Abusive Language on Social Media Through the Legal Looking Glass

Thales Bertaglia[1,3], Andreea Grigoriu[2], Michel Dumontier [1], and Gijs van Dijck [2]

{t.costabertaglia,a.grigoriu,michel.dumontier,gijs.vandijck}@maastrichtuniversity.nl
[1]Institute of Data Science, Maastricht, The Netherlands
[2]Maastricht Law and Tech Lab, Maastricht, The Netherlands
[3]Studio Europa, Maastricht, The Netherlands

Abstract

Abusive language is a growing phenomenon on social media platforms. Its effects can reach beyond the online context, contributing to mental or emotional stress on users. Automatic tools for detecting abuse can alleviate the issue. In practice, developing automated methods to detect abusive language relies on good quality data. However, there is currently a lack of standards for creating datasets in the field. These standards include definitions of what is considered abusive language, annotation guidelines and reporting on the process. This paper introduces an annotation framework inspired by legal concepts to define abusive language in the context of online harassment. The framework uses a 7-point Likert scale for labelling instead of class labels. We also present ALYT – a dataset of **A**busive **L**anguage on **Y**ou**T**ube. ALYT includes YouTube comments in English extracted from videos on different controversial topics and labelled by Law students. The comments were sampled from the actual collected data, without artificial methods for increasing the abusive content. The paper describes the annotation process thoroughly, including all its guidelines and training steps.

1 Introduction

The increased use of social media can worsen the issue of online harassment. Nowadays, more than half of online harassment cases happen on social media platforms (Center, 2017). A specific popular form of online harassment is the use of abusive language. One abusive or toxic statement is being sent every 30 seconds across the globe[1]. The use of abusive language on social media contributes to mental or emotional stress, with one in ten people developing such issues (Center, 2017).

Automatic detection tools for detecting abusive language are used for combating online harassment. These tools are mainly based on machine learning algorithms that rely on training data. Therefore, there is a need for good quality datasets to create high performing algorithms to alleviate online harassment. There are various datasets in the field of online harassment research. However, there is a lack of standards for developing these resources. These standards include the definitions used to determine what content is abusive and the steps of the annotation process (including the annotators). The lack of standards leads to conflicting definitions, which ultimately results in disagreement within the field regarding which tasks to solve, creating annotation guidelines, and terminology.

Our Contribution In this project, we introduce ALYT – a dataset of 20k YouTube comments in English labelled for abusive language. The dataset and its data statement are available online[2]. We manually selected videos focusing on a range of controversial topics and included different video types. Rather than artificially balancing the data for abusive content, we randomly sampled the collected data. We developed an annotation framework inspired by legal definitions by analysing various European provisions and case law ranging from insults, defamation and incitement to hatred. Instead of class labels, we use a 7-point Likert scale to encapsulate the complexity of the labelling decisions. We analyse the n-grams in our corpus to characterise its content and understand the abusive language's nature. The results show that the dataset contains diverse topics, targets, and expressions of abuse.

2 Related Work

Creating dataset for online harassment research, including abusive language, has been a challeng-

[1]https://decoders.amnesty.org/projects/troll-patrol/findings. For all links, the content refers to the page version last accessed on 8 June 2021.

[2]https://github.com/thalesbertaglia/ALYT

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 191–200

ing task. Vidgen and Derczynski (2020) review a wide range of datasets – and their creation process – within the field and identify many issues. Various approaches have been explored over the years, including different data collection strategies, labelling methodologies and employing different views and definitions of online harassment.

In terms of annotation methods, crowdsourcing is a popular option for labelling abusive language data (Burnap and Williams, 2015; Zhong et al., 2016; Chatzakou et al., 2017; Ribeiro et al., 2017; Zampieri et al., 2019). However, in some instances, a small group of non-experts in harassment research (Bretschneider et al., 2014; Mathew et al., 2018; van Rosendaal et al., 2020) or domain experts annotate the data (Golbeck et al., 2017; Waseem and Hovy, 2016). The definitions used to label the data can vary as well. At times, definitions are derived from literature (Chatzakou et al., 2017) on the topic, or existing social media platform's guidelines (Ribeiro et al., 2017). In other instances, annotators decide by themselves when abuse is present in the text (Walker et al., 2012).

A recent direction in the field has been applying legal provisions to decide whether content should be removed, given criminal law provisions on hate speech or incitement to hatred. These approaches represent legal provisions as decision trees that guide the annotation process. Zufall et al. (2020) apply this methodology focusing on the German provision related to incitement to hatred. Two non-expert annotators label the data, guided by the created decision tree. The experiments show that there was little difference between using expert and non-expert annotators in this case.

3 Data Collection

We aimed to include a representative sample of abusive language on social media in our dataset. Therefore, we did not search directly for abusive content. Instead, we chose topics likely to contain abusive comments. We chose three different topics before the video selection: Gender Identity (GI), Veganism (VG), and Workplace Diversity (WD). The topics generate controversial videos on YouTube while not being limited to one type of controversy (e.g. gender identity, diet choices, socio-economical issues). The videos in GI focus on the disclosure of transgender identity and the impact of transgender people in sports. The videos in the VG category concentrate on describing the vegan movement and influencers deciding to become vegan. In WD, the videos illustrate the gender wage gap and its implications.

We searched for content in one language; therefore, the videos and majority of the comments are in English. We manually searched for videos using the topics as keywords. We selected popular videos (considering the number of views) made by well-known influencers posting controversial content. We included three types of videos: personal videos (posted by influencers on the topic), reaction videos (videos in which the author reacts to another video) and official videos (posted by news and media channels).

To create our dataset, we retrieved all comments from the selected videos, excluding replies. We removed comments containing URLs because these are often spam or make reference to external content. We also removed comments with fewer than three tokens. In total, we obtained 879,000 comments after these steps. Out of this sample, we selected 20,215 to annotate. We randomly sampled comments from the total distribution, not attempting to balance the data according to content. We aimed to balance video topics and types equally, but as the total number of comments was not even per video category, the final sample was not perfectly balanced. Table 1 shows the distribution per video category of the comments included in the dataset.

Category	%	#
VG	34.75	6967
GI	34.46	7024
WD	30.79	6224
Official	50.31	10171
Personal	31.38	6343
Reaction	18.31	3701

Table 1: Distribution of comments per video category

Collecting Abusive Content

Searching for keywords related to harassment is a common approach to increase the amount of abusive content in datasets. We do not employ any method to balance the data artificially – i.e., we do not try to search for abusive content directly. Instead, we randomly select comments from the total distribution of comments, resulting in a realistic data sample, similar to what is available on

the platform. To compare our sampling approach to keyword search, we conduct two experiments comparing our dataset to others. First, we compare the final distribution of abusive content. Then, we compare the prevalence of hateful keywords. We use Hatebase[3] as a source of keywords, limiting it to the 500 most frequent terms (by the number of sightings).

We compare our dataset to three others, all containing tweets: Davidson et al. (2017) (HSOL), Waseem and Hovy (2016) (HSHP), and Zampieri et al. (2019) (OLID). Twitter is the most popular social media platform for online harassment research, so most datasets contain tweets. HSHP is distributed as tweet ids, so all experiments refer to the distribution of the tweets we were able to retrieve in April 2021. These datasets use different definitions of abusive content. To harmonise the definitions and compare the data distributions, we consider that the following classes match our definition of *abuse*: tweets labelled as *hateful* on HSOL; *sexist* or *racist* on HSHP; and *offensive* **and** *targeted* on OLID. Table 2 presents the distribution of abusive content on each dataset.

Dataset	%	#
ALYT	11.42	2274
HSOL	5.77	1430
HSHP	25.78	2715
OLID	29.00	4089

Table 2: Distribution of abusive content in each dataset

The datasets that use hateful keyword search have a higher prevalence of hate. ALYT has a lower, but comparable, proportion of abusive content. Considering that we do not explicitly try to balance the data, our approach leads to a more representative sample of the actual scenario of social media while still having a significant amount of abusive comments. HSOL uses keywords from Hatebase to search for tweets. Davidson et al. (2017) conclude that Hatebase is imprecise and leads to a small amount of actual hate speech; therefore, this sampling approach is inefficient. HSHP and OLID use a few hateful keywords and others associated with abusive tweets, such as messages directed to political accounts and hashtags about tv shows. This approach allows increasing the amount of abusive content without biasing it to specific keywords. However, the content may still correlate to the hashtags or accounts being used to search for tweets. Our approach is similar in the sense that the video topics delimit the scope of the comments, but the comments are not filtered; thus, they provide a representative sample of the entire data. To further investigate the prevalence of hateful keywords on the datasets, we analyse the proportion of content that contains at least one term from Hatebase. Table 3 presents the results.

Dataset	%	#
ALYT	8.81	246
HSOL	87.55	1252
HSHP	7.51	204
OLID	6.43	263

Table 3: Distribution of comments containing at least one term from Hatebase

ALYT has a low prevalence of Hatebase keywords, with a distribution similar to HSHP and OLID. This result shows that the abusive content in our dataset is not limited to frequent hateful words. Therefore, not searching for specific keywords leads to higher lexical diversity. The distribution of HSOL further confirms this observation: abusive content from the dataset predominantly contains terms from Hatebase. Although this experiment is limited to a single lexicon, it provides evidence that our sampling approach does not result in abusive content defined by specific keywords. section 6 will discuss the content of the dataset in details.

4 Annotation Framework

Datasets presented in state-of-the-art research mainly use several definitions for abusive language or focus on specific phenomena – such as hate speech, racism, and sexism. What constitutes *abusive content* is often not precisely defined. Dataset creators – in an attempt to make these concepts clear – rely on various definitions, ranging from platform guidelines to dictionary entries. Our goal is to develop an annotation framework inspired by legal definitions and to define abusive content concretely in light of these concepts. Using legal definitions as inspiration can provide a consistent and stable background for deciding which content is abusive, since most occurrences of abusive language are already covered in legal systems.

[3] https://hatebase.org/

Definitions

We collected legislative resources (provisions and case law) in the context of abusive language expressed online. We focused on the European landscape by studying four countries: The Netherlands, France, Germany and the UK. These countries include both civil and common law, providing a comprehensive sample of legal traditions. The legislative sources focus both on offensive language towards an individual and towards a specific group/minority. In this project, we also focus on both types of offences.

For Germany, we selected the following provisions using the Criminal Code[4]: incitement to hatred (Article 130); insulting (Section 185); malicious gossip defined as "degrading that person or negatively affecting public opinion about that person" (Section 186); and defamation (Section 187). Similarly, for the Netherlands, using the Criminal Code[5], we included: Article 137, which focuses on incitement to hatred and general defamation (Section 261), slander (Section 262), and insults (Section 266). For France, we used the Press Freedom Act of 29 July 1881[6], focusing on actions such as discrimination, hate, violence (Article 24), defamation (Article 32) and insult (Article 33). In the UK, the Public Order Act 1986[7] defines offensive messages and threats, specifically in Part 3 (focusing on racial grounds) and 3A (religious and sexual orientation grounds). After selecting the sources, we harmonised the elements present in the provisions such as the targets of the attack, protected attributes (grounds on which the targets are attacked such as race, religion etc) and the harmful acts specified to be performed (such as insult, defamation). Even though the countries might have elements in common which can be easy to harmonise, we also found elements specific to some countries only (for example specifically mentioning the effect caused by the attack to the victim in the UK, such as distress and anxiety). The analysis resulted in three main abstract categories found in provisions: incitement of hatred towards specific protected groups, acts which cause distress and anxiety and acts involving public opinion such as degradement or belittlement.

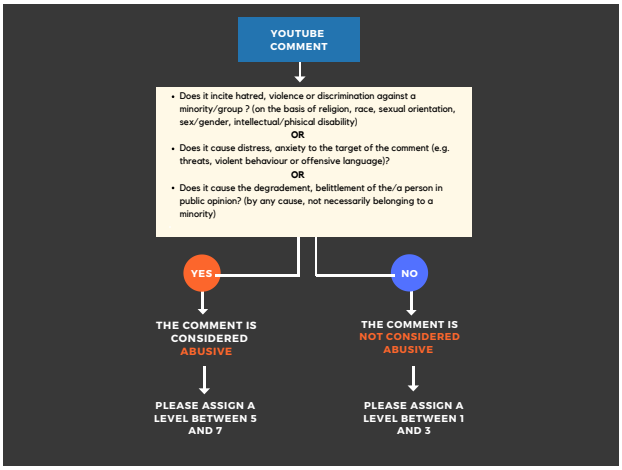

Figure 1: Annotation process diagram

We developed three questions comprising elements found in all the mentioned provisions to define *abusive language*:

- Does it incite hatred, violence or discrimination against a minority/group (on the basis of religion, race, sexual orientation, sex/gender, intellectual/physical disability)?
- Does it cause distress, anxiety to the target of the comment (e.g. threats, violent behaviour or offensive language)?
- Does it cause the degradement, belittlement of the/a person in public opinion? (by any cause, not necessarily belonging to a minority)

The annotators used these questions to determine whether a comment is abusive, as described in Figure 1. The full version of the annotation manual included examples and is available online[8].

Annotation Scale

For labelling, we used a 7-point Likert scale (Joshi et al., 2015) instead of class labels. The scale represents a mix of two features: the intensity of the abuse present in the comment and how confident the annotator is about the labelling decision. Specifically, numbers from 1 to 3 represent non-abusive content, with 1 describing comments with no abusive content at all. Comments labelled with 2 or 3 might contain sarcasm or jokes that could be considered abusive. Number 4 indicates comments that fall between abusive and non-abusive – so it also

[4] https://www.gesetze-im-internet.de/stgb/

[5] https://www.legislationline.org/documents/section/criminal-codes/country/12/Netherlands/show

[6] https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000877119/

[7] https://www.legislation.gov.uk/ukpga/1986/64

[8] http://bit.ly/alyt-manual

encodes labelling uncertainty. Numbers between 5 and 7 represent abusive comments, with 7 describing clearly abusive comments. Comments labelled with 5 or 6 might contain less obvious abuse.

5 Annotation Process

Throughout the data annotation process, annotators encode our framework and apply it to label the data. A proper annotation methodology, therefore, is fundamental to ensure data quality. Yet, most works presenting abusive language datasets fail to report this process in details. We follow the corpus annotation pipeline proposed by Hovy and Lavid (2010) and thoroughly describe how we conducted its main steps when creating ALYT. To measure the reliability of the annotation, we use Krippendorf's alpha (α) with ordinal level of measurement (Krippendorff, 2011) and majority agreement to calculate the overall inter-annotator agreement. Antoine et al. (2014) highlight that α is a reliable metric for ordinal annotations, such as the 7-point scale from our framework.

Training

A team of six Law students enrolled in an European university labelled ALYT. Before the main annotation stage, we conducted a careful training phase, actively engaging in discussions with the annotators to improve the annotation framework and manual. We instructed the team to watch the videos included in the dataset before labelling the comments. We organised three training meetings with the annotators. Also, we evaluated their performance by giving an assignment sample of comments after each meeting. We annotated 50 comments together during the first meeting, aiming to familiarise the team with the task and annotation platform. The inter-annotator agreement for the first round of training was $\alpha = 45.0$.

For the second meeting, we created a sample of the comments that had the most disagreements in the previous assignment. Then, we asked the annotators to specify which questions (from our annotation framework) they used to decide whether a comment was abusive. For the second assignment sample, we also required the team to mention the questions used for each comment. This round had $\alpha = 51.2$.

In the third meeting, we decided to change the labelling process. We used a shared document for annotation, in which each annotator added their label for the comments. Then, we discussed the examples that had disagreements with the whole group. This discussion allowed us to answer a variety of questions and incorporate feedback into the annotation manual. This round achieved $\alpha = 65.8$. After this meeting, we reached a satisfactory agreement and also noticed less confusion in the annotations.

The training phase showed that the interaction with annotators is fundamental to improve the annotation process. We received feedback about the framework, improved the manual, and clarified concepts related to the labelling decisions. The improvement in inter-annotator agreement allied to our empirical observations showed that the training phase led to higher data quality.

Main Annotation Phase

After the training phase, we proceeded to label the entire dataset. We randomly split the dataset into six samples, one for each annotator. A single annotator labelled each comment. Given the extensive training process, we consider that each annotator is qualified to apply the framework properly; thus, we opt not to have multiple annotations on the same comment to allow a more significant number of labelled samples in total. The annotation interface displayed the text of each comment (including emojis and special symbols) and the video id; therefore, annotators could check the video to understand the context. We sorted comments by their video ids, displaying comments belonging to the same videos in sequence to reduce context-switching during labelling. Annotators could also access a summary of the annotation framework within the platform.

We randomly selected 300 comments to be labelled by all annotators; we used this sample to calculate the inter-annotator agreement. In addition to α and majority agreement, we also compute the majority agreement for the class labels (i.e., the scale numbers grouped into three classes). This metric is relevant to verify whether annotators agree on the polarity of a comment – represented by the extremes of the scale. It also illustrates the annotation reliability for applications that would use class labels instead of the full scale. Table 4 presents the inter-annotator agreement metrics for the whole dataset and per video category. # indicates the number of comments in a given category; *Majority* shows the percentage of comments with at least four matching annotations (out of a total

of six); *Grouped* refers to majority agreement over class labels.

Topic	#	α	Majority	Grouped
All	300	73.8	64.3%	92.3%
VG	134	64.9	88.1%	98.5%
GI	135	55.2	44.4%	84.4%
WD	31	24.0	48.4%	100%
Official	76	60.0	50.0%	90.8%
Personal	179	77.9	76.5%	95.0%
Reaction	45	47.3	40.0%	84.4%

Table 4: Inter-annotator agreement metrics per category

The overall value of alpha indicates substantial agreement. The majority agreement was lower, which is expected given the 7-point scale. The grouped majority shows a high agreement about the polarity of comments, confirming that annotators agree whether a comment is abusive. There are significant differences between video categories. Disparities in sample size can lead to the difference in metrics: WD, for instance, had only 31 comments. For categories with similar size, a lower agreement can be attributed to controversial topics, confusing comments, or conflicting views. section 6 further investigates the content of each category and analyses how abuse is expressed in each one. In general, the annotation achieved significant inter-annotator agreement – which indicates that the annotation process was consistent and the dataset is reliable.

6 Dataset

The labelled dataset includes 19,915 comments, excluding the comments used in the training phase and the sample used for calculating the inter-annotator agreement. Each comment has a label ranging from 1 to 7, corresponding to the scale used in the annotation framework. We also aggregate the labels into classes: values from 1 to 3 correspond to *non-abusive* content; 5 to 7, *abusive*; and 4, *uncertain*. In this section, we analyse the aggregated labels. Table 5 presents the class distribution per category of the dataset. The percentage refers to the distribution over the specific category (video topic or type).

The annotators labelled 2274 comments as *Abusive*. This number represents 11.42% of the total distribution, showing a low prevalence of abusive content. Considering that we selected random samples and did not balance the data according to content, these comments potentially represent the actual distribution. However, since we balanced the number of comments per category, the dataset might misrepresent some video topics and types. The distribution of abusive content per category shows evidence of this imbalance. Videos about gender identity include 76.17% of the total amount of *abusive* comments and videos from an official source, 67.81%. To investigate the difference in content between categories, we analyse the lexical distribution within each topic and type.

Category	Abusive	Non-Abusive	Uncertain
Total	11.42%	85.98%	2.61%
VG	9.19%	38.02%	22.16%
GI	76.17%	28.55%	51.83%
WD	14.64%	33.44%	26.01%
Official	67.81%	47.99%	64.74%
Personal	17.46%	33.03%	21.39%
Reaction	14.73%	18.98%	13.87%

Table 5: Distribution of classes in the dataset per category

Lexical Analysis

We preprocess the comments by removing stopwords, punctuation marks, and character repetitions over three. First, we analyse the average length (in number of tokens) of comments in each class. *Abusive* comments have on average 31.67 tokens; *non-abusive*, 31.23; and *uncertain*, 41.25. Comments labelled as *uncertain* tend to be 30% longer than the other classes. However, sequences of short tokens, such as emojis, may impact the mean length. To avoid this issue, we also compute the average number of characters per comment, subtracting whitespaces. *Abusive* comments have on average 137.18 characters; *non-abusive*, 132.36; and *uncertain*, 179.02. Again, the *uncertain* class contains longer comments. These comments might be less readable and confusing, leading annotators to choose to label them as *uncertain*.

To analyse the content of the comments in depth, we identify the most frequent unigrams in the *abusive* class for each video category. Table 6 presents the ten most frequent unigrams.

In general, slurs and hateful terms are not prevalent among the most frequent unigrams. Each topic contains words related to videos from that cate-

VG	GI	WD	Official	Personal	Reaction
vegan	girls	women	women	trans	trisha
freelee	like	men	men	like	like
meat	men	gap	girls	f*cking	trans
like	trans	work	boys	b*tch	people
eating	women	wage	like	f*ck	think
b*tch	people	less	compete	people	needs
video	boys	make	people	video	i'm
go	compete	feminists	unfair	get	b*tch
eat	transgender	get	male	trisha	video
fat	male	pay	get	i'm	even

Table 6: Ten most frequent unigrams on abusive comments per category

gory, but there is some lexical overlap. *Veganism* includes neutral terms (meat, eat, vegan) and some derogatory words (fat, b*itch). The second most common unigram, Freelee, refers to a popular YouTuber – which shows that the *abusive* comments may target a specific person. *Gender Identity* and *Workplace Diversity* contain many gender-related words, which potentially occur in sexist comments.

For video types, *Personal* and *Reaction* have similar distributions. Personal has a higher prevalence of offensive words, and both include "Trisha" (a YouTuber) – indicating targeted comments. The dataset has both a video by Trisha and a reaction video to it, so mentions about the YouTuber are expected. Unigrams from *Official* videos are primarily about the video topics, following a pattern analogous to the topics of GI and WD.

Unigram distributions enable the identification of potentially relevant keywords related to abusive content. Understanding how abusive comments are expressed, however, requires more context. Therefore, we also identify the most frequent trigrams for each class to examine larger language constructs. We exclude trigrams consisting entirely of sequences of symbols or emojis. Many trigrams had the same frequency, so we highlight a sample of the total for each category. For the topic of *Veganism*, frequent trigrams include "*freele shut f*ck*", "*b*tch going vegan*", and "*vegans hate asians*". The first two phrases confirm that some abusive comments target content creators. *Gender Identity* contains "*boys competing girls*", "*make trans league*", and "*natural born gender*". The video with the most comments on this topic is about transgender athletes in sports – and these trigrams expose the prevalence of discriminatory comments against them. *Workplace Diversity* includes "*gender wage gap*", "*work long hours*", and "*take care children*". Interestingly, "*work less hours*" is also among the most frequent phrases, which indicates that the topic is controversial. Trigrams such as "*take care children*" show that comments about WD often express sexism.

Official videos, in general, combine trigrams from GI and WD. "*compelling argument sides*" and "*men better women*" are among the most frequent phrases; the former shows that comments contain civilised discussion; the latter, however, indicates the predominance of sexism. While the unigram distributions of *Personal* and *Reaction* videos are similar, their trigram frequencies exhibit different patterns. *Personal* includes "*identify natural born*", "*b*tch going vegan*", and "*whole trans community*", showing a combination of comments about GI and VG. *Reaction* displays a high prevalence of targeted comments with phrases such as "*trisha looks like*", "*trisha mentally ill*", and "*needs mental help*". Although these trigrams are the most frequent, their absolute number of occurrences is low. Therefore, lexical analysis indicates general trends about the content of comments but does not represent the entirety of abusive content in the dataset.

Classification Experiments

We perform classification experiments to explore ALYT as a source of training data for abusive language detection models. We frame the task as binary classification, using the grouped class labels *Abusive* and *Not Abusive*. We experiment with two models: logistic regression and a BERT-based classifier (Devlin et al., 2019).

The first baseline is a logistic regression clas-

Model	Class	P	R	F1
LogReg	NOT	$.914 \pm .014$	$.976 \pm .019$	$.944 \pm .003$
	ABU	$.678 \pm .081$	$.307 \pm .132$	$.395 \pm .102$
	AVG	$.796 \pm .034$	$.641 \pm .057$	$.670 \pm .050$
BERT	NOT	$.944 \pm .002$	$.952 \pm .004$	$.948 \pm .001$
	ABU	$.588 \pm .013$	$.546 \pm .017$	$.566 \pm .006$
	AVG	$.766 \pm .006$	$.749 \pm .007$	$\mathbf{.757} \pm .003$

Table 7: Results for abusive language detection

sifier trained on word n-grams ranging from 1 to 3. We preprocessed all comments using the steps described in section 6. We used the implementation from scikit-learn with default hyperparameters (Pedregosa et al., 2011). We trained the model using 5-fold cross-validation and report the metrics averaged over the folds, along with standard deviation.

The second baseline is a BERT model pretrained on English tweets (BERTweet) (Nguyen et al., 2020). In a preliminary experiment, BERTweet outperformed BERT-base by 3 points of macro F1. In addition to this result, we chose to use BERTweet because its vocabulary is more similar to ALYT's than BERT-base. We tokenised comments using TweetTokenizer from NLTK and translated emojis into strings using the emoji[9] package. We fine-tuned BERTweet for classification by training it with ALYT. We used a learning rate of 2^{-5}, a batch size of 32, and trained the model for a single epoch to avoid overfitting. We trained the model using a 80/20 train and test split; the results are averaged over five runs.

Table 7 presents the classification results. We report Precision (P), Recall (R), and F1 for each model on all classes (Not Abusive (NOT) and Abusive (ABU)) and macro averages (AVG). Values after $\pm$ represent the standard deviation.

The BERT-based model outperformed logistic regression by 8.6 points in macro F1 on average; the difference in the *Abusive* class was 17 points. Both models perform considerably worse when predicting abusive comments – which is expected given the data imbalance. Interestingly, logistic regression achieved higher precision but much lower recall than BERT. This result indicates that the classifier is making safe predictions based on surface-level patterns. To further investigate this effect, we compute the ten most relevant n-grams for the logistic regression (based on the model coefficients summed over all folds) and analyse their distribution over both classes. The top ten n-grams are *b*tch*, *dudes*, *femin*zis*, *d*ck*, *f*ck*, *idiot*, *drugs*, *f*cking*, *fair*, and *insane*. We then identify all comments that contain at least one of these terms and check their class distribution. 50.89% belong to *Not-Abusive* and 49.11% to *Abusive*. Although this percentage shows that these n-grams discriminate abusive comments above their actual distribution (11.42%), they are still frequent in non-abusive contexts. Therefore, the logistic regression classifier relies on lexical clues and fails to capture context. In conclusion, the higher recall that BERT achieves shows it can capture higher-level features.

7 Conclusion

This paper presented a dataset of YouTube comments in English, labelled as abusive by law students, using a 7-point Likert scale. The comments were collected from videos on three controversial topics: *Gender Identity*, *Veganism*, and *Workplace Diversity*. The dataset includes a sample of the actual amount of extracted comments, without any artificial balancing of the abusive content distribution.

We developed an annotation framework that includes legally inspired labelling rules based on European provisions and case law. Our annotation process includes developing and refining guidelines through various training sessions with active discussions. Our data sampling analysis shows that not purposefully searching for abusive content still leads to a considerable amount of abusive comments, while maintaining the characteristics of the social media platform's data distribution.

The content analyses show that ALYT contains various expressions of abuse, ranging from different topics and targets. The abusive content is not limited to specific keywords or slurs associated

[9] https://pypi.org/project/emoji/

with hateful content. Using a scale to label the content has the potential to capture multiple nuances of abusive language. However, we did not explore the implications of using a scale versus binary labels in this paper. This comparison might be a relevant research direction for future work.

We believe ALYT can be a valuable resource for training machine learning algorithms for abusive language detection and understanding online abuse on social media. Our annotation framework is a significant contribution toward the standardisation of practices in the field.

References

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559.

Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. 2014. Detecting online harassment in social networks. In *Proceedings of the ICIS 2014 conference*.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Pew Research Center. 2017. Online harassment 2017.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *Current Journal of Applied Science and Technology*, pages 396–403.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. 2011. *Annenberg School for Communication Departmental Papers: Philadelphia.*

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2017. " like sheep among wolves": Characterizing hateful users on twitter. *arXiv preprint arXiv:1801.00317*.

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

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 812, page 817. Istanbul.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

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

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958.

Frederike Zufall, Huangpan Zhang, Katharina Kloppenborg, and Torsten Zesch. 2020. Operationalizing the legal concept of'incitement to hatred'as an nlp task. *arXiv preprint arXiv:2004.03422*.