. Patti, Do you really want to hurt me? Predicting abusive swearing in social media, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6237–6246, URL https://www.aclweb.org/anthology/2020.lrec-1.765.

[61] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, D. Yang, Latent hatred: A benchmark for understanding implicit hate speech, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 345–363, http://dx.doi.org/10.18653/v1/2021.emnlp-main.29, URL https://aclanthology.org/2021.emnlp-main.29.

[62] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-European languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, in: FIRE '19, Association for Computing Machinery, Kolkata, India, 2019, pp. 14–17, http://dx.doi.org/10.1145/3368567.3368584.

[63] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at IberEval 2018, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), 2018, pp. 214–228.

[64] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Proces. Leng. Nat. 67 (2021) 195–207.