

RESEARCH ARTICLE

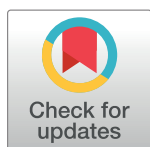
Automatic detection of cyberbullying in social media text

Cynthia Van Hee¹*, Gilles Jacobs¹, Chris Emmery², Bart Desmet¹, Els Lefever¹, Ben Verhoeven², Guy De Pauw², Walter Daelemans², Véronique Hoste¹

1 Department of Translation, Interpreting and Communication - Faculty of Arts and Philosophy, Ghent University, Ghent, Belgium, **2** Department of Linguistics - Faculty of Arts, University of Antwerp, Antwerp, Belgium

* These authors contributed equally to this work.

* cynthia.vanhee@ugent.be



OPEN ACCESS

Citation: Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, et al. (2018) Automatic detection of cyberbullying in social media text. PLoS ONE 13(10): e0203794. <https://doi.org/10.1371/journal.pone.0203794>

Editor: Hussein Suleman, University of Cape Town, SOUTH AFRICA

Received: February 6, 2017

Accepted: August 28, 2018

Published: October 8, 2018

Copyright: © 2018 Van Hee et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Because the actual posts in our corpus could contain names or other identifying information, we cannot share them publicly in a repository. They can, however be obtained upon request, for academic purposes solely and via gillesm.jacobs@ugent.be or cynthia.vanhee@ugent.be. The replication data are available through the Open Science Framework repository <https://osf.io/rgqw8/> with DOI [10.17605/OSF.IO/RGQW8](https://doi.org/10.17605/OSF.IO/RGQW8). This replication dataset allows interested researchers to download 1) the feature vectors of the corpus underlying the experiments described in this paper, 2) the indices

Abstract

While social media offer great communication opportunities, they also increase the vulnerability of young people to threatening situations online. Recent studies report that cyberbullying constitutes a growing problem among youngsters. Successful prevention depends on the adequate detection of potentially harmful messages and the information overload on the Web requires intelligent systems to identify potential risks automatically. The focus of this paper is on automatic cyberbullying detection in social media text by modelling posts written by bullies, victims, and bystanders of online bullying. We describe the collection and fine-grained annotation of a cyberbullying corpus for English and Dutch and perform a series of binary classification experiments to determine the feasibility of automatic cyberbullying detection. We make use of linear support vector machines exploiting a rich feature set and investigate which information sources contribute the most for the task. Experiments on a hold-out test set reveal promising results for the detection of cyberbullying-related posts. After optimisation of the hyperparameters, the classifier yields an F_1 score of 64% and 61% for English and Dutch respectively, and considerably outperforms baseline systems.

Introduction

Web 2.0 has had a substantial impact on communication and relationships in today's society. Children and teenagers go online more frequently, at younger ages, and in more diverse ways (e.g. smartphones, laptops and tablets). Although most of teenagers' Internet use is harmless and the benefits of digital communication are evident, the freedom and anonymity experienced online makes young people vulnerable with cyberbullying being one of the major threats [1, 2].

Bullying is not a new phenomenon and cyberbullying has manifested itself as soon as digital technologies have become primary communication tools. On the positive side, social media like blogs, social networking sites (e.g. Facebook), and instant messaging platforms (e.g. WhatsApp) make it possible to communicate with anyone and at any time. Moreover, they are a place where people engage in social interaction, offering the possibility to establish new

corresponding to instances that were kept separately to test the experimental design (referred to as the "hold-out test set" in the paper), 3) a feature mapping dictionary that allows to trace all indices in the feature vector files back to the corresponding feature types (e.g. the feature indices 0 to 14,230 represent word 3-gram features). We also share the seed terms that were used to construct the corpora for our topic model features. Lastly, we provide an Excel spreadsheet presenting a results overview of all the tested systems. All of this information is made available for both the Dutch and English experiments.

Funding: The work presented in this paper was carried out in the framework of the AMiCA IWT SBO-project 120007 project to WD and VH, funded by the government Flanders Innovation & Entrepreneurship (VLAIO) agency; <http://www.vlaio.be>. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

relationships and maintain existing friendships [3, 4]. On the negative side however, social media increase the risk of children being confronted with threatening situations including grooming or sexually transgressive behaviour, signals of depression and suicidal thoughts, and cyberbullying. Users are reachable 24/7 and are often able to remain anonymous if desired: this makes social media a convenient way for bullies to target their victims outside the school yard.

With regard to cyberbullying, a number of national and international initiatives have been launched over the past few years to increase children's online safety. Examples include *KiVa* (<http://www.kivaprogram.net/>), a Finnish cyberbullying prevention programme, the '*Non au harcèlement*' campaign in France, Belgian governmental initiatives and helplines (e.g. *clicksafe.be*, *veiligonline.be*, *mediawijs.be*) that provide information about online safety, and so on.

In spite of these efforts, a lot of undesirable and hurtful content remains online. [2] analysed a body of quantitative research on cyberbullying and observed cybervictimisation rates among teenagers between 20% and 40%. [5] focused on 12 to 17 year olds living in the United States and found that no less than 72% of them had encountered cyberbullying at least once within the year preceding the questionnaire. [6] surveyed 9 to 26 year olds in the United States, Canada, the United Kingdom and Australia, and found that 29% of the respondents had ever been victimised online. A study among 2,000 Flemish secondary school students (age 12 to 18) revealed that 11% of them had been bullied online at least once in the six months preceding the survey [7]. Finally, the 2014 large-scale EU Kids Online Report [8] published that 20% of 11 to 16 year olds had been exposed to hate messages online. In addition, youngsters were 12% more likely to be exposed to cyberbullying as compared to 2010, which clearly demonstrates that cyberbullying is a growing problem.

The prevalence of cybervictimisation depends on the conceptualisation used in describing cyberbullying, but also on research variables such as location and the number and age span of the participants. Nevertheless, the above studies demonstrate that online platforms are increasingly used for bullying, which is a cause for concern given its impact. As shown by [9–11], cyberbullying may negatively impact the victim's self-esteem, academic achievement and emotional well-being. [12] found that self-reported effects of cyberbullying include negative effects on school grades and feelings of sadness, anger, fear, and depression. In extreme cases, cyberbullying could even lead to self-harm and suicidal thoughts.

These findings demonstrate that cyberbullying is a serious problem the consequences of which can be dramatic. Early detection of cyberbullying attempts is therefore of key importance to youngsters' mental well-being. Successful detection depends on effective monitoring of online content, but the amount of information on the Web makes it practically unfeasible for moderators to monitor all user-generated content manually. To tackle this problem, intelligent systems are required that process this information in a fast way and automatically signal potential threats. This way, moderators can respond quickly and prevent threatening situations from escalating. According to recent research, teenagers are generally in favour of such automatic monitoring, provided that effective follow-up strategies are formulated, and that privacy and autonomy are guaranteed [13].

Parental control tools (e.g. *NetNanny*, <https://www.netnanny.com/>) already block unsuited or undesirable content and some social networks make use of keyword-based moderation tools (i.e. using lists of profane and insulting words to flag harmful content). However, such approaches typically fail to detect implicit and subtle forms of cyberbullying in which no explicit vocabulary is used. This creates the need for intelligent and self-learning systems that go beyond keyword spotting and hence improve the recall of cyberbullying detection.

The ultimate goal of this type of research is to develop models that could improve manual monitoring for cyberbullying on social networks. We explore the automatic detection of

textual signals of cyberbullying, in which cyberbullying is approached as a complex phenomenon that can be realised in various ways (see the Annotation guidelines section for a detailed overview). While the vast majority of the related research focuses on detecting cyberbullying ‘attacks’ (i.e. verbal aggression), the present study takes different types of cyberbullying into account, including more implicit posts from the bully, but also posts written by victims and bystanders. This is a more inclusive conceptualisation for the task of cyberbullying detection and should aid in moderation and prevention efforts by capturing different and more implicit signals of bullying.

To tackle this problem, we propose a machine learning method based on a linear SVM classifier [14, 15] exploiting a rich feature set. The contribution we make is twofold: first, we develop a complex classifier to detect *signals* of cyberbullying, which allows us to detect different types of cyberbullying that are related to different social roles involved in a cyberbullying event. Second, we demonstrate that the methodology is easily portable to other languages, provided there is annotated data available, by performing experiments on an English and Dutch dataset.

The remainder of this paper is structured as follows: the next section presents a definition of cyberbullying and its participant roles and provides an overview of the state of the art in cyberbullying detection. The *Data collection and annotation* section describes the corpus construction and annotation. Next, we present the experimental setup and discuss our experimental results for English and Dutch. Finally, the *Conclusion and future research* section concludes this paper and provides some perspectives for further research.

Related research

Both offline and online bullying are widely covered in the realm of social sciences and psychology, and the increasing number of cyberbullying cases in recent years [16] has stimulated research efforts to detect cyberbullying automatically. In the following section, we present a definition of cyberbullying and identify its participant roles and we provide a brief overview of automatic approaches to cyberbullying detection.

Cyberbullying definition and participant roles

A common starting point for conceptualising cyberbullying are definitions of traditional (i.e. *offline*) bullying, one of the most influential ones being formulated by [17]. The researcher described bullying based on three main criteria, including i) **intention** (i.e. a bully intends to inflict harm on the victim), ii) **repetition** (i.e. bullying acts take place repeatedly over time) and iii) a **power imbalance** between the bully and the victim (i.e. a more powerful bully attacks a less powerful victim). With respect to cyberbullying, a number of definitions are based on the above criteria. A popular definition is that of [18, p. 376], which describes cyberbullying 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”. However, opinion on the applicability of the above characteristics to cyberbullying is very much divided [19], and besides theoretical objections, a number of practical limitations have been observed. Firstly, while [17] claims intention to be inherent to traditional bullying, this is much harder to ascertain in an online environment. Online conversations lack the signals of a face-to-face interaction like intonation, facial expressions and gestures, which makes them more ambiguous than real-life conversations. The receiver may therefore get the wrong impression that they are being offended or ridiculed [20]. Another criterion for bullying that might not hold in online situations is the power imbalance between the bully and the victim. This can be evident in real life (e.g. the bully is taller, stronger or older than the victim), but it

is hard to conceptualise or measure online, where power may be related to technological skills, anonymity or the inability of the victim to escape from the bullying [19, 21]. Also empowering for the bully are inherent characteristics of the Web: once defamatory or confidential information is made public through the Internet, it is hard to remove.

Finally, while arguing that repetition distinguishes bullying from single acts of aggression, [17] himself states that such a single aggressive action can be considered bullying under certain circumstances. Accordingly, [21] claim that repetition in cyberbullying is problematic to operationalise, as it is unclear what the consequences are of a single derogatory message on a public page. A single act of aggression or humiliation may cause continued distress and humiliation for the victim if it is shared or liked by a large audience [21]. [22, p. 26] compare this with the “snowball effect”: one post may be repeated or distributed by other people so that it becomes out of the control of the initial bully and has larger effects than was originally intended.

Given these arguments, a number of less ‘strict’ definitions of cyberbullying were proposed by among others [2, 5, 6], where a power imbalance and repetition are not deemed necessary conditions for cyberbullying.

The above paragraphs demonstrate that defining cyberbullying is far from trivial, and varying prevalence rates (see the [Introduction](#) section) confirm that a univocal definition of the phenomenon is still lacking in the literature [2]. Based on existing conceptualisations, we define cyberbullying as *content that is published online by an individual and that is aggressive or hurtful against a victim*. Based on this definition, an annotation scheme was developed [23] to signal textual characteristics of cyberbullying, including posts from bullies, as well as reactions from victims and bystanders.

Cyberbullying research also involves the identification of its participant roles. [24] were among the first to define the roles in a bullying situation. **Based on surveys among teenagers involved in real-life bullying situations, they defined six participant roles: victims (i.e. who are the target of repeated harassment), bullies (i.e. who are the initiative-taking perpetrators), assistants of the bully (i.e. who encourage the bullying), reinforcers of the bully (i.e. who reinforce the bullying), defenders (i.e. who comfort the victim, take their side or try to stop the bullying) and outsiders (i.e. who ignore or distance themselves from the situation).** In sum, in addition to the bully and victim, the researchers distinguish four bystanders (i.e. assistants, reinforcers, defenders and outsiders). [25], however, do not distinguish between reinforcers and assistants of the bullying. Their typology includes victims, bullies and three types of bystanders: i) bystanders who participate in the bullying, ii) bystanders who help or support the victim and iii) bystanders who ignore the bullying. The cyberbullying roles that are identified in our annotation scheme are based on existing bullying role typologies, given that traditional bullying roles are applicable to cyberbullying as well [26, 27]. More details about the different roles that we take into account are provided in the Data collection and annotation section.

Bystanders and -to a lesser extent- victims are often overlooked in the related research. As a result, these studies can be better characterised as verbal aggression detection concerned with retrieving bully attacks. By taking bystanders into account, we capture different and more subtle signals of a bullying episode. Note that while in this work we did not include classification of the participant roles as such, they are essential to the conceptualisation of the current detection task.

Detecting and preventing cyberbullying

As mentioned earlier, although research on cyberbullying detection is more limited than social studies on the phenomenon, some important advances have been made in recent years. In

what follows, we present a brief overview of the most important natural language processing approaches to cyberbullying detection, but we refer to the survey paper by [28] for a more detailed overview.

Although some studies have investigated the effectiveness of rule-based modelling [29], the dominant approach to cyberbullying detection involves machine learning. Most machine learning approaches are based on supervised [30, 30–32] or semi-supervised learning [33]. The former involves the construction of a classifier based on labelled training data, whereas semi-supervised approaches rely on classifiers that are built from a training corpus containing a small set of labelled and a large set of unlabelled instances. Semi-supervised methods are often used to handle data sparsity, a typical issue in cyberbullying research. As cyberbullying detection essentially involves the distinction between bullying and non-bullying posts, the problem is generally approached as a binary classification task where the positive class is represented by instances containing (textual) cyberbullying, while the negative class is devoid of bullying signals.

A key challenge in cyberbullying research is the availability of suitable data, which is necessary to develop models that characterise cyberbullying. In recent years, only a few datasets have become publicly available for this particular task, such as the training sets provided in the context of the CAW 2.0 workshop (<http://caw2.barcelonamedia.org>), a MySpace (<https://myspace.com>) [34] and Formspring (<http://www.formspring.me>) cyberbullying corpus annotated with the help of Mechanical Turk [29], and more recently, the Twitter Bullying Traces dataset [35]. Many studies have therefore constructed their own corpus from social media websites that are prone to bullying content, such as YouTube [30, 32], Twitter [36, 37], Instagram [38], MySpace [31, 34], FormSpring [29, 39], Kaggle [40] and ASKfm [41]. Despite the bottleneck of data availability, cyberbullying detection approaches have been successfully implemented over the past years and the relevance of automatic text analysis techniques to ensure child safety online has been recognised [42].

Among the first studies on cyberbullying detection are [29–31], who explored the predictive power of n -grams (with and without tf-idf weighting), part-of-speech information (e.g. first and second pronouns), and sentiment information based on (polarity and profanity) lexicons for this task. Similar features were not only exploited for coarse-grained cyberbullying detection, but also for the detection of more fine-grained cyberbullying categories [41]. Despite their apparent simplicity, content-based features (i.e. lexical, syntactic and sentiment information) are very often exploited in recent approaches to cyberbullying detection [33, 43]. In fact, as observed by [28], more than 41 papers have approached cyberbullying detection using content-based features, which confirms that this type of information is crucial for the task.

More and more, however, content-based features are combined with semantic features derived from topic model information [44], word embeddings and representation learning [43, 45]. More recent studies have also demonstrated the added value of user-based information for the task, more specifically by including users' activities (i.e. the number of posts) on a social network, their age, gender, location, number of friends and followers, and so on [32, 33, 46, 47]. A final feature type that gains increasing popularity in cyberbullying detection are network-based features, whose application is motivated by the frequent use of social media data for the task. By using network information, researchers aim to capture social relations between participants in a conversation (e.g. bully versus victim), and other relevant information such as the popularity of a person (i.e. which can indicate the power of a potential bully) on a social network, the number of (historical) interactions between two people, and so on. [48] for instance used network-based features to take the behavioural history of a potential bully into account. [49] detected cyberbullying in tweets and included network features inspired by

Olweus' [17] bullying conditions (see *supra*). More specifically, they measured the power imbalance between a bully and victim, as well as the bully's popularity based on interaction graphs and the bully's position in the network.

As mentioned earlier, social media are a commonly used genre for this type of tasks. More recently, researchers have investigated cyberbullying detection in multi-modal data offered by specific platforms. For instance [38] explored cyberbullying detection using multi-modal data extracted from the social network Instagram. More precisely, they combined textual features derived from the posts themselves with user metadata and image features and showed that integrating the latter enhanced the classification performance. [37] also detected cyberbullying in different data genres, including ASKfm, Twitter, and Instagram. They took role information into account by integrating bully and victim scores as features, based on the occurrence of bully-related keywords in their sent or received posts.

With respect to the datasets used in cyberbullying research, it can be observed that corpora are often composed by keyword search (e.g. [43, 44]), which produces a biased dataset of positive (i.e. bullying) instances. To balance these corpora, negative data are often added from a background corpus or data resampling [50] techniques are adopted [33, 47]. For this research, data were randomly crawled across ASKfm and no keyword search was used to collect bullying data. Instead, all instances were manually annotated for the presence of bullying. As a result, our corpus contains a realistic distribution of bullying instances.

When looking at the performance of automatic cyberbullying, we see that scores vary greatly and do not only depend on the implemented algorithm and parameter settings, but also on a number of other variables. These include the metrics that are used to evaluate the system (i.e. micro- or macro-averaged F_1 , precision, recall, AUC, etc.), the corpus genre (i.e. Facebook, Twitter, ASKfm, Instagram) and class distribution (i.e. balanced or unbalanced), the annotation method (i.e. automatic annotations or manual annotations using crowdsourcing or by experts) and, perhaps the most important distinguishing factor, the conceptualisation of cyberbullying that is used. More concretely, while some approaches identify sensitive topics [30] or insulting language [29], others propose a more comprehensive approach by capturing different types of cyberbullying [41] or by modelling the bully-victim communications involved in a cyberbullying incident [37].

The studies discussed in this section demonstrated the variety of approaches that have been used to tackle cyberbullying detection. However, most of them focused on cyberbullying 'attacks', or posts written by a bully. Moreover, it is not entirely clear if different forms of cyberbullying were taken into account (e.g. sexual intimidation or harassment, or psychological threats), in addition to derogatory language or insults. In the present study, cyberbullying is considered a complex phenomenon comprising different forms of harmful online behaviour, which are described in more detail in our annotation scheme [23]. Purposing to facilitate manual monitoring efforts on social networks, we developed a system that automatically detects signals of cyberbullying, including attacks from bullies, as well as victim and bystander reactions, the latter of which are generally overlooked in related research.

Most similar to this research is the work by [44], [43, 45], who investigated bullying traces posted by different author roles (e.g. bully, victim, bystander, assistant, defender, reporter, accuser, reinforcer). However, they collected tweets using the keywords *bully*, *bullied* and *bullying*. As a result, their corpus contained many reports or testimonials of cyberbullying (example 1), instead of actual cyberbullying. Moreover, their method implies that cyberbullying signals that are devoid of such keywords are not included in the training corpus.

1. "Some tweens got violent on the n train, the one boy got off after blows 2 the chest. . . Saw him cryin as he walkd away: (bullying not cool)" [44, p. 658]

What clearly distinguishes these works from the present is that their conceptualisation of cyberbullying is not explained. It is, in other words, not clear which type of posts are considered bullying and which are not. In the present research, we identify different types of bullying and all are included in the positive class of our experimental corpus.

For this research, English and Dutch social media data were annotated for fine-grained forms of cyberbullying, based on the actors involved in a cyberbullying incident. After preliminary experiments for Dutch [41, 51], we currently present an optimised cyberbullying detection method for English and Dutch and hereby show that the proposed methodology can easily be applied to different languages, provided that annotated data are available.

Data collection and annotation

To be able to build representative models for cyberbullying, a suitable dataset is required. This section describes the construction of two corpora, English and Dutch, containing social media posts that are manually annotated for cyberbullying according to our fine-grained annotation scheme. This allows us to cover different forms and participants (or *roles*) involved in a cyberbullying event.

Data collection

Two corpora were constructed by collecting data from the social networking site ASKfm, where users can create profiles and ask or answer questions, with the option of doing so anonymously. ASKfm data typically consists 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 (<http://www.gnu.org/software/wget/>) in April and October, 2013. After language filtering (i.e. non-English or non-Dutch content was removed), the experimental corpora comprised 113,698 and 78,387 posts for English and Dutch, respectively.

Data annotation

Cyberbullying has been a widely covered research topic recently and studies have shed light on direct and indirect types of cyberbullying, implicit and explicit forms, verbal and non-verbal cyberbullying, and so on. This is important from a sociolinguistic point of view, but knowing what cyberbullying involves is also crucial to build models for automatic cyberbullying detection. In the following paragraphs, we present our data annotation guidelines [23] and focus on different types and roles related to the phenomenon.

Types of cyberbullying

Cyberbullying research is mainly centered around the conceptualisation, occurrence and prevention of the phenomenon [1, 52, 53]. Sociolinguistic studies have identified different types of cyberbullying [12, 54, 55] and compared these types with forms of traditional or offline bullying [20]. Like traditional bullying, direct and indirect forms of cyberbullying have been identified. Direct cyberbullying refers to actions in which the victim is directly involved (e.g. sending a virus-infected file, excluding someone from an online group, insulting and threatening), whereas indirect cyberbullying can take place without awareness of the victim (e.g. *outing* or publishing confidential information, spreading gossip, creating a hate page on social networking sites) [20].

The present annotation scheme describes some specific textual categories related to cyberbullying, including threats, insults, defensive statements from a victim, encouragements to the harasser, etc. (see the [Data collection and annotation](#) section for a complete overview). All of

these forms were inspired by social studies on cyberbullying [7, 20] and manual inspection of cyberbullying examples.

Roles in cyberbullying

Similarly to traditional bullying, cyberbullying involves a number of participants that adopt well-defined roles. Researchers have identified several roles in (cyber)bullying interactions. Although traditional studies on bullying have mainly concentrated on bullies and victims [24], the importance of bystanders in a bullying episode has been acknowledged [56, 57]. Bystanders can support the victim and mitigate the negative effects caused by the bullying [57], especially on social networking sites, where they hold higher intentions to help the victim than in real life conversations [58]. [25] distinguish three main types of bystanders: i) bystanders who participate in the bullying, ii) who help or support the victim and iii) those who ignore the bullying. Given that passive bystanders are hard to recognise in online text, only the former two are included in our annotation scheme.

Annotation guidelines

To operationalise the task of automatic cyberbullying detection, we elaborated a detailed annotation scheme for cyberbullying that is strongly embedded in the literature and applied it to our corpora. The applicability of the scheme was iteratively tested. Our final guidelines for the fine-grained annotation of cyberbullying are described in a technical report [23]. The objective of the scheme was to indicate several types of textual cyberbullying and verbal aggression, their severity, and the author participant roles. The scheme is formulated to be generic and is not limited to a specific social media platform. All messages were annotated in context (i.e. presented within their original content or conversation event) when available.

Essentially, the annotation scheme describes two levels of annotation. Firstly, the annotators were asked to indicate, at the message or post level, whether the text under investigation was related to cyberbullying. If the message was considered harmful and thus contained indications of cyberbullying, annotators identified the author's participant role. Based on the literature on role-allocation in cyberbullying episodes [25, 59], four roles are distinguished in the annotation scheme, including victim, bully, and two types of bystanders.

1. **Harasser or bully:** person who initiates the bullying.
2. **Victim:** person who is harassed.
3. **Bystander-defender:** person who helps the victim and discourages the harasser from continuing his actions.
4. **Bystander-assistant:** person who does not initiate, but helps or encourages the harasser.

Secondly, at the sub-sentence level, the annotators were tasked with the identification of fine-grained text categories related to cyberbullying. In the literature, different forms of cyberbullying are identified [12, 54, 55] and compared with traditional bullying [20]. Based on these forms, the annotation scheme describes a number of textual categories that are often inherent to a cyberbullying event, such as threats, insults, defensive statements from a victim, encouragements to the harasser, etc. Most of the categories are related to direct forms of cyberbullying (as defined by [25]), while one is related to *outing* [25], an indirect form of cyberbullying, namely *defamation*. Additionally, a number of subcategories were defined to make the annotation scheme as concrete and distinctive as possible (e.g., *discrimination* as a subcategory of *insult*). All cyberbullying-related categories in the scheme are listed below, and an example post for each category is presented in Table 1.

Table 1. Definitions and brat annotation examples of more fine-grained text categories related to cyberbullying.

Annotation category	Annotation example
Threat/blackmail	[I am going to find out who you are & I swear you are going to regret it.] ^{THREAT}
Insult	[Kill yourself] ^{CURS} [you fucking mc slut!!!!] ^{GEN. INSULT} [NO ONE LIKES YOU!!!!] ^{GEN. INSULT} [You are an ugly useless little whore!!!!] ^{GEN. INSULT}
Curse/Exclusion	[Fuck you.] ^{GEN. INSULT} [Now shush I don't wanna hear anything.] ^{CURSE OR EXCLUSION}
Defamation	[She slept with her ex behind his girlfriends back and she and him had broken up.] ^{DEFAMATION}
Sexual Talk	[Naked pic of you now.] ^{SEXUAL HARASSMENT}
Defense	[I would appreciate if you didn't talk shit about my bestfriend.] ^{GEN. VICTIM DEFENSE} He has enough to deal with already.
Encour. to har.	[She is a massive slut] ^{GEN. INSULT} [i agree with you @user she is!] ^{ENCOUR. HARASSER} [LOL AT HER mate, im on your side] ^{ENCOUR. HARASSER}

<https://doi.org/10.1371/journal.pone.0203794.t001>

- **Threat/blackmail:** expressions containing physical or psychological threats or indications of blackmail.
- **Insult:** expressions meant to hurt or offend the victim.
 - **General insult:** general expressions containing abusive, degrading or offensive language that are meant to insult the addressee.
 - **Attacking relatives:** insulting expressions towards relatives or friends of the victim.
 - **Discrimination:** expressions of unjust or prejudicial treatment of the victim. Two types of discrimination are distinguished (i.e. sexism and racism). Other forms of discrimination should be categorised as general insults.
- **Curse/exclusion:** expressions of a wish that some form of adversity or misfortune will befall the victim and expressions that exclude the victim from a conversation or a social group.
- **Defamation:** expressions that reveal confident or defamatory information about the victim to a large public.
- **Sexual Talk:** expressions with a sexual meaning or connotation. A distinction is made between innocent sexual talk and sexual harassment.
- **Defense:** expressions in support of the victim, expressed by the victim himself or by a bystander.
 - **Bystander defense:** expressions by which a bystander shows support for the victim or discourages the harasser from continuing his actions.
 - **Victim defense:** assertive or powerless reactions from the victim.
- **Encouragement to the harasser:** expressions in support of the harasser.
- **Other:** expressions that contain any other form of cyberbullying-related behaviour than the ones described here.

It is important to note that the categories were always indicated in text, even if the post in which they occurred was not considered harmful, for instance in the post “hi bitches, in for a movie?”, “bitches” was annotated as an insult while the post itself was not considered cyberbullying.

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 [60], some examples of which are presented in Table 1.

As can be deduced from the examples in the table, there were no restrictions as to what form the annotations could take. They could be adjectives, noun phrases, verb phrases, and so on. The only condition was that the annotation could not span more than one sentence and less than one word. Posts that were (primarily) written in another language than the corpus language (i.e. Dutch and English) were marked as such and required no further annotations.

We examined the validity of our guidelines and the annotations with an inter-annotator agreement experiment that is described in the following section.

Annotation statistics

The English and Dutch corpora were independently annotated for cyberbullying by trained linguists. All were Dutch native speakers and English second-language speakers. To demonstrate the validity of our guidelines, inter-annotator agreement scores were calculated using Kappa on a subset of each corpus. Inter-rater agreement for Dutch (2 raters) is calculated using Cohen's Kappa [61]. Fleiss' Kappa [62] is used for the English corpus (> 2 raters). Kappa scores for the identification of cyberbullying are $\kappa = 0.69$ (Dutch) and $\kappa = 0.59$ (English).

As shown in Table 2, inter-annotator agreement for the identification of the more fine-grained categories for English varies from fair to substantial [63], except for *defamation*, which appears to be more difficult to recognise. No encouragements to the harasser were present in this subset of the corpus. For Dutch, the inter-annotator agreement is fair to substantial, except for *curse* and *defamation*. Analysis revealed that one of both annotators often annotated the latter as an insult, and in some cases even did not consider it as cyberbullying-related.

In short, the inter-rater reliability study shows that the annotation of cyberbullying is not trivial and that more fine-grained categories like *defamation*, *curse* and *encouragements* are sometimes hard to recognise. It appears that defamations were sometimes hard to distinguish from insults, whereas curses and exclusions were sometimes considered insults or threats. The analysis further reveals that encouragements to the harasser are subject to interpretation. Some are straightforward (e.g. "I agree we should send her hate"), whereas others are subject to the annotator's judgment and interpretation (e.g. "hahaha", "LOL").

Experimental setup

In this paper, we explore the feasibility of automatically recognising signals of cyberbullying. A crucial difference with related research is that we do not only model bully 'attacks', but also more implicit forms of cyberbullying and reactions from victims and bystanders (i.e. all under one binary label 'signals of cyberbullying'), since these could likewise indicate that cyberbullying is going on. The experiments described in this paper focus on the automatic detection of such cyberbullying signals that need to be further investigated by human moderators when applied in a real-life moderation loop.

The English and Dutch corpus contain 113,698 and 78,387 posts, respectively. As shown in Table 3, the experimental corpus features a heavily **imbalanced** class distribution with the

Table 2. Inter-annotator agreement on the fine-grained categories related to cyberbullying.

	Threat	Insult	Defense	Sexual talk	Curse/exclusion	Defamation	Encouragements to the harasser
English	0.65	0.63	0.45	0.38	0.58	0.15	N/A
Dutch	0.52	0.66	0.63	0.53	0.19	0.00	0.21

<https://doi.org/10.1371/journal.pone.0203794.t002>

Table 3. Statistics of the English and Dutch cyberbullying corpus.

	Corpus size	Number(ratio) of bullying posts
English	113,698	5,375(4.73%)
Dutch	78,387	5,106(6.97%)

<https://doi.org/10.1371/journal.pone.0203794.t003>

large majority of posts not being part of cyberbullying. In classification, this class imbalance can lead to decreased performance. We apply cost-sensitive SVM as a possible hyperparameter in optimisation to counter this. The cost-sensitive SVM reweighs the penalty parameter C of the error term by the inverse class-ratio. This means that misclassifications of the minority positive class are penalised more than classification errors on the majority negative class. Other pre-processing methods to handle data imbalance in classification include feature filtering metrics and data resampling [64]. These methods were omitted as they were found to be too computationally expensive given our high-dimensional dataset.

For the automatic detection of cyberbullying, we performed **binary classification experiments** using a linear kernel support vector machine (SVM) implemented in LIBLINEAR [65] by making use of Scikit-learn [66], a machine learning library for Python. The motivation behind this is twofold: i) support vector machines (SVMs) have proven to work well for tasks similar to the ones under investigation [67] and ii) LIBLINEAR allows fast training of large-scale data which allow for a linear mapping (which was confirmed after a series of preliminary experiments using LIBSVM with linear, RBF and polynomial kernels).

The classifier was optimised for feature type (see the Pre-processing and feature engineering section) and hyperparameter combinations (see Table 4). Model selection was done using 10-fold cross validation in grid search over all possible feature types (i.e. groups of similar features, like different orders of n -gram bag-of-words features) and hyperparameter configurations. The best performing hyperparameters are selected by F_1 score on the positive class. The winning model is then retrained on all held-in data and subsequently tested on a hold-out test set to assess whether the classifier is over- or under-fitting. The hold-out set represents a random sample (10%) of all data. The folds were randomly stratified splits over the hold-in class distribution. Testing all feature type combinations is a rudimentary form of feature selection and provides insight into which types of features work best for this particular task.

Feature selection over all individual features was not performed because of the large feature space (NL: 795,072 and EN: 871,296 individual features). [68], among other researchers, demonstrated the importance of joint optimisation, where feature selection and hyperparameter optimisation are performed simultaneously, since the techniques mutually influence each other.

The optimised models are evaluated against **two baseline systems**: i) an unoptimised linear-kernel SVM (configured with default parameter settings) based on word n -grams only and, ii) a keyword-based system that marks posts as positive for cyberbullying if they contain a word from existing vocabulary lists composed by aggressive language and profanity terms.

Table 4. Hyperparameters in grid-search model selection.

Hyperparameter	Values
Penalty of error term C	$1e^{\{-3, -2, \dots, 2, 3\}}$
Loss function	Hinge, squared hinge
Penalty: norm used in penalisation	'l1' ('least absolute deviations') or 'l2' ('least squares')
Class weight (sets penalty C of class i to $\text{weight} \cdot C$)	None or 'balanced', i.e. weight inversely proportional to class frequencies

<https://doi.org/10.1371/journal.pone.0203794.t004>

Pre-processing and feature engineering

As pre-processing, we applied tokenisation, PoS-tagging and lemmatisation to the data using the LeT's Preprocess Toolkit [69]. In supervised learning, a machine learning algorithm takes a set of training instances (of which the label 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 cyberbullying from non-cyberbullying content.

We experimentally tested whether cyberbullying events can be recognised automatically 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 an existing mapping dictionary: <http://www.chatslang.com/terms/abbreviations/>). Additionally, tokenisation was applied before *n*-gram extraction and sentiment lexicon matching, and stemming was applied prior to extracting topic model features.

After pre-processing of the corpus, the following feature types were extracted:

- **Word *n*-gram bag-of-words:** binary features indicating the presence of word unigrams, bigrams and trigrams.
- **Character *n*-gram bag-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 characterises social media data.
- **Term lists:** one binary feature derived for each one out of six lists, indicating the presence of an item from the list in a post:
 - proper names: a gazetteer of named entities collected from several resources (e.g. Wikipedia).
 - 'allness' indicators (e.g. "always", "everybody"): forms which indicate rhetorical superlativity [70] which can be helpful in identifying the often hyperbolic bullying language.
 - diminishers (e.g. "slightly", "relatively"): diminishers, intensifiers and negation words were all obtained from an English grammar describing these lexical classes and existing sentiment lexicons (see further).
 - intensifiers (e.g. "absolutely", "amazingly")
 - negation words
 - aggressive language and profanity words: for English, we used the Google Profanity list (<https://code.google.com/archive/p/badwordslist/downloads>). For Dutch, a public profanity lexicon was consulted (<http://scheldwoorden.goedbegin.nl>).
Person alternation is a binary feature indicating whether the combination of a first and second person pronoun occurs in order to capture interpersonal intent.
- **Subjectivity lexicon features:** positive and negative opinion word ratios, as well as the overall post polarity were calculated using existing sentiment lexicons. For Dutch, we made use of the Duoman [71] and Pattern [72] lexicons. For English, we included the Liu and Hu opinion lexicon [73], the MPQA lexicon [74], the General Inquirer Sentiment Lexicon [75],

AFINN [76], and MSOL [77]. For both languages, we included the relative frequency of all 68 psychometric categories in the Linguistic Inquiry and Word Count (LIWC) dictionary for English [78] and Dutch [79].

- **Topic model features:** by making use of the Gensim topic modelling library [80], several LDA [81] and LSI [82] topic models with varying granularity ($k = 20, 50, 100$ and 200) were trained on data corresponding to each fine-grained category of a cyberbullying event (e.g. threats, defamations, insults, defenses). The topic models were based on a background corpus (EN: $\pm 1,200,000$ tokens, NL: $\pm 1,400,000$ tokens) scraped with the BootCaT [83] web-corpus toolkit. BootCaT collected ASKfm user profiles using lists of manually determined seed words that are characteristic of the cyberbullying categories.

When applied to the training data, this resulted in 871,296 and 795,072 features for English and Dutch, respectively.

Results

In this section, we present the results of our experiments to automatically detect cyberbullying signals in an English and Dutch corpus of ASKfm posts. Ten-fold cross-validation was performed in exhaustive grid search over different feature type and hyperparameter combinations (see the [Experimental setup](#) section). The *unoptimised word n -gram-based* classifier and *key-word-matching* system serve as baselines for comparison. Precision, Recall and F_1 performance metrics were calculated on the positive class. We also report Area Under the Receiver Operator Curve (AUROC) scores, a performance metric that is more robust to data imbalance than precision, recall and F score [84].

[Table 5](#) gives us an indication of which feature type combinations score best and hence contribute most to this task. It presents the cross-validation and hold-out scores of a set of feature

Table 5. Cross-validated and hold-out scores (%) according to different metrics (F_1 , precision, recall, accuracy and area under the curve) for the English and Dutch three best and worst combined feature type systems.

	Feature combination	Cross-validation scores					Hold-out scores				
		F ₁	P	R	Acc	AUROC	F ₁	P	R	Acc	AUROC
English											
Best three	B + C + D + E	64.26	73.32	57.19	96.97	78.07	63.69	74.13	55.82	97.21	77.47
	A + B + C	64.24	73.22	57.23	96.96	78.09	64.32	74.08	56.83	97.24	77.96
	A + C + E	63.84	73.21	56.59	96.94	77.78	62.94	72.82	55.42	97.14	77.24
Worst three	D	40.48	38.98	42.12	94.10	69.41	39.56	39.56	39.56	94.71	68.39
	A + D + E	38.95	31.47	51.10	92.37	72.76	40.71	33.87	51.00	93.49	73.22
	E	17.35	9.73	79.91	63.72	71.41	15.70	8.72	78.51	63.07	70.44
Baseline	word n -gram	58.17	67.55	51.07	96.54	74.93	59.63	69.57	52.17	96.57	75.50
	profanity	17.17	9.61	80.14	63.73	71.53	17.61	9.90	78.51	63.79	71.34
Dutch											
Best three	A + B + C + E	61.20	56.76	66.40	94.47	81.42	58.13	54.03	62.90	94.58	79.75
	A + B + C + D + E	61.03	71.55	53.20	95.53	75.86	58.72	67.40	52.03	95.62	75.21
	A + C + E	60.82	71.66	52.84	95.53	75.68	58.15	67.71	50.96	95.61	74.71
Worst three	D + B	32.90	29.23	37.63	89.91	65.61	30.16	34.72	26.65	92.61	61.73
	D	28.65	19.36	55.10	81.97	69.48	25.13	16.73	50.53	81.99	67.26
	B	24.74	21.24	29.61	88.16	60.94	17.99	23.15	14.71	91.98	55.80
Baseline	word n -gram	50.39	67.80	40.09	94.81	69.38	49.54	64.29	40.30	95.09	69.44
	profanity	28.46	19.24	54.66	81.99	69.28	25.13	16.73	50.53	81.99	67.26

<https://doi.org/10.1371/journal.pone.0203794.t005>

Table 6. Feature group mapping (Table 5).

A	word <i>n</i> -grams
B	subjectivity lexicons
C	character <i>n</i> -grams
D	term lists
E	topic models

<https://doi.org/10.1371/journal.pone.0203794.t006>

combinations, which are explained in the feature groups legend (Table 6). A total of 31 feature type combinations, each with 28 different hyperparameter sets have been tested. Table 5 shows the results for the three best scoring systems by included feature types with optimised hyperparameters. The maximum obtained F_1 score in cross-validation is 64.26% for English and 61.20% for Dutch and shows that the classifier benefits from a variety of feature types. The results on the hold-out test set show that the trained systems generalise well on unseen data, indicating little under- or overfitting. The simple keyword-matching baseline system has the lowest performance for both languages even though it obtains high recall for both languages, especially for English (80.14%), suggesting that profane language characterises many cyberbullying-related posts. Feature group and hyperparameter optimisation provides a considerable performance increase over the unoptimised word *n*-gram baseline system. The top-scoring systems for each language do not differ a lot in performance, except the best system for Dutch, which trades recall for precision when compared to the runner-ups.

Table 7 presents the scores of the (hyperparameter-optimised) single feature type systems, to gain insight into the performance of these feature types when used individually. Analysis of the combined and single feature type sets reveals that **word *n*-grams**, **character *n*-grams**, and **subjectivity lexicons** prove to be strong features for this task. In effect, adding character *n*-grams always improved classification performance for both languages. They are likely to provide robustness to lexical variation in social media text, as compared to word *n*-grams. While subjectivity lexicons appear to be discriminative features, term lists perform badly on their own as well as in combinations for both languages. This shows once again (see the profanity baseline) that cyberbullying detection requires more sophisticated information sources than profanity lists. Topic models seem to do badly for both languages on their own, but in

Table 7. Cross-validated and hold-out scores (%) according to different metrics (F_1 , precision, recall, accuracy and area under the ROC curve) for English and Dutch single feature type systems.

	Feature type	Cross-validation scores					hold-out scores				
		F ₁	P	R	Acc	AUROC	F ₁	P	R	Acc	AUROC
English											
	word <i>n</i> -grams	60.09	60.49	59.69	96.22	78.87	58.35	57.12	59.64	96.27	78.79
	subjectivity lexicons	56.82	73.32	46.38	96.64	72.77	56.16	72.61	45.78	96.87	72.50
	character <i>n</i> -grams	52.69	58.70	47.80	95.91	73.06	53.33	62.37	46.59	96.43	72.65
	term lists	40.48	38.98	42.12	94.10	69.41	39.56	39.56	39.56	94.71	68.39
	topic models	17.35	9.73	79.91	63.72	71.41	15.70	8.72	78.51	63.07	70.44
Dutch											
	word <i>n</i> -grams	55.53	72.64	44.94	95.27	71.88	54.99	70.20	45.20	95.57	71.99
	subjectivity lexicons	54.34	54.12	54.56	93.97	75.65	51.82	50.61	53.09	94.09	74.90
	character <i>n</i> -grams	51.70	67.58	41.86	94.86	70.22	50.46	65.20	41.15	95.17	69.88
	term lists	28.65	19.36	55.10	81.97	69.48	25.13	16.73	50.53	81.99	67.26
	topic models	24.74	21.24	29.61	88.16	60.94	17.99	23.15	14.71	91.98	55.80

<https://doi.org/10.1371/journal.pone.0203794.t007>

combination with other features, they improve Dutch performance consistently. A possible explanation for their varying performance in both languages would be that the topic models trained on the Dutch background corpus are of better quality than the English ones. In effect, a random selection of background corpus texts reveals that the English scrape contains more noisy data (i.e. low word-count posts and non-English posts) compared to the Dutch scraped corpus.

A shallow **qualitative analysis** of the classification output provided insight into some of the classification mistakes.

Table 8 gives an overview of the error rates per cyberbullying category of the best performing and baseline systems. This could give an indication of the types of bullying are hard to detect by the current classifier. All categories are always considered positive for cyberbullying (i.e. the error rate equals the false negative rate), except for *Sexual* and *Insult* which can also be negative (in case of harmless sexual talk and ‘socially acceptable’ insulting language like “hi bitches, in for a movie?” the corresponding category was indicated, but the post itself was not annotated as cyberbullying) and *Not cyberbullying*, which is always negative. Error rates often being lowest for the profanity baseline confirms that it performs particularly well in terms of recall (at the expense of precision, see Table 5). When looking at the best system for both languages, we see that *Defense* is the hardest category to classify. This should not be a surprise as the category comprises defensive posts from bystanders and victims, which contain less aggressive language than cyberbullying attacks and are often shorter in length than the latter. Assertive defensive posts (i.e. a subcategory of *Defense*) which attack the bully are, however, more often correctly classified. There are not sufficient instances of the *Encouragement* class for either language in the hold-out set to be representative. In both languages, threats, curses and incidences of sexual harassment are most easily recognisable, showing (far) lower error rates than the categories *Defamation*, *Defense*, *Encouragements to the harasser*, and *Insult*.

A qualitative error analysis of the English and Dutch predictions reveals that false positives often contain aggressive language directed at a second person, often denoting personal flaws

Table 8. Error rates (%) per cyberbullying subcategory on hold-out for English and Dutch systems.

	Category	Nr. occurrences in hold-out	Profanity baseline	Word <i>n</i> -gram baseline	Best system
English	Curse	<i>n</i> = 109	14.68	30.28	24.77
	Defamation	<i>n</i> = 21	23.81	47.62	38.10
	Defense	<i>n</i> = 165	22.42	52.12	43.64
	Encouragement	<i>n</i> = 1	0.00	100.00	100.00
	Insult	<i>n</i> = 345	26.67	41.74	35.94
	Sexual	<i>n</i> = 165	63.80	21.47	21.47
	Threat	<i>n</i> = 12	8.33	41.67	25.00
	Not cyberbullying	<i>n</i> = 10,714	36.94	1.10	0.76
Dutch	Curse	<i>n</i> = 96	39.58	50.00	22.92
	Defamation	<i>n</i> = 6	100.00	66.67	33.33
	Defense	<i>n</i> = 200	52.50	63.50	46.00
	Encouragement	<i>n</i> = 5	40.00	60.00	40.00
	Insult	<i>n</i> = 355	43.38	47.89	28.17
	Sexual	<i>n</i> = 37	37.84	21.62	27.03
	Threat	<i>n</i> = 15	33.33	46.67	20.00
	Not cyberbullying	<i>n</i> = 7,295	15.63	1.23	3.07

<https://doi.org/10.1371/journal.pone.0203794.t008>

or containing sexual and profanity words. We see that misclassifications are often short posts containing just a few words and that false negatives often lack explicit verbal signs of cyberbullying (e.g. insulting or profane words) or are ironic (examples 2 and 3). Additionally, we see that cyberbullying posts containing misspellings or grammatical errors and incomplete words are also hard to recognise as such (examples 4 and 5). The Dutch and English corpus are overall similar with respect to qualitative properties of classification errors.

2. *You might want to do some sports ahah x*
3. *Look who is there... my thousandth anonymous hater, congratulations!*
4. *ivegot 1 word foryou... yknow whatit is? → slut*
5. *One word for you: G—A—...*

In short, the experiments show that our classifier clearly outperforms both a keyword-based and word n -gram baseline. However, analysis of the classifier output reveals that false negatives often lack explicit clues that cyberbullying is going on, indicating that our system might benefit from irony recognition and integrating world knowledge to capture such implicit realisations of cyberbullying.

Our annotation scheme allowed to indicate different author roles, which provides better insight into the realisation of cyberbullying. Table 9 presents the error rates of our classifier for the different author roles, being harasser, victim, and two types of bystanders. We observe that the error rates are high for *bystander assistant* and *victim*, but there are not sufficient instances in the hold-out set of the former role for either language to be representative. Error rates for the *victim* class of 50.39% and 54% in English and Dutch respectively indicate that the role is hard to recognise by the classifier. A possible explanation for this could be that victim posts in our corpus either expressed powerlessness facing the bully (example 6) or either contained explicit aggressive language as well (example 7).

6. *Your the one going round saying im a cunt and a twat and im ugly. tbh all im doing is sticking up for myself.*
7. *You're fucked up saying I smell from sweat, because unlike some other people I shower every day BITCH*

According to the figures, the most straightforward roles in detection are *bystander defender* and *harasser*.

Table 9. Error rates (%) per cyberbullying participant role on hold-out for English and Dutch systems.

	Participant role	Nr. occurrences in hold-out	Profanity baseline	Word n -gram baseline	Best system
English					
	Harasser	$n = 328$	20.43	48.48	43.60
	Bystander assistant	$n = 2$	50.00	100.00	100.00
	Bystander defender	$n = 39$	7.69	38.46	25.64
	Victim	$n = 129$	27.91	57.36	50.39
	Not cyberbullying	$n = 10872$	37.64	1.24	0.89
Dutch					
	Harasser	$n = 261$	47.13	56.70	29.89
	Bystander assistant	$n = 6$	50.00	66.67	50.00
	Bystander defender	$n = 52$	25.00	38.46	23.08
	Victim	$n = 150$	62.00	72.00	54.00
	Not cyberbullying	$n = 7370$	16.01	1.42	3.41

<https://doi.org/10.1371/journal.pone.0203794.t009>

Table 10. Overview of the most related cyberbullying detection approaches.

Reference	Classifier	Corpus	Bully rate	F ₁ score
[44]	SVM	1,762 tweets	39%	77%
[43]	wvec+SVM	1,762 tweets	39%	78%
[45]	smSDA+SVM	7,321 tweets	29%	72%
[45]	smSDA+SVM	1,539 MySpace posts	26%	78%

<https://doi.org/10.1371/journal.pone.0203794.t010>

In the light of comparison with state-of-the-art approaches to cyberbullying, we observe that competitive results are obtained with regard to [30–32, 41]. However, the fundamental differences with respect to data collection, sources, and conceptualisations of bullying hardly allow for direct comparison. Table 10 presents the experimental results obtained by [43–45] who, like the current study, approach the task as detecting posts from bullies as well as from victims and bystanders. Given their experimental setup (i.e. task description, data genre and classifier), their work can be considered most similar to ours so their results might function as benchmarks. Also here, a number of crucial differences with the current approach can be observed: Firstly, their corpora were collected using the keywords “bully”, “bullying” and “bullied”, which may bias the dataset towards the positive class and ensures that many explicit lexicalisations are present in the positive class. Second, it is not clear which types of cyberbullying (i.e. explicit and implicit bullying, threats, insults, sexual harassment) are included in the positive class. Furthermore, as can be deduced from Table 10, the datasets are considerably smaller than ours and show a more balanced class distribution (respectively 39% cyberbullying posts in [43] and [44], and 29%/26% in [45]) than the ratio of bullying posts in our corpus (see Table 3: 5% for English, 7% for Dutch). Hence, any comparison should be made with caution due to these differences.

These studies obtain higher scores on similar task but vastly different datasets. Notably, [45] shows a great improvement in classification performance using deep representational learning with a semantic-enhanced marginalized denoising auto-encoder over traditional *n*-gram and topic modelling features.

Conclusions and future research

The goal of the current research was to investigate the automatic detection of cyberbullying-related posts on social media. Given the information overload on the web, manual monitoring for cyberbullying has become unfeasible. Automatic detection of signals of cyberbullying would enhance moderation and allow to respond quickly when necessary.

Cyberbullying research has often focused on detecting cyberbullying ‘attacks’ and hence overlook other or more implicit forms of cyberbullying and posts written by victims and bystanders. However, these posts could just as well indicate that cyberbullying is going on. The main contribution of this paper is that it presents a system to automatically detect **signals of cyberbullying** on social media, including different types of cyberbullying, covering posts from bullies, victims and bystanders. We evaluated our system on a manually annotated cyberbullying corpus for English and Dutch and hereby demonstrated that our approach can easily be applied to different languages, provided that annotated data for these languages are available.

A set of binary classification experiments were conducted to explore the feasibility of automatic cyberbullying detection on social media. In addition, we sought to determine which information sources contribute most to the task. Two classifiers were trained on an English and Dutch ASKfm corpus and evaluated on a hold-out test of the same genre. Our experiments reveal that the current approach is a promising strategy for detecting signals of cyberbullying

on social media automatically. After feature and hyperparameter optimisation of our models, a maximum F_1 score of 64.32% and 58.72% was obtained for English and Dutch, respectively. The classifiers hereby significantly outperformed a keyword and an (unoptimised) n -gram baseline. A qualitative analysis of the results revealed that false positives often include implicit cyberbullying or offenses through irony, the challenge of which will constitute an important area for future work. Error rates on the different author roles in our corpus revealed that especially victims are hard to recognise, as they react differently in our corpus, showing either powerlessness facing the bully or reacting in an assertive and sometimes even aggressive way.

As shown in [45] deep representation learning is a promising avenue for this task. We therefore intent to apply deep learning techniques to improve classifier performance.

Another interesting direction for future work would be the detection of fine-grained cyberbullying categories such as threats, curses and expressions of racism and hate. When applied in a cascaded model, the system could find severe cases of cyberbullying with high precision. This would be particularly interesting for monitoring purposes. Additionally, our dataset allows for detection of participant roles typically involved in cyberbullying. When applied as moderation support on online platforms, such a system enables feedback in function of the recipient (i.e. a bully, victim, or bystander).

Author Contributions

Conceptualization: Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Véronique Hoste.

Data curation: Cynthia Van Hee, Gilles Jacobs, Els Lefever, Ben Verhoeven, Guy De Pauw.

Funding acquisition: Walter Daelemans, Véronique Hoste.

Methodology: Cynthia Van Hee, Gilles Jacobs, Bart Desmet, Els Lefever, Walter Daelemans, Véronique Hoste.

Project administration: Walter Daelemans, Véronique Hoste.

Resources: Cynthia Van Hee, Gilles Jacobs, Els Lefever.

Software: Cynthia Van Hee, Gilles Jacobs, Bart Desmet.

Supervision: Walter Daelemans, Véronique Hoste.

Writing – original draft: Cynthia Van Hee, Gilles Jacobs.

Writing – review & editing: Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Véronique Hoste.

References

1. Livingstone S, Haddon L, Görzig A, Ólafsson K. Risks and safety on the internet: The perspective of European children. Initial Findings. London: EU Kids Online; 2010.
2. Tokunaga RS. Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. *Computers in Human Behavior*. 2010; 26(3):277–287. <https://doi.org/10.1016/j.chb.2009.11.014>
3. McKenna KY, Bargh JA. Plan 9 From Cyberspace: The Implications of the Internet for Personality and Social Psychology. *Personality & Social Psychology Review*. 1999; 4(1):57–75. https://doi.org/10.1207/S15327957PSPR0401_6
4. Gross EF, Juvonen J, Gable SL. Internet Use and Well-Being in Adolescence. *Journal of Social Issues*. 2002; 58(1):75–90. <https://doi.org/10.1111/1540-4560.00249>

5. Juvonen J, Gross EF. Extending the school grounds?—Bullying experiences in cyberspace. *Journal of School Health*. 2008; 78(9):496–505. <https://doi.org/10.1111/j.1746-1561.2008.00335.x> PMID: 18786042
6. Hinduja S, Patchin JW. Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying. *Youth Violence And Juvenile Justice*. 2006; 4(2):148–169. <https://doi.org/10.1177/1541204006286288>
7. Van Cleemput K, Bastiaensens S, Vandebosch H, Poels K, Deboutte G, DeSmet A, et al. Zes jaar onderzoek naar cyberpesten in Vlaanderen, België en daarbuiten: een overzicht van de bevindingen. (Six years of research on cyberbullying in Flanders, Belgium and beyond: an overview of the findings.) (White Paper). University of Antwerp & Ghent University; 2013.
8. Livingstone S, Haddon L, Vincent J, Giovanna M, Ólafsson K. Net Children Go Mobile: The Uk report; 2014. Available from: <http://netchildrengomobile.eu/reports>. [Accessed 30th March 2018].
9. O'Moore M, Kirkham C. Self-esteem and its relationship to bullying behaviour. *Aggressive Behavior*. 2001; 27(4):269–283. <https://doi.org/10.1002/ab.1010>
10. Fekkes M, Pijpers FIM, Fredriks AM, Vogels T, Verloove-Vanhorick SP. Do Bullied Children Get Ill, or Do Ill Children Get Bullied? A Prospective Cohort Study on the Relationship Between Bullying and Health-Related Symptoms. *Pediatrics*. 2006; 117(5):1568–1574. <https://doi.org/10.1542/peds.2005-0187> PMID: 16651310
11. Cowie H. Cyberbullying and its impact on young people's emotional health and well-being. *The Psychiatrist*. 2013; 37(5):167–170. <https://doi.org/10.1192/pb.bp.112.040840>
12. Price M, Dalgleish J. Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People. *Youth Studies Australia*. 2010; 29(2):51–59.
13. Van Royen K, Poels K, Daelemans W, Vandebosch H. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*. 2014;.
14. Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning*. 1995; 20(3):273–297. <https://doi.org/10.1007/BF00994018>
15. Chang CC, Lin CJ. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2011; 2(3):27:1–27:27.
16. Dalla Pozza V, Di Pietro A, Morel S, Emma P. Cyberbullying among Young People. European Parliament; 2016.
17. Olweus D. Bullying at School: What We Know and What We Can Do. 2nd ed. Wiley; 1993.
18. Smith PK, Mahdavi J, Carvalho M, Fisher S, Russell S, Tippett N. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*. 2008; 49(4):376–385. <https://doi.org/10.1111/j.1469-7610.2007.01846.x> PMID: 18363945
19. Vandebosch H, Van Cleemput K. Defining cyberbullying: a qualitative research into the perceptions of youngsters. *Cyberpsychology and behavior: the impact of the Internet, multimedia and virtual reality on behavior and society*. 2008; 11(4):499–503. <https://doi.org/10.1089/cpb.2007.0042>
20. Vandebosch H, Van Cleemput K. Cyberbullying among youngsters: profiles of bullies and victims. *New Media & Society*. 2009; 11(8):1349–1371. <https://doi.org/10.1177/1461444809341263>
21. Dooley JJ, Cross D. Cyberbullying versus face-to-face bullying: A review of the similarities and differences. *Journal of Psychology*. 2009; 217:182–188.
22. Slonje R, Smith PK, Frisén A. The Nature of Cyberbullying, and Strategies for Prevention. *Computers in Human Behavior*. 2013; 29(1):26–32. <https://doi.org/10.1016/j.chb.2012.05.024>
23. Van Hee C, Verhoeven B, Lefever E, De Pauw G, Daelemans W, Hoste V. Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0. LT3, Language and Translation Technology Team—Ghent University; 2015. LT3 15-01.
24. Salmivalli C, Lagerspetz K, Björkqvist K, Österman K, Kaukiainen A. Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior*. 1996; 22(1):1–15. [https://doi.org/10.1002/\(SICI\)1098-2337\(1996\)22:1%3C1::AID-AB1%3E3.0.CO;2-T](https://doi.org/10.1002/(SICI)1098-2337(1996)22:1%3C1::AID-AB1%3E3.0.CO;2-T)
25. Vandebosch H, Van Cleemput K, Mortelmans D, Walrave M. Cyberpesten bij jongeren in Vlaanderen: Een studie in opdracht van het viWTA (Cyberbullying among youngsters in Flanders: a study commissioned by the viWTA). Brussels: viWTA; 2006. Available from: <https://wise.vub.ac.be/fattac/mios/Eindrapport%20cyberpesten%20viwta%202006.pdf>. [Accessed 30th March 2018].
26. Dehue F, Bolman C, Völlink T. Cyberbullying: Youngsters' Experiences and Parental Perception. *Cyberpsychology and Behavior*. 2008; 11(2):217–223. <https://doi.org/10.1089/cpb.2007.0008> PMID: 18422417
27. Salmivalli C, Pöyhönen V. In: *Cyberbullying in Finland*. 2nd ed. Wiley-Blackwell; 2012. p. 57–72.
28. Salawu S, he Y, Lumsden J. Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Transactions on Affective Computing*. forthcoming;p. 1.

29. Reynolds K, Kontostathis A, Edwards L. 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; 2011. p. 241–244.
30. Dinakar K, Reichart R, Lieberman H. Modeling the Detection of Textual Cyberbullying. In: The Social Mobile Web. vol. WS-11-02 of AAAI Workshops. AAAI; 2011. p. 11–17.
31. Yin D, Davison BD, Xue Z, Hong L, Kontostathis A, Edwards L. Detection of Harassment on Web 2.0. In: Proceedings of the Content Analysis in the Web 2.0 (CAW2.0). Madrid, Spain; 2009.
32. Dadvar M. Experts and machines united against cyberbullying [PhD thesis]. University of Twente; 2014.
33. Nahar V, Al-Maskari S, Li X, Pang C. Semi-supervised Learning for Cyberbullying Detection in Social Networks. In: ADC.Databases Theory and Applications. Springer International Publishing; 2014. p. 160–171.
34. Bayzick J, Kontostathis A, Edwards L. Detecting the Presence of Cyberbullying Using Computer Software. In: Proceedings of the 3rd International Web Science Conference. WebSci11; 2011.
35. Sui J. Understanding and Fighting Bullying with Machine Learning [PhD thesis]. Department of Computer Sciences, University of Wisconsin-Madison; 2015.
36. Galán-García P, Puerta JGdl, Gómez CL, Santos I, Bringas PG. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. Logic Journal of the IGPL. 2016; 24(1):42–53. Available from: <http://dx.doi.org/10.1093/jigpal/jzv048>.
37. Raisi E, Huang B. Cyberbullying Detection with Weakly Supervised Machine Learning. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. ASONAM'17. New York, NY, USA: ACM; 2017. p. 409–416.
38. Hosseinmardi H. Survey of Computational Methods in Cyberbullying Research. In: Proceedings of the First International Workshop on Computational Methods for CyberSafety. CyberSafety'16. New York, NY, USA: ACM; 2016. p. 4–4.
39. Nandhini B Sri, Sheeba JI. Online Social Network Bullying Detection Using Intelligence Techniques. Procedia Computer Science. 2015; 45:485–492. <https://doi.org/10.1016/j.procs.2015.03.085>
40. Chavan VS, S SS. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2015. p. 2354–2358.
41. Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, et al. Detection and fine-grained classification of cyberbullying events. In: Angelova G, Bontcheva K, Mitkov R, editors. Proceedings of Recent Advances in Natural Language Processing, Proceedings; 2015. p. 672–680.
42. Van Royen K, Poels K, Vandebosch H. Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites. Children and Youth Services Review. 2016; 64:35–41. <https://doi.org/10.1016/j.childyouth.2016.02.024>
43. Zhao R, Zhou A, Mao K. Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. In: Proceedings of the 17th International Conference on Distributed Computing and Networking. No. 43 in ICDCN'16. New York, NY, USA: ACM; 2016. p. 43:1–43:6.
44. Xu JM, Jun KS, Zhu X, Bellmore A. 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; 2012. p. 656–666.
45. Zhao R, Mao K. Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. IEEE Transactions on Affective Computing. 2017; 8(3):328–339. <https://doi.org/10.1109/TAFFC.2016.2531682>
46. Huang Q, Singh VK, Atrey PK. Cyber Bullying Detection Using Social and Textual Analysis. In: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. SAM'14. New York, NY, USA: ACM; 2014. p. 3–6.
47. Al-garadi MA, Dewi Varathan K, Devi Ravana S. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior. 2016; 63:433–443. <https://doi.org/10.1016/j.chb.2016.05.051>
48. Squicciarini A, Rajtmajer S, Liu Yh, Griffin C. Identification and Characterization of Cyberbullying Dynamics in an Online Social Network. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ASONAM'15. New York, NY, USA: ACM; 2015. p. 280–285.
49. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A. Mean Birds: Detecting Aggression and Bullying on Twitter. In: Proceedings of the 2017 ACM on Web Science Conference. WebSci'17. New York, NY, USA: ACM; 2017. p. 13–22.

50. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer PW. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research (JAIR)*. 2002; 16:321–357. <https://doi.org/10.1613/jair.953>
51. Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, et al. Automatic detection and prevention of cyberbullying. In: Lorenz P, Bourret C, editors. *International Conference on Human and Social Analytics, Proceedings. IARIA*; 2015. p. 13–18.
52. Hinduja S, Patchin JW. Cyberbullying: Neither an epidemic nor a rarity. *European Journal of Developmental Psychology*. 2012; 9(5):539–543. <https://doi.org/10.1080/17405629.2012.706448>
53. Slonje R, Smith PK. Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*. 2008; 49(2):147–154. <https://doi.org/10.1111/j.1467-9450.2007.00611.x> PMID: 18352984
54. O'Sullivan PB, Flanagan AJ. Reconceptualizing 'flaming' and other problematic messages. *New Media & Society*. 2003; 5(1):69–94. <https://doi.org/10.1177/1461444803005001908>
55. Willard NE. *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*. 2nd ed. Research Publishers LLC; 2007.
56. Bastiaensens S, Vandebosch H, Poels K, Van Cleemput K, DeSmet A, De Bourdeaudhuij I. Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*. 2014; 31:259–271. <https://doi.org/10.1016/j.chb.2013.10.036>
57. Salmivalli C. Bullying and the peer group: A review. *Aggression and Violent Behavior*. 2010; 15(2):112–120. <https://doi.org/10.1016/j.avb.2009.08.007>
58. Bastiaensens S, Vandebosch H, Poels K, Van Cleemput K, DeSmet A, De Bourdeaudhuij I. 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology*. 2015; 34(4):425–435. <https://doi.org/10.1080/0144929X.2014.983979>
59. Salmivalli C, Voeten M, Poskiparta E. Bystanders Matter: Associations Between Reinforcing, Defending, and the Frequency of Bullying Behavior in Classrooms. *Journal of Clinical Child & Adolescent Psychology*. 2011; 40(5):668–676. <https://doi.org/10.1080/15374416.2011.597090>
60. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation. In: *Proceedings of the Demonstrations Session at EACL 2012*. Avignon, France; 2012. p. 102–107.
61. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960; 20(1):37–46. <https://doi.org/10.1177/001316446002000104>
62. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 1971; 76(5):378–382. <https://doi.org/10.1037/h0031619>
63. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica*. 2012; 22(3):276–282. <https://doi.org/10.11613/BM.2012.031> PMID: 23092060
64. He H, Garcia EA. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*. 2009; 21(9):1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
65. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*. 2008; 9:1871–1874.
66. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011; 12:2825–2830.
67. Desmet B. Finding the online cry for help: automatic text classification for suicide prevention [PhD thesis]. Ghent University; 2014.
68. Hoste V. *Optimization Issues in Machine Learning of Coreference Resolution* [PhD thesis]. Antwerp University; 2005.
69. van de Kauter M, Coorman G, Lefever E, Desmet B, Macken L, Hoste V. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*. 2013; 3:103–120.
70. Osgood CE, Walker EG. Motivation and language behavior: A content analysis of suicide notes. *The Journal of Abnormal and Social Psychology*. 1959; 59(1):58. <https://doi.org/10.1037/h0047078>
71. Jijkoun V, Hofmann K. 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; 2009. p. 398–405.
72. De Smedt T, Daelemans W. "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; 2012. p. 3568–3572.

73. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD04. ACM; 2004. p. 168–177.
74. Wilson T, Wiebe J, Hoffmann P. 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. Association for Computational Linguistics; 2005. p. 347–354.
75. Stone PJ, Dunphy DCD, Smith MS, Ogilvie DM. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press; 1966.
76. Nielsen FÅ. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In: Rowe M, Stankovic M, Dadzie AS, Hardey M, editors. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. vol. 718 of CEUR Workshop Proceedings. CEUR-WS.org; 2011. p. 93–98.
77. Mohammad S, Dunne C, Dorr B. 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; 2009. p. 599–608.
78. Pennebaker JW, Francis ME, Booth RJ. *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.
79. Zijlstra H, Van Meerveld T, Van Middendorp H, Pennebaker JW, Geenen R. De Nederlandse versie van de 'linguistic inquiry and word count' (LIWC). *Gedrag Gezond*. 2004; 32:271–281.
80. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: The LREC 2010 Workshop on new Challenges for NLP Frameworks. University of Malta; 2010. p. 45–50.
81. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of Machine Learning Research*. 2003; 3:993–1022.
82. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. 1990; 41:391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6%3C391::AID-AS11%3E3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-AS11%3E3.0.CO;2-9)
83. Baroni M, Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. LREC'04; 2004. p. 1313–1316.
84. Fawcett T. An introduction to ROC analysis. *Pattern recognition letters*. 2006; 27(8):861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>