*Natural Language Engineering* (2022), **28**, pp. 141–166
doi:10.1017/S135132492000056X

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

# Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?

Gilles Jacobs* Cynthia Van Hee and Véronique Hoste

Language & Translation Technology Team (LT3), Translating, Interpreting & Communications Department, Ghent University, Ghent, Belgium
*Corresponding author. E-mail: gilles.jacobs@ugent.be

(Received 22 February 2019; revised 28 September 2020; accepted 5 October 2020; first published online 18 November 2020)

**Abstract**

Successful prevention of cyberbullying depends on the adequate detection of harmful messages. Given the impossibility of human moderation on the Social Web, intelligent systems are required to identify clues of cyberbullying automatically. Much work on cyberbullying detection focuses on detecting abusive language without analyzing the severity of the event nor the participants involved. Automatic analysis of participant roles in cyberbullying traces enables targeted bullying prevention strategies. In this paper, we aim to automatically detect different participant roles involved in textual cyberbullying traces, including bullies, victims, and bystanders. We describe the construction of two cyberbullying corpora (a Dutch and English corpus) that were both manually annotated with bullying types and participant roles and we perform a series of multiclass classification experiments to determine the feasibility of text-based cyberbullying participant role detection. The representative datasets present a data imbalance problem for which we investigate feature filtering and data resampling as skew mitigation techniques. We investigate the performance of feature-engineered single and ensemble classifier setups as well as transformer-based pretrained language models (PLMs). Cross-validation experiments revealed promising results for the detection of cyberbullying roles using PLM fine-tuning techniques, with the best classifier for English (RoBERTa) yielding a macro-averaged $F_1$-score of 55.84%, and the best one for Dutch (RobBERT) yielding an $F_1$-score of 56.73%. Experiment replication data and source code are available at https://osf.io/nb2r3.

**Keywords:** Cyberbullying; Bullying participants; Text classification; Social media text

## 1. Introduction

Cyberbullying is a prevalent issue that comes with the rise of internet-based mass communication. Preventing and understanding cyberbullying requires automated detection of bullying and analysis of bullying situations.

Web 2.0 has a substantial impact on communication and relationships in today's society. Adolescents spend a substantial amount of time online, and more specifically on social networking sites (SNSs). Although most of teenagers' internet use is harmless, the freedom and anonymity experienced online makes young people vulnerable with cyberbullying being one of the major threats (Livingstone *et al.* 2014). Despite multiple (national and international) anti-bullying initiatives that have been launched to increase children's online safety (e.g., KiVa Salmivalli, Kärnä, and Poskiparta 2011a, ClickSafe Childfocus 2018, *Non au harcèlement* Ministre de l'ducation nationale 2018), much undesirable and hurtful content remains online. Tokunaga (2010) analyzed a body of quantitative research on cyberbullying and observed cybervictimization rates among teenagers between 20% and 40%. A study among 2000 Flemish secondary school students (age 12–18 years) revealed that 11% of them had been bullied online at least once in the 6 months preceding the survey (Van Cleemput *et al.* 2013). The 2014 large-scale EU Kids Online Report (EU Kids Online 2014) stated that 20% of 11–16-year-olds had been exposed to online hate messages that year, and according to a recent report by the European Commission, European Schoolnet and EU Kids Online cyberbullying remains one of the most prevalent risks as reported to helplines (O'Neill and Dinh 2018). While there is increased awareness concerning cyberbullying and its consequences, the large amount of online content generated each day makes manual monitoring practically impossible. Therefore, SNSs have a need for automated detection of harmful content by means of text and image mining techniques allowing administrators to remove content, block users, or take legal action.

Bullying episodes are complex social interactions and their psychosocial dimensions have been studied extensively in the social sciences by means of surveys (e.g., Kaltiala-Heino *et al.* 1999; Nansel *et al.* 2001; Klein, Cornell, and Konold 2012). Time-consuming surveys in schools are the typical method for data collection which usually result in small sample sizes and short textual descriptions. Xu *et al.* (2012) note that natural language processing methods have the ability to automatically analyze more data and therefore can more accurately capture the prevalence and impact of the incidence. Moreover, analyzing the roles of participants in bullying enables prevention strategies: DeSmet *et al.* (2012) discuss how identifying and mobilizing bystander participants that defend the victim have a positive impact on bullying prevention. In short, much of previous work on automated cyberbullying detection can more accurately be called "verbal aggression detection," as the conceptualization of cyberbullying is generally limited to aggressive interactions. These studies do not include cyberbullying traces that do not stem from anyone other than the bully. Automated processing of participant roles goes beyond this limited conceptualization and provides more granular insights in cyberbullying contexts, which are necessary for effective prevention.

In this research, we approach cyberbullying detection by identifying different *participant roles* in a cyberbullying event to distinguish between bullies, victims, and bystanders. Fine-grained annotation guidelines (Van Hee *et al.* 2015c) were developed to enable the annotation of participant roles in cyberbullying. While much of the related research focuses on detecting bully "attacks" in cyberbullying, the present study is the first to classify participant roles based on a representative real-world dataset that is manually annotated. Given the high skew of the dataset, both at the level of cyberbullying instances and at the level of the roles, we can present satisfactory results with the best classifier configuration yielding a macro-averaged $F_1$-score of 55.84% for English and 56.73% for Dutch.

In the remainder of this paper, we discuss related research on the conceptualization of cyberbullying, the definition of participant roles, and detection approaches in Section 2. Section 3 presents the corpus collection and annotation guidelines together with the results of an inter-annotator agreement (IAA) study to demonstrate the validity of the guidelines. In Section 4, we present the setup and optimization steps of the linear classification experiments. The pretrained language models (PLMs) are described in Section 5. Section 6 presents a general overview of the classification results obtained in the two experimental setups for both languages and we provide a results discussion and comparison between the two corpora in Section 7. Section 8 recapitulates our main findings and presents some prospects for future work.

## 2. Related research

As shown by scholars such as Cowie (2013) and Price and Dalgleish (2010), the negative effects of cyberbullying include a lower self-esteem, worse academic achievement, feelings like sadness, anger, fear, depression, and—in extreme cases—cyberbullying could lead to self-harm and suicidal thoughts. As a response to these threats, automated cyberbullying detection has received increased

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

interest resulting in several detection systems (Dinakar *et al.* 2012; Dadvar 2014; Van Hee *et al.* 2015b; Chen, Mckeever, and Delany 2017) and sociological studies investigating the desirability of online monitoring tools (Tucker 2010; Van Royen, Poels, and Vandebosch 2016). In fact, social media users, and specifically teenagers, highly value their privacy and autonomy on social media platforms and underline that priorities must be set related to the detection of harmful content (Van Royen *et al.* 2016). This has important consequences for developers of automatic monitoring systems, as it means that such systems should enable to balance protection and autonomy and hence should be optimized for precision so that cyberbullying is not flagged more than necessary. The following subsections provide theoretical background related to the definition and conceptualization of cyberbullying, the analysis of the participant roles, and automatic cyberbullying detection using text mining techniques.

### 2.1 Working definition of cyberbullying

A common starting point for conceptualizing cyberbullying are definitions of traditional (or *offline*) bullying. Seminal work has been published by Olweus (1993) and Salmivalli *et al.* (1996), who describe bullying based on three main criteria, including (i) intention, that is a bully intends to inflict harm on the victim, (ii) repetition, that is bullying acts take place repeatedly over time, and (iii) a power imbalance between the bully and the victim, that is a more powerful bully attacks a less powerful victim. The same criteria are often used to define cyberbullying. Smith *et al.* (2008), p. 376, for instance, proposed the popular definition of cyberbullying, identifying the phenomenon as "an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who cannot easily defend him or herself."

Other studies, however, have questioned the relevance of the three criteria to define online bullying based on theoretical objections and practical limitations. Firstly, while Olweus (1993) claims intention to be inherent to bullying, this is hard to ascertain in online conversations, which lack the signals of a face-to-face interaction like intonation, facial expressions, and gestures. In other words, the receiver might get the unintended impression of being offended or ridiculed (Vandebosch and Van Cleemput 2009). Another criterion that would not hold in online bullying is the power imbalance between bully and victim. In real life, this could mean that the bully is larger, stronger, or older than the victim, but it is difficult to measure online, where it can be related to technological skills, anonymity, or the inability to escape the bullying (Slonje and Smith 2008). Anonymity and persistence of content are empowering characteristics of the web for the bully: once defamatory or confidential information is posted online, it is hard—if not impossible—to remove. Finally, Dooley and Cross (2009) claim that repetition in cyberbullying is problematic to operationalize, as it is hard to estimate the consequences of a single derogatory message on a public page. Even a single act of aggression or humiliation may result in continued distress and humiliation for the victim if it is shared, liked, or read by a large audience (Dooley and Cross 2009).

The above paragraphs demonstrate that defining cyberbullying is far from trivial, and varying prevalence rates (cf. supra) show that a univocal definition of the phenomenon is still lacking in the literature. Starting from existing conceptualizations, we define cyberbullying as *content that is published online by an individual and that is aggressive or hurtful against a victim*. The motivation for a rather broad definition of cyberbullying is twofold: first, as mentioned earlier, quantifying repeated offense is difficult, especially in an online environment. Second, context-rich data would allow for a better delineation of cyberbullying, but severe challenges are faced given the sensitivity of the topic and strict general data protection regulations.

Based on this working definition, however, Van Hee *et al.* (2015c) developed a fine-grained annotation scheme to signal textual characteristics of cyberbullying that can be considered specific forms of bullying and identify different participant roles in cyberbullying.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

### 2.2 Participant roles in bullying situations

In order to capture the complex social interactions involved in bullying events, we set out to automatically detect and classify bullying utterances as belonging to a specific participant role in the bullying interaction. *Participant roles* conceptualize typical behavior patterns in bullying situations as social roles. Salmivalli (1999) were among the first to define bullying in these terms. They distinguish six roles: the *victims* (who are the target of repeated harassment), the *bullies* (who are the initiative-taking, active perpetrators), the *assistants of the bully* (who encourage the bullying), the *reinforcers of the bully* (who reinforce the bullying), the *defenders* (who comfort the victim, take his/her side, or try to stop the bullying), and the *outsiders* (who distance themselves from the situation). Their seminal work on bullying relied on surveys with children at schools in real-life bullying situations, but the conceptual framework and sociometrics used in these are later applied to cyberbullying (Salmivalli and Pöyhönen 2012).

Although traditional studies on bullying have mainly concentrated on bullies and victims, the importance of bystanders in a bullying episode has also been acknowledged (Salmivalli 2010; Bastiaensens *et al.* 2014). When it comes to prevention many people can take an active role to intervene, especially bystanders. Bystanders can support the victim and mitigate the negative effects caused by the bullying (Salmivalli 2010), especially on SNSs, where they have shown to hold higher intentions to help the victim than in real-life conversations (Bastiaensens *et al.* 2015). While Salmivalli *et al.* (1996) distinguish four bystanders, Vandebosch *et al.* (2006) identify three main types, namely bystanders who (i) participate in the bullying, (ii) help or support the victim, and (iii) ignore the bullying.

The lack of social context in near-anonymous online interactions entails a different operationalization of author participant roles: anonymity makes it challenging to link roles to actual persons. It is impossible to ascertain the frequency and different types of roles a person takes as social behavior. The conversational structure in our dataset is simple compared to full dialogue due to the limitations of the ASKfm platform (http://ask.fm/) in which a dialogue consists of only two utterances: a question and a response (which is optional and therefore often missing). Note that users do hold longer continuous conversations by posting several question–answer pairs, but reconstructing a dialog is impossible because of anonymous posting on the platform. Due to limited dialog information, we do not rely on conversational structure but on purely textual utterances that characterize typical social roles in cyberbullying. This allows for social role detection even if little context is available and user information is anonymized.

### 2.3 Automated detection and analysis of cyberbullying

Although some studies have investigated rule-based approaches (Reynolds, Kontostathis, and Edwards 2011), the dominant approach to cyberbullying detection involves machine learning, mostly based on supervised (Dinakar, Reichart, and Lieberman 2011; Dadvar 2014) or semi-supervised learning (Nahar *et al.* 2014). The former constructs a classifier using labeled training data, whereas semi-supervised approaches rely on classifiers that are built from a small set of labeled and a large set of unlabeled instances. As cyberbullying detection essentially involves distinguishing bullying from "not bullying" posts, the problem is generally approached as a binary classification task.

A key challenge in cyberbullying research is the availability of suitable data. In recent years, only a few datasets have become publicly available for this task, such as the training sets provided by the CAW 2.0 workshop (Yin *et al.* 2009), the Twitter Bullying Traces dataset (Sui 2015), and more recently the Polish cyberbullying corpus that is made available in the framework of PolEval 2019.[a] Most studies have, therefore, constructed their own corpus from platforms that are prone to bullying content, such as YouTube (Dinakar *et al.* 2011), Formspring.com (Reynolds *et al.* 2011),

[a] http://poleval.pl/.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

ASKfm (Van Hee *et al.* 2015b),[b] Instagram (Hosseinmardi *et al.* 2016), and WhatsApp (Sprugnoli *et al.* 2018). To overcome the problem of limited data accessibility, some studies use simulated cyberbullying data obtained from carefully setup experiments (Van Hee *et al.* 2015b; Sprugnoli *et al.* 2018). Furthermore, corpora were also compiled covering related subtasks such as aggressive language, offensive language, hate speech, and other abusive language detection. Many of these datasets were compiled in the framework of shared tasks, such as the TRAC shared task on aggression identification (Kumar *et al.* 2018), the HatEval shared task on multilingual detection of hate speech against immigrants and women on Twitter (Basile *et al.* 2019), and the OffensEval shared task on offensive language identification (Zampieri *et al.* 2019b). The latter task is based on the Offensive Language Identification Dataset (OLID) (Zampieri *et al.* 2019a), which also includes the coarse-grained two-class annotation of the target of the offensive language (individual vs. group). Insults and threats targeted at individuals are often defined as cyberbulling.

Among the first studies on cyberbullying detection are Yin *et al.* (2009), Reynolds *et al.* (2011), Dinakar *et al.* (2011), who explored the predictive power of *n*-grams, part-of-speech information (e.g., first and second pronouns), and dictionary-based (i.e., profanity lexicons) information for this task. Similar features were also exploited for the detection of fine-grained cyberbullying categories (Van Hee *et al.* 2015b). Studies have also demonstrated the benefits of combining such content-based features with user-based information including previous posts, the user's age, gender, location, number of friends and followers, and so on (Dadvar 2014; Nahar *et al.* 2014; Al-Garadi, Varathan, and Ravana 2016; Chatzakou *et al.* 2017). Recently, deep neural network (DNN) models have also been applied to cyberbullying (Zhang *et al.* 2016; Zhao and Mao 2017; Agrawal and Awekar 2018); in case of comparison with traditional machine learning methods in benchmarking tasks, they currently seem to slightly outperform or perform on par with classifiers like support vector machine (SVM) and even random forest and logistic regression (LR) (Zhang *et al.* 2016; Kumar *et al.* 2018; Emmery *et al.* 2019; Basile *et al.* 2019). The current state of the art in deep-learning text classification is obtained by large-scale PLMs using the transformer architecture such as BERT (Devlin *et al.* 2019) and its derivatives. For offensive language classification in OffensEval (Zampieri *et al.* 2019b), BERT-based approaches obtained best performance out of all deep-learning techniques.

The previously discussed studies have in common that they are aimed at detecting cyberbullying or merely abusive language, without processing information about the participant roles in a cyberbullying event. We discuss the only two studies—to our knowledge—in modeling some form of participant roles:

Xu *et al.* (2012) introduced bully role labeling as a challenge for the field of natural language processing when laying out future paths for research on bullying. The authors experimented with tweets containing keywords like "bully" and "bullied" as bullying traces. They studied traces of real-life bully events, which lead them to define five participant roles: bully, victim, reporter (who reports a bully event on social media), accuser (who accuses someone of bullying), and other. They performed supervised classification on 648 role-annotated tweets. A linear-kernel SVM trained on token uni- and bigrams performed best, yielding a cross-validated accuracy of 61% (no F-score or ROC-AUC scores were given). Furthermore, they experimented with token-level labeling of participant roles where a Conditional Random Field classifier ($F_1$= 0.47) outperformed an SVM classifier ($F_1$= 0.36). Their dataset is small compared to ours and is composed using keyword searches, thus limiting the number of relevant posts (i.e., less implicit cyberbullying) and excluding a lot of negative data as well, which biases the dataset (Cheng and Wicks 2014; González-Bailón *et al.* 2014). Because our corpus is randomly crawled instead of collected by keyword search, it contains a more realistic distribution of cyberbullying posts. Moreover, all posts were manually annotated with fine-grained information on cyberbullying types and participant

[b] Both the former Formspring and ASKfm are social networking sites where users can send each other questions or respond to them.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

roles, which allowed to identify implicit forms of cyberbullying that would not appear in keyword search.

Raisi and Huang (2018) defined a weakly supervised *n*-gram-based model which includes the sender's bullying and receiver's victim scores as parameters in optimization. Their algorithm optimizes parameters for all users and *n*-gram features that characterize the tendency of each user to send and receive harassing keyphrases as well as the tendency of a keyphrase to indicate harassment. They use a human-curated list of keyphrases as indicators of cyberbullying for weak supervision. This requires that the user profiles are known and available. The authors evaluate their approach on three datasets consisting of Twitter, ASKfm, and Instagram. Anonymous user interactions in their ASKfm corpus were discarded, since they essentially use a communicative sender–receiver model as a proxy for bully and victim participant roles.

In the present research, cyberbullying is considered a complex phenomenon consisting of several types of harmful behavior and realized by different participant roles (Van Hee *et al.* 2015c). In our previous work on binary and fine-grained cyberbullying detection (Van Hee *et al.* 2015b; 2018), we did not differentiate between participant roles in cyberbullying. To our knowledge, no previous study has been done that aims to automatically predict the participation role in a cyberbullying event on a representative corpus (i.e., containing real-world data and not biased by keyword search).

## 3. Data collection and annotation

To be able to build representative models for cyberbullying, a suitable dataset is required. This section describes the construction and fine-grained annotation of two cyberbullying corpora, for English and Dutch.

### *3.1 Data collection*

A Dutch and English corpus were constructed by collecting data from the SNS ASKfm, where users can create profiles and ask or answer questions, with the option of doing so anonymously. ASKfm data typically consist of question–answer pairs published on a user's profile. The data were retrieved by crawling a number of seed profiles using the GNU Wget software in April and October 2013. The terms of service of ASKfm did not forbid automated crawling at that time. Question–answer pairs were kept together and pairs occurring on the same user page were presented as a conversational thread for annotation as to provide the annotators with as much context as possible. Basic cleaning was applied to the corpus (e.g., removal of non-ASCII characters, multiple white spaces, duplicate and mass spam posts), as well as language filtering. Although the seed profiles were chosen to be of users with Dutch and English as mother tongue, the crawled corpora both contained a fair amount of non-Dutch and non-English data. Non-English and non-Dutch posts were removed, which resulted in 113,698 and 78,387 posts for English and Dutch, respectively.

For more detail on the annotation scheme and dataset, we refer to Van Hee *et al.* (2015b).

### *3.2 Data annotation*

In the following paragraphs, we present our data annotation guidelines as described in Van Hee *et al.* (2015c) and focus on different types and roles related to the phenomenon.

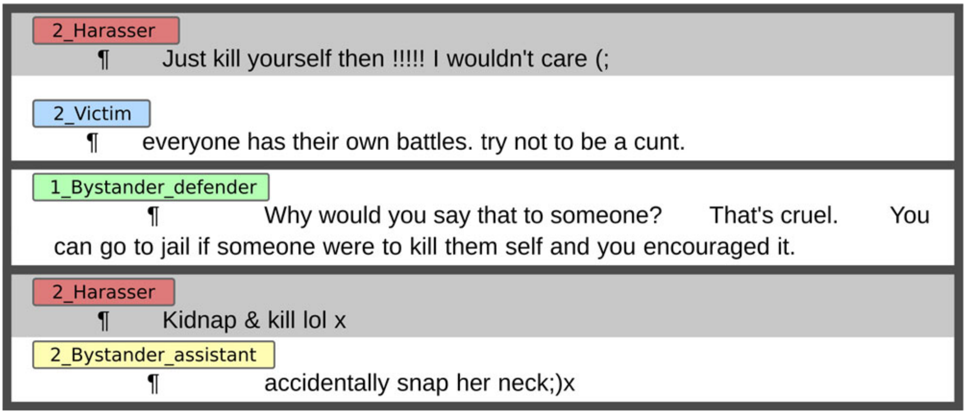
#### *3.2.1 Types of cyberbullying*

The guidelines used to annotate our corpora describe specific textual categories related to cyberbullying, including threats, insults, defensive statements from a victim, encouragements to the harasser, sexual talk, defamation, etc. (we refer to Van Hee *et al.* 2015c for a complete overview). All of these forms were inspired by social studies on cyberbullying (Vandebosch and Van Cleemput 2009; Van Cleemput *et al.* 2013) and manual inspection of cyberbullying examples.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

**Table 1.** Distribution of participant roles in the English and Dutch cyberbullying corpus

| | English | Dutch |
|---|---|---|
| Harasser | 3572 (66.46%) | 2890 (56.6%) |
| Victim | 1354 (25.19%) | 1603 (31.39%) |
| Bystander-defender | 425 (7.90%) | 575 (11.26%) |
| Bystander-assistant | 24 (0.45%) | 38 (0.74%) |



2_Harasser
¶ Just kill yourself then !!!!! I wouldn't care (;

2_Victim
¶ everyone has their own battles. try not to be a cunt.

1_Bystander_defender
¶ Why would you say that to someone? That's cruel. You can go to jail if someone were to kill them self and you encouraged it.

2_Harasser
¶ Kidnap & kill lol x

2_Bystander_assistant
¶ accidentally snap her neck;)x

**Figure 1.** Examples of ASKfm posts including "harasser," "victim," "bystander-defender," and "bystander-assistant" instances in the Brat annotation tool.

### *3.2.2 Participant roles in cyberbullying*

In formulating an annotation scheme for supervised classification of participant roles, annotators were asked to infer the bullying role as an illocutionary act by means of the information present in the ASKfm corpus. The "outsider" from Salmivalli's participant classification (Salmivalli, Voeten, and Poskiparta 2011b) has been left out, given that passive bystanders are impossible to recognize in online text because they do not leave traces. Concretely, four cyberbullying roles were annotated in both corpora:

**Harasser:** person who initiates the harassment, that is the bully.
**Victim:** person who is harassed.
**Bystander-assistant:** person who does not initiate but takes part in the actions of the harasser.
**Bystander-defender:** person who helps the victim and discourages the harasser from continuing.

The annotation scheme describes two levels of annotation. Firstly, the annotators were asked to indicate, at the post level, whether the post under investigation contained traces of cyberbullying. If so, the annotators identified the author's role as one out of the four mentioned above. Secondly, at the subsentence level, the annotators were tasked with the identification of a number of fine-grained categories related to cyberbullying. More concretely, they identified all text spans corresponding to one of the categories described in the annotation scheme. To provide the annotators with some context, all posts were presented within their original conversation when possible. All annotations were done using the Brat rapid annotation tool (Stenetorp *et al.* 2012). Figure 1 shows example annotations of participant roles.

Because of the rare occurrence of bystander-assistants in the dataset (English $n = 24$ (0.45%), Dutch $n = 38$ (0.74%) cf. Table 1), this class was merged with the harasser class for the machine

**Table 2.** Statistics of the English and Dutch cyberbullying corpus

| | Corpus size (posts) | Number (ratio) of bullying posts |
|---|---|---|
| English | 113,694 | 5375 (4.73%) |
| Dutch | 78,387 | 5106 (6.97%) |

learning experiments. First, classification difficulty increases significantly with data imbalance. Given the low occurrence of bystander-assistant instances, there was too little data to learn a discriminative model for these posts. Second, the textual content of the post assigned to both classes is similar; it is from the (incomplete) context that the initiation nature of harasser posts was determined by annotators. Note that the four labels will be maintained in the discussion of role distribution and dataset statistics presented in the next section.

### *3.3 Dataset and annotation statistics*

The English and Dutch corpora were independently annotated for cyberbullying by trained linguists after supervised instruction and practice with the guidelines. All were Dutch native speakers or English second-language speakers. To demonstrate the validity of our annotations, IAA scores were calculated using Kappa ($\kappa$) on a subset of the English and Dutch ASKfm corpus. For English, 3882 posts were annotated by four raters and for Dutch 6498 posts were annotated by two raters. For both IAA-corpora, all posts were annotated by each rater to have full overlap. Inter-rater agreement for Dutch is calculated using Cohen's Kappa (Cohen 1960), while Fleiss' Kappa (Fleiss 1971) is used for the English corpus due to there being more than two raters. Kappa scores for the identification of cyberbullying are $\kappa = 0.69$ (Dutch) and $\kappa = 0.59$ (English), which point to substantial and moderate agreement (Landis and Koch 1977). Kappa scores for participant roles are $\kappa = 0.65$ (Dutch) and $\kappa = 0.57$ (English), again pointing to respectively substantial and moderate agreement. We also computed cross-averaged $F_1$-score of the annotators: Dutch annotations have a 0.63 and English annotations a 0.59 macro-averaged $F_1$-score.

As the corpus consists of a random crawl of the ASKfm website, we have at our disposal a realistic dataset with regard to the occurrence of cyberbullying traces, which has been annotated at a fine-grained level. This stands in contrast with many previous studies that use keyword-based search to collect bully traces. The English and Dutch corpus contain 113,694 and 78,387 posts, respectively (cf. Table 2). A similar skewed distribution of bullying versus not bullying posts can be observed in both languages, as well as a comparable distribution of bullying roles.

## 4. Feature engineering and linear classification experiments

Given our annotated dataset, we approached the task of participant role detection as a multiclass classification task. As discussed in the previous section, the minority class "bystander-assistant" was merged with the "harasser" class for practical reasons. We set out to classify each textual post as either "not bullying" or as an instance where one out of the following bully participants is speaking: "harasser," "victim," or "bystander-defender." In this section, we describe the information sources we used as input to our classification algorithms (Section 4.1), we then give an overview of the different single classifiers and two ensemble learning techniques (Section 4.3) we investigated to obtain optimal performance (Section 4.2) and we outline how we used feature selection and data resampling to tackle the large imbalance in our dataset (Section 4.4).

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

### 4.1 Text preprocessing and feature engineering

As preprocessing, we applied tokenization, part-of-speech-tagging, and lemmatization to the data using the LeTs Preprocess Toolkit (van de Kauter *et al.* 2013). In supervised learning, a machine learning algorithm takes a set of training instances (the label of which is known) and seeks to build a model that generates a desired prediction for an unseen instance. To enable the model construction, all instances are represented as a vector of features (i.e., inherent characteristics of the data) that contain information that is potentially useful to distinguish bullying from not bullying content.

We experimentally tested whether cyberbullying events can be automatically recognized by lexical markers in a post. To this end, all posts were represented by a number of information sources (or *features*) including lexical features like bags of words, sentiment lexicon features, and topic model features, which are described in more detail below. Prior to feature extraction, some data cleaning steps were executed, such as the replacement of hyperlinks and @-replies, removal of superfluous white spaces, and the replacement of abbreviations by their full form (based on the chatslang.com lexicon). Additionally, tokenization was applied before *n*-gram extraction and lemmatization for sentiment lexicon matching, and stemming was applied prior to extracting topic model features.

After preprocessing the corpus, the following feature types were extracted:

**Word *n*-gram bags of words:** binary features indicating the presence of word unigrams, bigrams, and trigrams.

**Character *n*-gram bags of words:** binary features indicating the presence of character bigrams, trigrams, and fourgrams (without crossing word boundaries). Character *n*-grams provide some abstraction from the word level and provide robustness to the spelling variation that characterizes social media data.

**Term lists:** one binary feature derived for each one out of six lists, indicating the presence of a term from the list in a post: proper names, "allness" indicators (e.g., *always*, *everybody*), diminishers (e.g., *slightly*, *relatively*), intensifiers (e.g., *absolutely, amazingly*), negation words, and aggressive language and profanity words. We hypothesize that "allness" indicators, negation words, and intensifiers are informative of often hyperbolic bullying style. Diminishers, intensifiers, and negation words were all obtained from an English grammar describing these lexical classes or existing sentiment lexicons (see further). Person alternation is a binary feature indicating whether the combination of the first and second person pronoun occurs in order to capture interpersonal intent. We hypothesize posts that contain person alternation and proper names to be informative because cyberbullying is a directed and interpersonal communicative act.

**Subjectivity lexicon features:** Subjectivity lexicons provide categorical or continuous values for the affective connotation of a word. Sentiment can correspond to sentiment polarity, that is positive, negative, or neutral, or as more fine-grained categories of emotions or psychological processes (as is the case for Linguistic Inquiry and Word Count (LIWC) Pennebaker *et al.* 2001). From the sentiment lexicons, we derive positive and negative opinion word ratios, as well as the overall post polarity. For Dutch, we made use of the Duoman (Jijkoun and Hofmann 2009) and Pattern (De Smedt and Daelemans 2012) sentiment lexicons. For English, we included the Liu and Hu opinion lexicon (Hu and Liu 2004), the MPQA lexicon (Wilson, Wiebe, and Hoffmann 2005), the General Inquirer Sentiment Lexicon (Stone *et al.* 1966), AFINN (Nielsen 2011), and MSOL (Mohammad, Dunne, and Dorr 2009). For both languages, we included the relative frequency of all 68 psychometric categories in the LIWC dictionary for English (Pennebaker *et al.* 2001) and Dutch (Zijlstra *et al.* 2004). We hypothesize that certain psychometric and sentiment categories are informative of different roles. For instance, we expect more positive opinion to be expressed by a bystander-defender, whereas a bully is more likely to express negative sentiment and LIWC affective categories such as anxiety and anger.

**Topic model features:** Topic models provide hidden semantic structures in collections of text and are used to uncover thematic information. The similarity score of a post to one of these models

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

provides a score for how semantically similar the post is to the topics in that text collection. By making use of the Gensim topic modeling library (Rehurek and Sojka 2010), several Latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) and latent semantic analysis (Deerwester *et al.* 1990) topic models with varying granularity ($k = 20$, 50, 100, and 200) were trained on data corresponding to each fine-grained cyberbullying category (e.g., threats, defamations, insults, defenses). The topic models were trained on a background corpus (EN: 1,200,000 tokens, NL: 1,400,000 tokens) scraped with the BootCAT (Baroni and Bernardini 2004) web-corpus toolkit. Using BootCaT, we collected ASKfm posts from user profiles (different than the ones used for corpus compilation) using lists of manually determined seed words that are characteristic of the different cyberbullying categories based on the training data (e.g., "slut" for the insult category and "ignore" for the defense category). BootCat is crawler software that collects text from the web if it contains seed words. In this manner, we collect posts that are likely to be representative of the cyberbullying category corresponding to the characteristic seed words. The resulting background corpus was cleaned by removing non-English and non-Dutch text and stop words before creating topic models.

When applied to the training data, this resulted in 871,044 and 795,072 features for English and Dutch, respectively. To prevent features with greater numerical values to be attributed more weight in the model than features with smaller values, we used feature scaling, that is all feature values were scaled to the range [0, 1].

### 4.2 *Classification algorithms*

The tested multiclass classification algorithms include a linear-kernel SVM, a LR, a passive-aggressive (PA), and a stochastic gradient descent (SGD) classifier.[c] We chose these regularized linear classification models as this type of algorithms is typically fast enough for handling large datasets. The linear SVM makes use of the dual l2-loss implementation in LIBLINEAR (Fan *et al.* 2008), which uses a one-vs-rest strategy for multiclass classification. We fixed the loss function to squared hinge. The LR classifier also uses a one-vs-rest scheme for multiclass and a coordinate descent algorithm. The PA classifier (Crammer *et al.* 2006) is a type of online margin-based learning meant for large-scale data such as ours. We use the hinge loss as the loss function in PA, which is equivalent to the PA-I algorithm described in Crammer *et al.* (2006). Other regularized linear models using SGD learning were tested with a modified-huber loss function (Zhang 2004) which is the equivalent of a quadratically smoothed SVM with $\gamma = 2$ and the perceptron loss function. We used $10^6/n\ instances$ iterations for SGD, as well as PA. Testing the effectiveness of online and batch algorithms such as SGD and PA are useful for large datasets, for dealing with memory restrictions or for use with streaming data (such as social media content).

Feature selection settings and classifier hyper-parameters for each classifier type were tuned in grid search in fivefold cross-validation. Table 3 gives an overview of the tested hyper-parameters for the different algorithms. The winning parametrization was chosen by means of macro-averaged $F_1$-score averaged over the folds.

### 4.3 *Ensemble approaches*

Apart from evaluating the above-described single-algorithm classifiers, we also investigated two ensemble learning techniques: a Voting classifier and a Cascading classifier. Ensemble learning has the potential to improve classification performance by combining different individual classifiers and additionally mitigates data imbalance (Galar *et al.* 2012).

We tuned an ensemble Voting classifier out of the three best-scoring classifiers. Hence, the final classification was determined by majority vote on the predicted label by each classifier.

[c] For the experiments in this paper, we make use of Scikit-learn (Pedregosa *et al.* 2011), a machine learning library for the Python programming language.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

**Table 3.** Tested hyper-parameters for each step in our machine learning pipeline

| Algorithm | Hyper-parameters |
|---|---|
| Feature filtering | Feature scoring: “Anova F-value” or “Mutual information” |
| | Percentile of features retained: 67% or 33% |
| SVM (linear-kernel) | Cost C: [0.02, 0.2, 1, 2, 20, 200] |
| | Balanced class weighting: enabled or disabled |
| Logistic regression | idem |
| Passive aggressive | Loss function: hinged or squared hinged |
| | Cost C: [0.02, 0.2, 1, 2, 20, 200] |
| | Balanced class weighting: enabled or disabled |
| SGD | Loss function: “modified huber” or “perceptron” |
| | Balanced class weighting: enabled or disabled |

We also tested a Cascading (also known as “multistage”) classifier approach in which the output of the first classifier for binary detection (i.e., bullying vs. not bullying) of cyberbullying instances is followed by a multiclass role classifier for the positive predictions (i.e., bullying posts). The first-stage classifier detects the presence of cyberbullying (i.e., bullying or not). The predicted cyberbullying instances serve as input for the second-stage classifier, which then predicts to which one out of three participant roles (i.e., harasser, victim, or bystander-defender) the post appears. In this manner, we test if it is worthwhile to first detect the presence of cyberbullying in a post and to proceed with the second classifier that assigns a participant role to the post. For both the Voting and the Cascading ensemble, the selection of best models is based on the average cross-validation and holdout $F_1$-scores of the tuned feature selection and resampling pipeline.

Our previous work on cyberbullying detection in ASKfm posts (Van Hee *et al*. 2018) revealed that linear-kernel SVMs work well for binary classification of cyberbullying posts. The highest-scoring classifier setup and hyper-parameters (i.e., out of single-algorithm and ensemble Voting approaches) were selected for the second-stage multiclass role classification in the Cascading ensemble: for Dutch, this was the LR classifier, while the Voting classifier performed best for English. The detection classifier and the second-stage classifier were tuned for the SVM hyper-parameters in Table 3. Hyper-parameter selection was optimized jointly for both stages. In this manner, we tested whether it is plausible to first automatically detect instances of cyberbullying and subsequently classify the participant roles of positive bullying instances.

### *4.4 Feature selection and resampling for imbalance*

Due to the realistic nature of our dataset compared to previously used balanced studies, we encountered severe class imbalance or skew, which could negatively affect machine learning performance. In cyberbullying detection experiments on Twitter data, Al-Garadi *et al.* (2016) encounter the same data imbalance problem with a similar ratio of bullying to not bullying posts.

In line with Al-Garadi *et al.* (2016), we investigate feature selection and data resampling, two techniques that can be used to enhance classification performance in general, but in particular when faced with (severe) class imbalance (Japkowicz and Stephen 2002).

For the multiclass role of classification experiments, we include random undersampling, which randomly removes instances of the majority classes to obtain a desired ratio with one or more minority classes. This technique is used to bias the classifier toward the minority class (i.e.,

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

bullying posts). After initial experimentation, we set the ratio to 1/100, so that the majority classes were undersampled until the minority "bystander-defender" class became 1% of the training set. Initial runs determined that a balanced ratio and a ratio of 10% and 5% consistently produced worse results, so a ratio of 1% was chosen. In cyberbullying detection experiments, Al-Garadi *et al.* (2016) have shown that synthetic super-sampling such as Synthetic Minority Oversampling Technique (SMOTE) effectively improve classifier performance. These algorithms create new observations of minority classes usually by means of some type of nearest neighbors or bootstrapping method. We tested SMOTE+Tomek and SMOTE+Edited Nearest Neighbours (Batista, Prati, and Monard 2004), as well as ADASYN (He *et al.* 2008) for synthetic super-sampling but found these algorithms to be too computationally expensive to work with our high-dimensional data.

For feature selection, we chose filter metrics that are used to characterize both the relevance and redundancy of variables (Guyon and Elisseeff 2003). For feature filtering, we relied on ANOVA F-value and mutual information (MI) score as metrics. The MI implementation in Scikit-learn relies on nonparametric methods based on entropy estimation from $k$-nearest neighbors distances as described in Ross (2014) ($k = 3$). Using grid searches over the different pipeline setups, we determined the feature filter scoring function and the percentile of retained features to construct the model.

Feature selection and undersampling were tuned and parameterized for each classification algorithm, except for the Voting and Cascading ensemble approaches where undersampling was not applied. We also investigated whether the sequential order of feature selection and data resampling had any effect on the classification performance. Each possible pipeline setup and parametrization were evaluated by means of fivefold cross-validation on our hold-in development set.

## 5. PLM experiments

The field of Natural Language Processing (NLP) has transitioned from developing task-specific models to fine-tuning approaches based on large general-purpose language models (Howard and Ruder 2018; Peters *et al.* 2018). PLMs of this type are based on the transformer architecture (Vaswani *et al.* 2017). Currently, the most widely used model of this type is BERT (Devlin *et al.* 2019) and its many variants. Fine-tuning BERT-like models obtain state-of-the-art performance in many NLP tasks including text classification. In the OffensEval shared task, Zampieri *et al.* (2019b) BERT-based approaches outperformed all other classification approaches.

We experimented with fine-tuning pretrained transformer models for both English and Dutch using the "huggingface/transformers" PyTorch library which provides models in their repository (Wolf *et al.* 2019). We used the provided default hyper-parameters, tokenizers, and configuration for all models. To add sentence classification capability to the language models, a sequence classification head was added to the original transformer architecture. The batch size was set to 32 instances and sequence length to 256 for all models. The only hyper-parameter set in cross-validation grid search is the number of training epochs ($e = \{4, 8, 16\}$). The best hyper-parametrization was chosen for each architecture by macro-averaged $F_1$-score over fivefold. Due to the more expensive compute of these Graphics Processing Unit accelerated models compared to linear classification algorithms, we did not experiment with ensembles, feature selection, and data resampling.

### *5.1 Pretrained models*

For English, we ran text classification experiments using pretrained BERT (Devlin *et al.* 2019), RoBERTa (Liu *et al.* 2019), and XLNet (Yang *et al.* 2019) models.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

BERT is an attention-based auto-encoding sequence-to-sequence model using two unsupervised task objectives. The first task is word masking, where the masked language model (MLM) has to guess which word is masked in its position in the text. The second task is next sentence prediction (NSP) performed by predicting whether two sentences are subsequent in the corpus or randomly sampled from the corpus. The specific English BERT model used is the "bert-base-uncased" as provided with the original paper[d] pretrained on the 3.3B word Wikipedia + BookCorpus corpus. We prefer using a lowercase ("uncased") model over a case-preserving ("cased") model as our cyberbullying dataset contains social media text with many mistakes against capitalization convention. However, a lowercased pretrained model was only available for English BERT.

The RoBERTa model (Liu *et al.* 2019) improved over BERT by dropping NSP and using only the MLM task on multiple sentences instead of single sentences. The authors argue that while NSP was intended to learn inter-sentence coherence, it actually learned topic similarity because of the random sampling of sentences in negative instances. The specific RoBERTa model used is the "roberta-base" cased model provided alongside the original paper[e] pretrained on the 160-GB Wikipedia + BooksCorpus + CommonCrawl-News + CommonCrawl-Stories + OpenTextWeb corpus.

XLNet is a permutation language model that combines strengths of auto-regressive and auto-encoding modeling approaches: permutation language models are trained to predict tokens given preceding context like a traditional unidirectional language model, but instead of predicting the tokens in sequential order, it predicts tokens in a random order sampling from both the left and right context. XLNet incorporates two key ideas from the TransformerXL architecture: relative positional embeddings and the recurrence mechanism. In combination with the permutation objective, these techniques effectively capture bidirectional context while avoiding the independence assumption and the discrepancy between pretraining and fine-tuning caused by the use of masked tokens in BERT. The specific XLNet model used is the "xlnet-base" cased model released alongside the original work[f] pretrained on a Wikipedia + BooksCorpus + Giga5 + ClueWeb 2012-B + Common Crawl totalling 339B SentencePiece tokens.

For Dutch, we tested BERTje (de Vries *et al.* 2019) and RobBERT (Delobelle, Winters, and Berendt 2020) which are the Dutch architectural equivalents of BERT and RobBERTa. The specific BERTje model is "bert-base-dutch-cased"[g] trained on a 2.4-B token Wikipedia + Books + SoNaR-500 + Web news + Wikipedia corpus. The specific RobBERT model is "robbert-base"[h] pretrained on the Dutch part of the OSCAR corpus (39 GB). The choice of these models was informed by their improved performance over another available monolingual Dutch Bert-NL model[i] or the multilingual mBERT (Devlin *et al.* 2019) model. As subword token input BERT and BERTje uses WordPiece, RoBERTa, and RobBERT use byte-level Byte Pair Encoding (BPE), and XLNet uses SentencePiece (Kudo and Richardson 2018).

## 6. Experimental results

In this section, we discuss the results of the parametrized classifier approaches. We chose macro-averaged $F_1$-score to evaluate the models so as to attribute equal weight to each class in the evaluation. We compared the scores of the winning parametrized model by classifier type on a

[d] https://github.com/google-research/bert.
[e] https://github.com/pytorch/fairseq/tree/master/examples/roberta.
[f] https://github.com/zihangdai/xlnet/.
[g] https://github.com/wietsedv/bertje.
[h] https://github.com/iPieter/RobBERT.
[i] http://textdata.nl/bert-nl.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

**Table 4.** Results (%) for role classification using a single linear classification algorithm approach. **Boldface** indicates highest score for the relevant metric

| English | | | | | | |
|---|---|---|---|---|---|---|
| | Cross-validation | | | Holdout test | | |
| Classifier | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| SVM | 59.47 | 51.54 | 54.29 | 58.95 | 51.95 | 54.29 |
| Logistic regression | 58.60 | **53.48** | **55.47** | 57.13 | **53.26** | **54.54** |
| Passive aggressive | **60.94** | 46.81 | 51.03 | **67.97** | 47.50 | 53.31 |
| SGD | 59.78 | 46.39 | 50.38 | 60.83 | 47.55 | 51.97 |
| Random BL | 24.98 | 25.84 | 12.07 | 24.96 | 21.89 | 12.14 |
| Majority BL | 23.81 | 25.00 | 24.39 | 23.88 | 25.00 | 24.43 |

| Dutch | | | | | | |
|---|---|---|---|---|---|---|
| | Cross-validation | | | Holdout test | | |
| Classifier | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| SVM | 59.51 | 49.00 | 52.78 | 59.41 | 49.65 | 53.21 |
| Logistic regression | 58.90 | **50.78** | **54.03** | 58.53 | **50.95** | **53.92** |
| Passive aggressive | **59.64** | 46.80 | 51.34 | **61.52** | 45.18 | 50.31 |
| SGD | 57.56 | 47.36 | 51.15 | 55.98 | 45.50 | 49.40 |
| Random BL | 25.01 | 24.90 | 12.79 | 24.80 | 26.31 | 12.55 |
| Majority BL | 23.37 | 25.00 | 24.16 | 23.37 | 25.00 | 24.16 |

holdout test set comprising 10% of the entire corpus. Hence, the holdout scores are $F_1$-scores for 10,000 random corpus samples.

For baseline comparison, we chose a majority baseline (a.k.a. the zero-rule baseline) in which the negative majority class "not bullying" is always predicted, as well as a random baseline in which random predictions are made.

We first compare the result of hyper-parameter-optimized single-algorithm classifier approaches without feature filtering or resampling. Then, we discuss the best results obtained by the models where feature selection and resampling were included. Finally, we discuss the ensemble approaches and recapitulate our main findings in the conclusion. We consider $F_1$-score on the randomly split 10% holdout test set as the comparative measure in the discussion of the best-performing system. As is common practice, cross-validation scores served for establishing the best hyper-parameters and as an indicator of under- or over-fitting of the model when compared to the holdout test scores.

### *6.1 Single-algorithm classifier results*

As can be deduced from Table 4, for the English dataset, the LR classifier obtained the best results with $F_1$= 55.47% in cross-validation and $F_1$= 54.54% on the holdout test set. The runner-up is the SVM classifier which performs comparably but trades in recall. The PA algorithm obtains a remarkably high precision. All classifiers show a marked improvement over the majority and random baselines with $F_1$-scores of, respectively, 24.43% and 12.14%.

**Table 5.** Results (%) for ensemble role classification, combining logistic regression, SVM, and SGD for English and logistic regression, SVM, and the passive aggressive classifier for Dutch. **Boldface** indicates the highest score for the relevant metric. All scores are macro-averaged

| English | | | | | | |
|---|---|---|---|---|---|---|
| | Cross-validation | | | Holdout test | | |
| Classifier | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| Voting | **59.90** | 51.85 | 54.71 | **60.06** | 51.95 | 54.70 |
| Cascading | 54.55 | **56.80** | **55.19** | 54.68 | **58.15** | **55.67** |
| Dutch | | | | | | |
| | Cross-validation | | | Holdout test | | |
| Classifier | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| Voting | **59.18** | 49.70 | **53.31** | **59.64** | 50.72 | 54.12 |
| Cascading | 54.12 | **50.48** | 51.79 | 53.77 | **55.26** | **54.30** |

Similar results are observed for cyberbullying role classification on the Dutch dataset: the LR classifier obtained the best results in both cross-validation ($F_1$= 54.03%) and on the holdout set ($F_1$= 53.92%). The SVM classifier is again a close second and the other algorithms performed comparably. Like we observed for the English corpus, PA obtained the highest precision for a trade-off in recall. Here as well, the classifiers substantially outperformed the majority ($F_1$= 24.16%) and random baseline ($F_1$= 12.55%).

The algorithms do not seem to under- or over-fit as evidenced by the similar scores obtained through cross-validation and on the holdout test set. When recall is important, for instance in a semi-automated moderation context where the system assists human moderators by flagging posts for manual review, LR classification seems to be the desired approach. However, when precision is important, for instance in fully automated moderation applications, PA appears to be a good choice. SGD obtains the worst performance on both datasets.

### *6.2 Ensemble classifier results*

In this section, we investigate the benefits of ensemble learning in two different types of ensemble classifiers: a Voting ensemble and a Cascading ensemble. The Voting ensemble classifier combines the three best-scoring classifiers by majority vote. Concretely, each individual algorithm provides a prediction and the majority vote is considered for the final prediction. For English, we combined LR, SVM, and SGD as the individual models. For Dutch, LR, SVM, and PA classifiers are considered. These were tuned by cross-validation with the same experimental settings and hyper-parameters as described earlier.

The Cascading ensemble consists of two stages where an SVM first predicts positive instances of cyberbullying. Second, we choose the best multiclass classifier to perform fine-grained role classification on these instances. The second-stage role classifier for English is the Voting classifier, whereas for Dutch it is the SVM. For both models, feature selection was applied. In this manner, we can examine whether it is worthwhile to detect if cyberbullying is present in the first step before classifying the participant role in the second step.

As shown in Table 5, for both languages, the Cascading approach outperforms the Voting classifier in terms of recall while the latter obtains better precision. The second-stage results shown in

**Table 6.** Results (%) for the initial cyberbullying detection stage and subsequent role classification stage on the holdout test set in the best-performing Cascading ensemble for the Dutch and English corpus. Macro-averaged scores are given for both stages; binary-averaged scores on the positive class are shown for the detection stage

| English: Cascading ensemble | | | | | | |
|---|---|---|---|---|---|---|
| | Holdout macro-averaged | | | Holdout binary-averaged | | |
| Stage in cascade | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| 1*st* stage: cyberbullying detection | 75.53 | 83.54 | 78.9 | 52.49 | 70.06 | 60.02 |
| 2*nd* stage: role classification | 54.43 | 33.88 | 41.28 | | | |

| Dutch: Cascading ensemble + feature selection | | | | | | |
|---|---|---|---|---|---|---|
| | Holdout macro-averaged | | | Holdout binary-averaged | | |
| Stage in cascade | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| 1*st* stage: cyberbullying detection | 78.52 | 80.33 | 79.39 | 59.6 | 63.67 | 61.57 |
| 2*nd* stage: role classification | 48.79 | 30.81 | 37.52 | | | |

Table 6 differ from the results in Table 5 because the former shows the macro-averaged scores on the three bullying roles (i.e., “harasser,” “victim,” and “bystander-defender”) without the negative “not bullying” class, whereas the latter shows the macro-averaged scores on all four classes.

Both ensemble classifiers outperform the best single classifier on the holdout test set (cf. Table 4). To conclude, ensemble methods obtain a higher recall than single-classifier approaches, but they provide marginal practical value taking into account their computational cost.

To examine the performance of the Cascade ensemble in closer detail, we present the scores obtained in the two stages of the classification in Table 6. The first detection stage obtains a macro-averaged $F_1$-score of 78.9% on the holdout test set for English and 79.39% for Dutch. Since this is a binary cyberbullying detection task, binary-averaged precision, recall, and $F_1$-score are also presented. These are the results for the positive bullying instances only. The detection scores (stage 1) slightly underperform scores obtained in previous research on cyberbullying detection in this dataset (Van Hee *et al.* 2018) using comparable linear classification algorithms (their binary-averaged $F_1$-score is 64% for English and 61% for Dutch). The second-stage role classifier obtains an $F_1$-score of 41.28% for English and 37.52% for Dutch. The results show the negative effect of both error percolation in the first stage and the difficulty of cyberbullying role classification.

However, in spite of the relatively poor performance of the second-stage classifier, final recall scores for role classification (in cross-validation) remain higher in Cascading ensembles when compared to the single-algorithm and Voting approaches.

### *6.3 Feature selection and resampling*

To mitigate the effects of data imbalance, we experimented with feature selection and random undersampling techniques. We created pipelines with each enabled and in sequence with both possible orders, that is resampling followed by feature selection and inversely, feature selection followed by resampling. When looking at the individual classifiers, we found that random undersampling rarely improved results and due to the added computational complexity this step was left out when testing the ensemble methods. Table 7 presents the results obtained by the best-performing classification pipelines of the above experiments. The scores were more often improved by feature

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

**Table 7.** Results (%) of the best configurations for both languages. **Boldface** indicates the highest score for the relevant metric. All scores are macro-averaged

English

| Classifier | Cross-validation | | | Holdout test | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| SVM + FS | 59.35 | 51.04 | 53.89 | 59.75 | 51.89 | 54.51 |
| LR + FS | 58.74 | 53.33 | 55.42 | 57.86 | 53.12 | 54.77 |
| PA | 61.54 | 46.13 | 50.68 | **73.02** | 46.15 | 52.66 |
| SGD + FS + RES | 49.49 | 53.20 | 50.01 | 57.33 | 56.45 | 56.42 |
| Voting + FS | 59.76 | 51.45 | 54.39 | 60.92 | 51.77 | 54.87 |
| Cascading | 54.55 | 56.80 | 55.19 | 54.68 | 58.15 | 55.67 |
| BERT | 61.95 | 57.61 | 59.48 | 60.62 | **59.89** | **60.04** |
| RoBERTa | **62.63** | **58.88** | **60.25** | 58.06 | 54.66 | 55.84 |
| XLNet | 59.66 | 54.64 | 56.85 | 63.73 | 56.92 | 59.88 |

Dutch

| Classifier | Cross-validation | | | Holdout test | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| SVM | 59.51 | 49.00 | 52.78 | 59.41 | 49.65 | 53.21 |
| LR | 58.90 | 50.78 | 54.03 | 58.53 | 50.95 | 53.92 |
| PA + RES | 54.11 | 49.94 | 51.22 | 51.46 | 51.00 | 50.87 |
| SGD + RES | 55.08 | 50.10 | 52.2 | 54.15 | 52.44 | 53.18 |
| Voting | 59.18 | 49.70 | 53.31 | 59.64 | 50.72 | 54.12 |
| Cascading + FS | 55.07 | 49.44 | 51.53 | 54.08 | **55.07** | 54.37 |
| BERTje | 57.52 | 48.65 | 52.06 | 56.54 | 50.62 | 52.96 |
| RobBERT | **60.17** | **51.60** | **54.92** | **61.93** | 53.58 | **56.73** |

selection than by random undersampling. The $F_1$-score improvements are minimal and are in line with the findings of Al-Garadi *et al.* (2016) who report similar increases.

The best score for English is obtained when feature selection is included as preprocessing, followed by random undersampling and SGD (modified huber loss) with $F_1 = 56.42\%$. For Dutch, the Cascading ensemble with ANOVA F-score feature selection obtains the best score with $F_1 = 54.37\%$ on the holdout set. It should be noted however, that SGD obtains the worst performance in cross-validation. More robust results are obtained by the Cascading classifier, where the cross-validation and holdout scores are consistent and obtain the highest recall.

### *6.4 PLM results*

Finally, we discuss the results of the PLM fine-tuning experiments. Table 7 presents all results obtained by the best-performing classification pipelines of the above experiments including transformer-based classifiers.

The default hyper-parameters and configurations were applied for all transformer models. The only hyper-parameter that was tuned in cross-validation is the number of training epochs. All Dutch and English models performed best with four epochs, except for the English XLNet model, which obtained the best results with eight epochs. The results of the cross-validation experiments in Table 7 show that, respectively, RoBERTa and RobBERT outperform the optimized single classifiers and the ensemble methods for English and Dutch, with $F_1$-scores of 60.25% and 54.92%, respectively.

We also observe a precision improvement over the Voting ensembles for both languages. After applying the best models to the holdout test set, however, we observe a remarkable decrease in the English RoBERTa model compared with the cross-validated scores and we see the model outperformed by BERT and XLNet.

Overall, the table shows a performance increase of about 5 points with RoBERTa compared to the Cascading model (i.e., $F_1$= 60.25% versus $\mathrm{F}_1$= 55.19%). For Dutch, the improvement is less outspoken, with RobBERT scoring about 1.6 points better than the best ensemble method (i.e., $F_1$= 54.92% versus $F_1$= 53.31%). The transformer-based models further show more balance between precision and recall than the single and ensemble classifiers for English, but they show larger differences for Dutch. This observation applies to the cross-validated results and the results on the holdout test set. In Section 7, we examine the results of the best models in closer detail and we discuss their performance on the different class labels.

## 7. Discussion

This section contains a discussion of the resulting scores and a qualitative error analysis providing insight into the types of misclassifications and areas of potential improvement. Figures 2 and 3 show the results per class on the holdout test set obtained by the best models as determined by cross-validation experiments, that is RoBERTa and RobBERT for English and Dutch, respectively.

The confusion matrices visualize the amount of mislabeling by class. We see that both models scored significantly better on the majority "not bullying" class compared to the three cyberbullying classes (i.e., "harasser," "victim," and "bystander-defender"). The best linear classifier pipeline (not pictured) still shows bias toward the majority class despite the use of imbalance mitigation techniques. However, the imbalance mitigated linear models perform better on the "harasser" and "victim" classes than the PLM models ("harasser" 62.9% correct for English and 56.8% for Dutch, "victim" 21.3% and 30.5% resp.). The PLM classifiers perform much better on the "bystander-defender" minority class than the linear classifier pipelines (English: 29% correct, Dutch: 35.8% correct) even without data imbalance mitigation.

While "harasser" and "bystander-defender" posts are mostly predicted correctly, the figures show much more confusion for the "victim" class. In the English corpus, such instances are most often confused with the "not bullying" class, followed by the "harasser" class. In the Dutch corpus, there is much confusion with the "not bullying" class, but considerably less with the "harasser" class. One explanation for these results could be that victim posts show important linguistic variation as they can range from assertive and self-defending utterances (which may be hard to distinguish from bullying in some contexts) to indifference and to reactions of desperation and distress.

It is also noteworthy that, in both languages, on average 4 in 10 harasser posts are predicted as "not bulling," which indicates that the bullying is rather lexically implicit or goes unnoticed given the limited context that is available.

Overall, the confusion matrices show that, on the one hand, all three cyberbullying classes are often confused with "not bullying," which is most likely due to (i) classifier bias toward the majority class, (ii) "masked" cyberbullying through implicit language, and (iii) lack of conversational context. On the other hand, not bullying posts as well as "victim" and "bystander-defender"

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

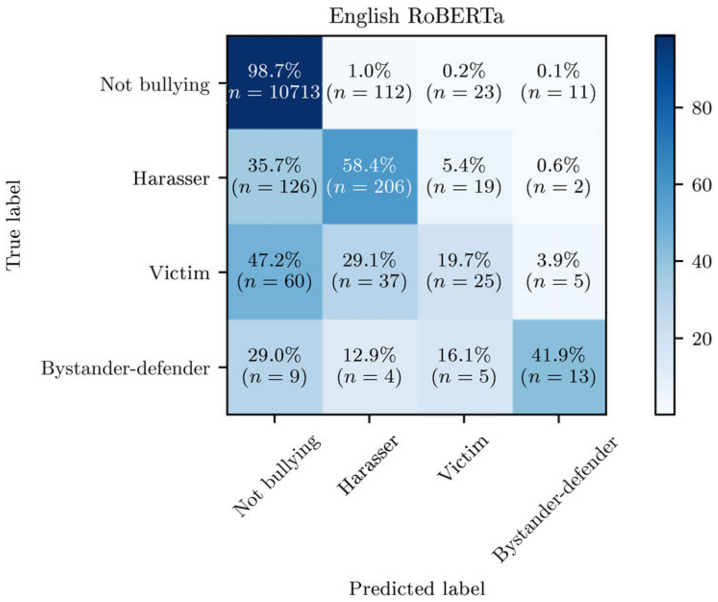


**Figure 2.** Confusion matrix for best English system.

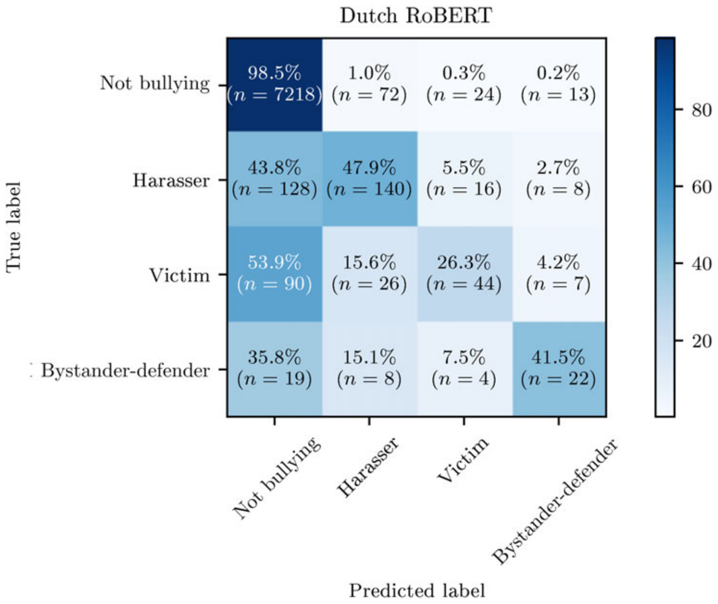


**Figure 3.** Confusion matrix for best Dutch system.

utterances that are classified as harassing may be due to users' socially accepted slang in non-harmful contexts.

To get a better insight into classification errors, we performed a qualitative analysis of the predictions by the English RoBERTa and the Dutch RobBERT model and included some corpus examples. As for harasser posts that were classified as "not bullying," we observed that such

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

posts often require more context to understand that bullying is taking place. Such context was available to the annotators because all posts were presented in their original thread when available. Recognizing cyberbullying at the post level is, however, much more challenging when the cyberbullying is implicit (example 1). Other examples of false negatives include utterances with sexting or sexual requests, the language of which is often suggestive and ambiguous (example 2).

Victim posts that went unnoticed are sensibly shorter (i.e., containing half the number of characters) than correctly classified victim posts and often contain irony (examples 3 and 4). "Bystander-defender" posts that were predicted as "not bullying" are considerably shorter than correct predictions as well (on average 153 versus 225 characters per post in the English corpus and 103 vs. 212 characters in the Dutch corpus). In addition, defender posts often contain different polarities and targets, which may complicate their classification. Sometimes both the bully and the victim are addressed in the same utterance. For instance, when directed to the bully, the tone of voice is negative, whereas positive words are for the victim (example 5).

(1) *Ik kus nog liever een everzwijn.* (EN: *I'd rather kiss a boar.*) [prediction: "not bullying"]
(2) *Pic of u in bikini?* [prediction: "not bullying"]
(3) *Kijk naar mijn gezicht, kan het mij iets schelen? NEEN.* (EN: *Look at my face, do you think I care? NO.*) [prediction: "not bullying"]
(4) *Criticising me? Thanks.* [prediction: "not bullying"])
(5) *You're pretty just ignore the hate.* [prediction: "not bullying"]

The confusion matrices also show a certain degree of confusion between the "victim" and "bystander-defender" classes and the "harasser" class. This will be discussed in the next section, where we focus on the differences between these misclassifications for English and Dutch.

In sum, our qualitative analysis revealed that false negatives are often examples of cyberbullying that lack explicit profane words (e.g., defamations, sexually inappropriate comments), but also posts that lack (historical) context to understand that bullying is going on, and posts that contain noisy language. In fact, the typical language and (rather unintended) spelling mistakes we observed in both corpora augments the sparsity of lexical features and may therefore affect classification performance. As mentioned earlier, we observed that false negatives for the "victim" and "bystander-defender" class contain on average half the number of characters compared to correctly identified posts, for Dutch as well as for English.

In a semi-automatic (i.e., machine plus human) moderation setup, false negatives are considered more problematic than false positives, and the matrices show they are in fact a bigger problem. However, false positives are undesirable as well, especially in fully automated monitoring. In this regard, the qualitative analysis showed that both the English and Dutch models are mislead by, among other things, socially accepted slang or profanity (examples 6 and 7), genuine kind statements as if they were meant to counter a negative one (example 8) and rude statements that address a large group of people rather than an individual victim (example 9). Here as well, more context would help to judge whether cyberbullying is actually going on. Another type of information that could be useful here could be world knowledge, for instance to differentiate between individual targets (which may be considered more urgent cyberbullying cases) and group or "mass" targets.

(6) *Job interview. Wish me luck you cunts!* [prediction: "harasser"])
(7) *Hahaha, ik bedoel dak u begrijp slet ;)* (EN: *Hahaha, I mean that I totally get you, slut ;)* [prediction: "harasser"]
(8) *Ok, Je bent niet lelijk :O !!!* (EN: *Ok, you are not ugly :O !!!*) [prediction: "victim"]
(9) *Ge hebt echt zo rotte, schijnheilige, achterlijke mensen op de wereld eh, ugh.* (EN: *There really are rotten, hypocritical and retarded people in this world, ugh.*) [prediction: "harasser"]

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

**Table 8.** Casing, flooding, and profanity statistics drawn from the victim posts ($n = 127$ and $n = 167$) in the English and Dutch cyberbullying corpus, respectively

| | English | Dutch |
|---|---|---|
| Uppercase words | 7.41% | 1.90% |
| Flooded punctuation tokens | 0.58% | 0.24% |
| Profane words toward bully | 2.14% | 0.87% |
| All profane words | 2.34% | 1.50% |

### 7.1 *English versus Dutch cyberbullying role detection*

Having at our disposal a similar cyberbullying corpus for English and Dutch allows us to compare cyberbullying role detection performance in the two languages. When looking at Table 7, we observe that the English RoBERTa outperforms the Dutch RobBERT in the cross-validation setup. Our analysis revealed that there is more lexical variety in the Dutch corpus compared to the English, which increases the model complexity of the former. However, the holdout test scores show a slightly better performance of the Dutch ($F_1 = 56.73\%$) compared to the English best model ($F_1 = 55.84\%$).

In the following paragraphs, we examine the most outspoken (i.e., > 5%) differences in misclassifications between the English and Dutch best system as visualized in Figures 2 and 3.

The most outstanding difference concerns the "victim" class. Nearly, one out of three is predicted as a harassing post in the English corpus, whereas for Dutch confusions with the "harasser" class count for only 15.6%. This suggests that victim posts in the English corpus contain more aggressive language that resembles harassing compared to Dutch.

In fact, Table 8 shows lexical differences between the two corpora that seems to support this finding. To account for varying post length, the statistics in Table 8 were calculated with the total number of corpus tokens as the denominator. "Profane words towards bully" differs from "All profane words" in that the former only includes profane words directed at the bully (e.g., *"Stfu all of you bitches, leave her alone (. . .)"*), whereas the latter also includes profanity in indirect speech (e.g., *"Get off her back, alright? You're calling her a bitch (. . .)"*). While examples like the latter are not necessarily violent comments, the presence of abusive words may have affected their classification.

Our qualitative analysis further revealed that 2% of all tokens in the English corpus are swear words (e.g., *"fuck"*, *"WTF"*), whereas this is only 0.6% in the Dutch corpus. At the post level, this means that more than 1 in 4 (26%) English victim posts contain swearing, as opposed to 1 in 10 (10%) Dutch posts.

Another noticeable difference between the two languages concerns the false negatives for the three bullying classes. Figure 3 shows that, for Dutch, the "harasser," "victim," and "bystander-defender" posts are more often confused with the "not bullying" class compared to English. This suggests that (i) the bullying in our Dutch corpus is more implicit or requires more context and (ii) the bullying classes in the English corpus contain more aggressive language compared to non-bullying posts.

Lastly, we observe a difference of about 8.5% for the "bystander-defender" posts predicted as "victim," with English defender posts being more often confused with victim posts compared to Dutch.

The above analysis leads us to the tentative conclusion that "victim" posts in the English corpus are more assertive or even "aggressive" compared to the Dutch corpus and are therefore harder to distinguish from "harasser" posts. For both languages, false negatives for all cyberbullying categories (i.e., "harasser," "victim," and "bystander-defender") are most likely due to the absence of lexical cues, and the little context we dispose of given the data genre. This lack of context also makes it hard to differentiate between "harassers" and assertive "bystander-defenders" or self-defending "victims." When considering a conversation of just two utterances (i.e., one question–answer pair), both participants are likely to use aggressive language. As social media are and will be a place of informal communication including slang and non-harmful swearing, being able to model conversational threads or keeping track of interaction history should allow to better estimate whether cyberbullying is going on. Another way classification performance could be improved is by capturing offenses that are implicit (e.g., including irony) or that require world knowledge to understand (i) their hurtful intent and (ii) who is targeted.

## 8. Conclusion

In this work, we investigated fine-grained cyberbullying role detection in a real-world social media corpus for two languages (viz. English and Dutch). Apart from differentiating between "bullying" and "not bullying" messages, we also aimed to detect three participant roles in the bullying messages (i.e., "harasser," "victim," and "bystander-defender"). To our knowledge, no previous study has been done to automatically predict the participant role in cyberbullying messages on a representative corpus (i.e., containing real-world data and not biased by keyword search collection). We presented two different experimental setups, one where we optimized and compared linear task-specific classification algorithms, and another one where we explored the performance of fine-tuning pretrained transformer models for this task. In the first setup, different classification algorithms were compared and evaluated as part of a combined ensemble architecture (i.e., Cascading and Voting). Given the highly imbalanced nature of the dataset, we also experimented with filter-based feature selection and random undersampling to facilitate the classifier to learn from the positive class. Feature selection seemed to generally have a minimal positive impact, whereas random undersampling more often had a negative effect on the system performance for both languages. In the second setup, we experimented with pretrained BERT, RoBERTa, and XLNet models for English and BERTje and RobBERT for Dutch.

With both experimental setups, we have shown that participant roles can be classified with satisfactory results. The transformer-based models RoBERTa and RobBERT achieved the highest score for English and Dutch. While RoBERTa outperformed the best Cascading classifier by 5%, the difference between RobBERT and the Voting classifier is less outspoken (1.6%). There is a practical trade-off to consider as the linear models are much less computationally expensive to train than fine-tuning large-scale language models. For English, the best model obtained an macro-averaged $F_1$-score of 60.25% in cross-validation and 55.84% on the holdout test set. For Dutch, the best $F_1$-score reached 54.92% in cross-validation and 56.73% on the holdout test set.

Both are a marked improvement over the majority and random baselines. This demonstrates that cyberbullying role detection is feasible, even with little to no conversational and author-based historic context. All bullying content is solely identified by textual content at the post level. We have shown that automated role classification methods can be assistive tools for moderation purposes. In addition, they can help in collecting data for bullying research, in a more efficient and effective way than manual or commonly used keyword-based search.

While our results show that cyberbullying role classification in English and Dutch text are feasible tasks, there is still room for improvement. As the data used for this research has limited context, it is often hard to distinguish between different roles, for example, discriminating between "harassers" that initiate the bullying and assertive "victims" or aggressive "bystander-defenders" often requires more context than what we had available (i.e., one question–answer pair per instance). An interesting future research direction will be to investigate other conversational data genres that allow to take user and context information into account.

Another challenge inherent to this type of task is data imbalance. In future research, we will therefore investigate different methodologies to use DNNs for problems containing a high level of class imbalance (see Johnson and Khoshgoftaar 2019 for a recent survey on the topic).

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

Lastly, we will investigate whether our approach for role detection could be complemented with a script-based approach (Schank 1975) and (Tomkins 1978), in which stereotypical sequences of (cyberbullying) events in a specific context are modeled.

**Data availability and reproducibility.** For reproducibility, we provide experiment replication data and source code for download at https://osf.io/nb2r3/. The full annotated dataset is available upon request.

## References

**Agrawal S. and Awekar A.** (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *European Conference on Information Retrieval*, Grenoble, France. Cham, Switzerland: Springer, pp. 141–153.

**Al-Garadi M.A., Varathan K.D. and Ravana S.D.** (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior* **63**, 433–443.

**Baroni M. and Bernardini S.** (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, LREC'04. Lisbon, Portugal: European Language Resources Association (ELRA), pp. 1313–1316.

**Basile V., Bosco C., Fersini E., Nozza D., Patti V., Pardo F.M.R., Rosso P. and Sanguinetti M.** (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 54–63.

**Bastiaensens S., Vandebosch H., Poels K., Van Cleemput K., DeSmet A. and De Bourdeaudhuij I.** (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior* **31**, 259–271.

**Bastiaensens S., Vandebosch H., Poels K., Van Cleemput K., DeSmet A. and De Bourdeaudhuij I.** (2015). 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology* **34**(4), 425–435.

**Batista G.E.A.P.A., Prati R.C. and Monard M.C.** (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations Newsletter* **6**(1), 20–29.

**Blei D.M., Ng A.Y. and Jordan M.I.** (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

**Chatzakou D., Kourtellis N., Blackburn J., De Cristofaro E., Stringhini G. and Vakali A.** (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci'17. Troy, New York, USA: ACM, pp. 13–22.

**Chen H., Mckeever S. and Delany S.J.** (2017). Harnessing the power of text mining for the detection of abusive content in social media. In *Advances in Computational Intelligence Systems*. Cham, Switzerland: Springer, pp. 187–205.

**Cheng T. and Wicks T.** (2014). Event detection using twitter: A spatio-temporal approach. *PloS ONE* **9**(6), e97807.

Childfocus (2018). Clicksafe: Veilig internetten. www.childfocus.be/nl/preventie/clicksafe-veilig-internetten (accessed 2018-05-14).

**Cohen J.** (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1), 37–46.

**Cowie H.** (2013). Cyberbullying and its impact on young people's emotional health and well-being. *The Psychiatrist* **37**(5), 167–170.

**Crammer K., Dekel O., Keshet J., Shalev-Shwartz S. and Singer Y.** (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research* **7**, 551–585.

**Dadvar M.** (2014). *Experts and Machines United Against Cyberbullying*. *Phd thesis*, University of Twente.

**Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. and Harshman R.** (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41**, 391–407.

**Delobelle P., Winters T. and Berendt B.** (2020). Robbert: A dutch roberta-based language model. ArXiv pre-print https://arxiv.org/abs/2001.06286.

**De Smedt T. and Daelemans W.** (2012). "Vreselijk mooi!" ("Terribly Beautiful!"): A subjectivity lexicon for dutch adjectives. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC'12. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 3568–3572.

**DeSmet A., Bastiaensens S., Van Cleemput K., Poels K., Vandebosch H. and De Bourdeaudhuij I.** (2012). Mobilizing bystanders of cyberbullying: An exploratory study into behavioural determinants of defending the victim. *Annual Review of Cybertherapy and Telemedicine* **10**, 58–63.

**de Vries W., van Cranenburgh A., Bisazza A., Caselli T., van Noord G. and Nissim M.** (2019). Bertje: A dutch bert model. ArXiv pre-print https://arxiv.org/abs/1912.09582.

**Devlin J., Chang M.-W., Lee K. and Toutanova K.** (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*, vol. 1. Minneapolis, Minnesota, US: Association for Computational Linguistics, pp. 4171–4186.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

**Dinakar K., Jones B., Havasi C., Lieberman H. and Picard R.** (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems* **2**(3), 18:1–18:30.

**Dinakar K., Reichart R. and Lieberman H.** (2011). Modeling the detection of textual cyberbullying. In *Proceedings of The Social Mobile Web*, volume WS-11-02 of *AAAI Workshops*. Barcelona, Catalonia, Spain: AAAI, pp. 11–17.

**Dooley J.J. and Cross D.** (2009). Cyberbullying versus face-to-face bullying: A review of the similarities and differences. *Journal of Psychology* **217**, 182–188.

**Emmery C., Verhoeven B., De Pauw G., Jacobs G., Van Hee C., Lefever E., Desmet B., Hoste V. and Daelemans W.** (2019). Current Limitations in Cyberbullying Detection: On Evaluation Criteria, Reproducibility, and Data Scarcity. ArXiv pre-print https://arxiv.org/abs/1801.05617.

**EU Kids Online**. (2014). EU Kids Online: Findings, methods, recommendations.

**Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R. and Lin C.-J.** (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* **9**, 1871–1874.

**Fleiss J.L.** (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**(5), 378–382.

**Galar M., Fernandez A., Barrenechea E., Bustince H. and Herrera F.** (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **42**(4), 463–484.

**González-Bailón S., Wang N., Rivero A., Borge-Holthoefer J. and Moreno Y.** (2014). Assessing the bias in samples of large online networks. *Social Networks* **38**, 16–27.

**Guyon I. and Elisseeff A.** (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1157–1182.

**He H., Bai Y., Garcia E.A. and Li S.** (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence)*. Hong Kong, China: IEEE, pp. 1322–1328.

**Hosseinmardi H., Rafiq R.I., Han R., Lv Q. and Mishra S.** (2016). Prediction of cyberbullying incidents in a media-based social network. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM'16. Piscataway, NJ, USA: IEEE Press, pp. 186–192.

**Howard J. and Ruder S.** (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339.

**Hu M. and Liu B.** (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD04. Seattle, WA, USA: ACM, pp. 168–177.

**Japkowicz N. and Stephen S.** (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis* **6**(5), 429–449.

**Jijkoun V. and Hofmann K.** (2009). Generating a non-English subjectivity lexicon: Relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 398–405.

**Johnson J.M. and Khoshgoftaar T.M.** (2019). Survey on deep learning with class imbalance. *Journal of Big Data* **27**(6), 1–54.

**Kaltiala-Heino R., Rimpelä M., Marttunen M., Rimpelä A. and Rantanen P.** (1999). Bullying, depression, and suicidal ideation in finnish adolescents: School survey. *BMJ* **319**(7206), 348–351.

**Klein J., Cornell D. and Konold T.** (2012). Relationships between bullying, school climate, and student risk behaviors. *School Psychology Quarterly* **27**(3), 154.

**Kudo T. and Richardson J.** (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71.

**Kumar R., Ojha A.K., Malmasi S. and Zampieri M.** (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1–11.

**Landis J.R. and Koch G.G.** (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174.

**Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L. and Stoyanov V.** (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

**Livingstone S., Kirwil L., Ponte C. and Staksrud E.** (2014). In their own words: What bothers children online? *European Journal of Communication* **29**(3), 271–288.

Ministre de l'ducation nationale (2018). Non au harcelment. www.nonauharcelement.education.gouv.fr (accessed 14 May 2018).

**Mohammad S., Dunne C. and Dorr B.** (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, EMNLP'09. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 599–608.

**Nahar V., Al-Maskari S., Li X. and Pang C.** (2014). Semi-supervised learning for cyberbullying detection in social networks. In **Wang H. and Sharaf M.A.** (eds), *Databases Theory and Applications*. Cham, Switzerland: Springer, pp. 160–171.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

**Nansel T.R., Overpeck M., Pilla R.S., Ruan W.J., Simons-Morton B. and Scheidt P.** (2001). Bullying behaviors among us youth: Prevalence and association with psychosocial adjustment. *Journal of the American Medical Association* **285**(16), 2094–2100.

**Nielsen F.å.** (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In Rowe M., Stankovic M., Dadzie A.-S. and Hardey M. (eds), *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, CEUR Workshop Proceedings, vol. 718. Heraklion, Crete: CEUR-WS.org, pp. 93–98.

**Olweus D.** (1993). *Bullying at School: What We Know and What We Can Do*, 2nd Edn. Hoboken, New Jersey, USA: Wiley.

**O'Neill B. and Dinh T.** (2018). The Better Internet for Kids Policy Map: Implementing the European Strategy for a Better Internet for Children in European Member States. Technical report, European Commission, European Schoolnet, EU Kids Online.

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

**Pennebaker J.W., Francis M.E. and Booth R.J.** (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates.

**Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L.** (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237.

**Price M. and Dalgleish J.** (2010). Cyberbullying: Experiences, impacts and coping strategies as described by Australian young people. *Youth Studies Australia* **29**(2), 51–59.

**Raisi E. and Huang B.** (2018). Weakly supervised cyberbullying detection with participant-vocabulary consistency. *Social Network Analysis and Mining* **8**(1), 1–17.

**Rehurek R. and Sojka P.** (2010). Software framework for topic modelling with large corpora. In *The LREC 2010 Workshop on new Challenges for NLP Frameworks*. Valletta, Malta: European Language Resources Association (ELRA), pp. 45–50.

**Reynolds K., Kontostathis A. and Edwards L.** (2011). Using machine learning to detect cyberbullying. In *Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops*, ICMLA'11. Washington, DC, USA: IEEE Computer Society, pp. 241–244.

**Ross B.C.** (2014). Mutual information between discrete and continuous data sets. *PloS One* **9**(2), e87357.

**Salmivalli C.** (1999). Participant role approach to school bullying: Implications for interventions. *Journal of Adolescence* **22**(4), 453–459.

**Salmivalli C.** (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior* **15**(2), 112–120.

**Salmivalli C., Kärnä A. and Poskiparta E.** (2011a). Counteracting bullying in finland: The kiva program and its effects on different forms of being bullied. *International Journal of Behavioral Development* **35**(5), 405–411.

**Salmivalli C., Lagerspetz K., Björkqvist K., Österman K. and Kaukiainen A.** (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior* **22**(1), 1–15.

**Salmivalli C. and Pöyhönen V.** (2012). Cyberbullying in finland. In *Cyberbullying in the Global Playground: Research from International Perspectives*, pp. 57–72.

**Salmivalli C., Voeten M. and Poskiparta E.** (2011b). Bystanders matter: Associations between reinforcing, defending, and the frequency of bullying behavior in classrooms. *Journal of Clinical Child & Adolescent Psychology* **40**(5), 668–676.

**Schank R.C.** (1975). *Conceptual Information Processing*. New York, NY, USA: Elsevier Science Inc.

**Slonje R. and Smith P.K.** (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology* **49**(2), 147–154.

**Smith P.K., Mahdavi J., Carvalho M., Fisher S., Russell S. and Tippett N.** (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry* **49**(4), 376–385.

**Sprugnoli R., Menini S., Tonelli S., Oncini F. and Piras E.** (2018). Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, pp. 51–59.

**Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S. and Tsujii J.** (2012). brat: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. Avignon, France: Association of Computational Linguistics, pp. 102–107.

**Stone P.J., Dunphy D.C.D., Smith M.S. and Ogilvie D.M.** (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, Massachusetts, USA: MIT Press.

**Sui J.** (2015). *Understanding and Fighting Bullying with Machine Learning*. PhD thesis, Department of Computer Sciences, University of Wisconsin-Madison.

**Tokunaga R.S.** (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior* **26**(3), 277–287.

**Tomkins S.S.** (1978). Script theory: Differential magnification of affects. *Nebraska Symposium on Motivation* **26**, 201–236.

**Tucker C.E.** (2010). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research* **51**(5), 546–562.

https://doi.org/10.1017/S135132492000056X Published online by Cambridge University Press

**Van Cleemput K., Bastiaensens S., Vandebosch H., Poels K., Deboutte G., DeSmet A. and De Bourdeaudhuij I.** (2013). Zes jaar onderzoek naar cyberpesten in Vlaanderen, België en daarbuiten: een overzicht van de bevindingen. (Six years of cyberbullying research in Flanders, Belgium and beyond: an overview of the findings.). Technical report, Brussels, Belgium: Friendly Attac, IWT-SBO.

**van de Kauter M., Coorman G., Lefever E., Desmet B., Macken L. and Hoste V.** (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal* **3**, 103–120.

**Van Hee C., Jacobs G., Emmery C., Desmet B., Lefever E., Verhoeven B., De Pauw G., Daelemans W. and Hoste V.** (2018). Automatic detection of cyberbullying in social media text. *PLOS ONE* **13**(10), 1–22.

**Van Hee C., Lefever E., Verhoeven B., Mennes J., Desmet B., De Pauw G., Daelemans W. and Hoste V.** (2015a). Automatic detection and prevention of cyberbullying. In Lorenz P. and Bourret C. (eds), *Proceedings of the International Conference on Human and Social Analytics*. St. Julians, Malta: IARIA, pp. 13–18.

**Van Hee C., Lefever E., Verhoeven B., Mennes J., Desmet B., De Pauw G., Daelemans W. and Hoste V.** (2015b). Detection and fine-grained classification of cyberbullying events. In Angelova G., Bontcheva K. and Mitkov R. (eds), *Proceedings of Recent Advances in Natural Language Processing*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, pp. 672–680.

**Van Hee C., Verhoeven B., Lefever E., De Pauw G., Daelemans W. and Hoste V.** (2015c). Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0. Technical Report LT3 15-01, LT3, Language and Translation Technology Team–Ghent University.

**Van Royen K., Poels K. and Vandebosch H.** (2016). Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites. *Children and Youth Services Review* **64**, 35–41.

**Vandebosch H. and Van Cleemput K.** (2009). Cyberbullying among youngsters: Profiles of bullies and victims. *New Media & Society* **11**(8), 1349–1371.

**Vandebosch H., Van Cleemput K., Mortelmans D. and Walrave M.** (2006). Cyberpesten bij jongeren in Vlaanderen: Een studie in opdracht van het viWTA (Cyberbullying among youngsters in Flanders: a study commissoned by the viWTA). Brussels: viWTA. Technical report, Universities of Antwerp & viWTA.

**Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł. and Polosukhin I.** (2017). Attention is all you need. *In Advances in Neural Information Processing Systems*, pp. 5998–6008.

**Wilson T., Wiebe J. and Hoffmann P.** (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT'05. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 347–354.

**Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M. and Brew J.** (2019). Huggingface's transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.

**Xu J.-M., Jun K.-S., Zhu X. and Bellmore A.** (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT'12. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 656–666.

**Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R. and Le Q.V.** (2019). XLNet: Generalized autoregressive pretraining for language understanding. ArXiv pre-print https://arxiv.org/abs/1906.08237.

**Yin D., Xue Z., Hong L., Davison B.D., Kontostathis A. and Edwards L.** (2009). Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the Web 2.0*, vol. 2. Madrid, Spain: CAW, pp. 1–7.

**Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N. and Kumar R.** (2019a). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1415–1420.

**Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N. and Kumar R.** (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 75–86.

**Zhang T.** (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 116. New York, NY, USA: ACM.

**Zhang X., Tong J., Vishwamitra N., Whittaker E., Mazer J.P., Kowalski R., Hu H., Luo F., Macbeth J. and Dillon E.** (2016). Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Anaheim, CA: IEEE, pp. 740–745.

**Zhao R. and Mao K.** (2017). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing* **8**(3), 328–339.

**Zijlstra H., Van Meerveld T., Van Middendorp H., Pennebaker J.W. and Geenen R.** (2004). De Nederlandse versie van de 'linguistic inquiry and word count' (LIWC). *Gedrag Gezond* **32**, 271–281.

**Cite this article:** Jacobs G, Van Hee C and Hoste V (2022). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? *Natural Language Engineering* **28**, 141–166. https://doi.org/10.1017/S135132492000056X