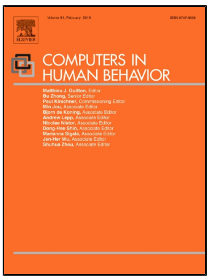
# Accepted Manuscript




Automatic cyberbullying detection: A systematic review.

H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.M. Veiga Simão, I. Trancoso

| | |
|---|---|
| PII: | S0747-5632(18)30607-1 |
| DOI: | 10.1016/j.chb.2018.12.021 |
| Reference: | CHB 5843 |
| To appear in: | *Computers in Human Behavior* |
| Received Date: | 02 October 2018 |
| Accepted Date: | 11 December 2018 |

Please cite this article as: H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.M. Veiga Simão, I. Trancoso, Automatic cyberbullying detection: A systematic review., *Computers in Human Behavior* (2018), doi: 10.1016/j.chb.2018.12.021

**Automatic cyberbullying detection: A systematic review.**

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A. M., Trancoso, I.

**Authors' names and affiliations:**

**Hugo Rosa,**

INESC-ID, Institute of System and Computer Engineering, Research and Development of Lisbon, Instituto Superior Técnico of Lisbon, Lisbon, Portugal

Foundation for Science and Technology (FPUL/FCT/BI/07)

**Nádia Pereira,**

Faculty of Psychology of the University of Lisbon, Lisbon, Portugal

CICPSI, Research Center for Psychological Science, Faculty of Psychology, University of Lisbon, Lisbon, Portugal

Foundation for Science and Technology (FPUL/FCT/BI/2017/12)

**Ricardo Ribeiro,**

INESC-ID, Institute of System and Computer Engineering, Research and Development of Lisbon, Instituto Superior Técnico of Lisbon, Lisbon, Portugal

ISCTE-IUL, Instituto Universitário de Lisboa, Lisbon, Portugal

**Paula da Costa Ferreira,**

Faculty of Psychology of the University of Lisbon, Lisbon, Portugal

CICPSI, Research Center for Psychological Science, Faculty of Psychology, University of Lisbon, Lisbon, Portugal

INESC-ID, Institute of System and Computer Engineering, Research and Development of Lisbon, Instituto Superior Técnico of Lisbon, Lisbon, Portugal

**João Paulo Carvalho,**

Instituto Superior Técnico, Lisbon, Portugal

INESC-ID, Institute of System and Computer Engineering, Research and Development of Lisbon, Instituto Superior Técnico of Lisbon, Lisbon, Portugal

**Sofia Oliveira,**

Faculty of Psychology of the University of Lisbon, Lisbon, Portugal

CICPSI, Research Center for Psychological Science, Faculty of Psychology, University of Lisbon, Lisbon, Portugal

Foundation for Science and Technology (FPUL/FCT/BI/2017/13)

**Luisa Coheur,**

Instituto Superior Técnico, Lisbon, Portugal

INESC-ID, Institute of System and Computer Engineering, Research and Development of Lisbon, Instituto Superior Técnico of Lisbon, Lisbon, Portugal

**Paula Paulino,**

CICPSI, Research Center for Psychological Science, Faculty of Psychology, University of Lisbon, Lisbon, Portugal

University Lusófona, Lisbon, Portugal

**Ana Margarida Veiga Simão,**

Faculty of Psychology of the University of Lisbon, Lisbon, Portugal

CICPSI, Research Center for Psychological Science, Faculty of Psychology, University of Lisbon, Lisbon, Portugal

**Isabel Trancoso,**

Instituto Superior Técnico, Lisbon, Portugal

INESC-ID, Institute of System and Computer Engineering, Research and Development of Lisbon, Instituto Superior Técnico of Lisbon, Lisbon, Portugal

**Corresponding author contact details:**

Nádia Pereira

+351 913731461

Email: nadia@campus.ul.pt

**Acknowledgements:**

This work was supported by the Foundation for Science and Technology (FCT) of the Science and Education Ministry of Portugal (PTDC/MHC/PED/3297/2014; SFRH/BPD/110695/2015), along with the Research Center for Psychological Science (CICPSI; UID/PSI/4527/2016), and in collaboration with INESC-ID via project reference UID/CEC/50021/2013.

**Declaration of conflicting interests:**

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

# Automatic cyberbullying detection: A systematic review.

Abstract

Automatic cyberbullying detection is a task of growing interest, particularly in the Natural Language Processing and Machine Learning communities. Not only is it challenging, but it is also a relevant need given how social networks have become a vital part of individuals' lives and how dire the consequences of cyberbullying can be, especially among adolescents. In this work, we conduct an in-depth analysis of 22 studies on automatic cyberbullying detection, complemented by an experiment to validate current practices through the analysis of two datasets. Results indicated that cyberbullying is often misrepresented in the literature, leading to inaccurate systems that would have little real-world application. Criteria concerning cyberbullying definitions and other methodological concerns seem to be often dismissed. Additionally, there is no uniformity regarding the methodology to evaluate said systems and the natural imbalance of datasets remains an issue. This paper aims to direct future research on the subject towards a viewpoint that is more coherent with the definition and representation of the phenomenon, so that future systems can have a practical and impactful application. Recommendations on future works are also made.

*Keywords:* cyberbullying; automatic cyberbullying detection; natural language processing; machine learning; abusive language; social networks

Automatic cyberbullying detection: A systematic review.

## 1. Defining and detecting cyberbullying online

With the development of new Information and Communication Technologies (ICT) and the vast proliferation of Social Network Systems among adolescents to communicate online, interpersonal relationships have gained a new medium through which communication is established. In these online intercommunications, it is common to see social interactions involving offensive online content, since this is one of the main expressions of aggression in cyber harassment situations, such as cyberbullying [Authors, 2017]. Cyberbullying has been defined in various ways but there is a large consensus in the literature that it involves intentional, cruel and repeated behavior among peers, by means of electronic media (Olweus, 2012; Wright, 2017). The content used by adolescents in online interactions has been examined in the literature with the intent of understanding what determines how they communicate with data gathered from self-reported data [e.g., Authors, 2018c]. Automatic cyberbullying detection can be used within digital tools to prevent and intervene in cyberbullying along with messages to promote self-reflection of behavior (Van Royen, Poels, Vandebosch, & Adam, 2017). Thus, classifiers need to be as accurate as possible to reduce the incidence of cyberbullying. It is still difficult however, to capture these incidents online almost in realtime, depending on the operationalization that is given to cyberbullying by automatic cyberbullying detection systems, among other factors. In light of this concern, we propose to understand whether cyberbullying has been automatically detected according to the criteria that constitute its definition and characteristics in order to improve current classifiers, and develop more effective cyberbullying digital interventions. To reach this objective, this study presents an in-depth analysis of research that has focused on automatic cyberbullying detection through a quantitative systematic review approach. In addition, we complement this approach with an extensive experiment to assess current

practices, namely, with the use of feature engineering. Moreover, this study provides guidelines for future research and suggests improvements to current datasets and classifiers in automatic cyberbullying detection in line with the findings from the systematic review presented.

Research has shown that cyberbullying may affect and be determined by social relationships, such as the sense of belonging to a social group (Glover, Gough, Johnson, & Cartwright, 2000; Spears, Slee, Owens, & Johnson, 2009), since the latter cannot be developed without the interference of the social technological world (Spears et al., 2009). Different forms of cyber harassment such as cyberbullying are increasing in online social interactions, particularly among youth, and may affect adolescents' mental health and well-being (Fridh, Lindström, & Rosvall, 2015; Nixon, 2014). Therefore, it is crucial to detect its occurrence almost in realtime so that effective intervention to resolve cyberbullying incidents is developed. In addition, effective intervention could include the phenomenon's detection online firstly and foremost so that victims may be aided in a timely fashion.

Cyberbullying can be considered a form of bullying in a new context (Li, 2006). Bullying refers to repeated abusive behavior in which there is an imbalance of power among peers with the intent of harming others in a prolonged manner (Olweus, 1993). Some of the main characteristics of bullying include the intention to physically, psychologically or socially harm the victim, the repetition of aggressive behavior over time, and the physical, mental and/or social imbalance of power between the bully and the victim (Olweus, 1993; Smith & Brain, 2000). Due to the negative psychological and physical consequences that may affect adolescents (Anderson & Sturm, 2007), and which have been reported in the literature as being suicide ideation, depression, anxiety, cutting, negative emotions, and psychosomatic symptoms (Fridh et al., 2015; Miller, 2017; Nixon, 2014), it is fundamental that its occurrence be tracked in order to prevent this type of harm on the victims. Moreover,

the consequences of cyberbullying may affect victims in a continuous manner, and therefore, research must focus on the severity of the incident within the specific context in which it occurs and the circumstances surrounding it (Hinduja & Patchin, 2009).

Since this paper aims to provide an in-depth analysis of research that has focused on automatic cyberbullying detection, it is crucial to identify how it has been operationalized in various studies focusing on this type of detection, according to the definitions provided in the cyberbullying literature. For instance, several studies in the field of automatic cyberbullying detection (e.g., Bayzick, Kontostathis, & Edwards, 2011; Reynolds, Kontostathis, & Edwards, 2011) have used Patchin and Hinduja's (2006, p.152) definition of cyberbullying which consists of "willful and repeated harm inflicted through the medium of electronic text". In addition, Nahar, Li, and Pang (2014) and Van Hee and colleagues (2015), used Smith and colleagues' (2008, p.376) definition, where they consider 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". Al-Garadi, Varathan, and Ravana (2016) used Salmivalli's definition of cyberbullying (2010, p.112), mentioning that it could "be a sub-type of aggressive behavior in which an individual or group of individuals repeatedly attack, humiliate and/or exclude a relatively powerless person". Most of these definitions include the use of ICT, and the aggressors' engagement in harmful behavior. Some also refer to the intentionality and repetition characteristics of cyberbullying, and to the fact that the behavior occurs among peers. Lastly, some authors provided their own definitions of cyberbullying with a focus also on intentionality, repetition and harmful behavior (e.g., Dadvar, de Jong, Ordelman, & Trieschnigg, 2012; Dadvar, Trieschnigg, Ordelman, & de Jong, 2013; Huang, Singh, & Atrey, 2014), while others highlighted harassment through malicious comments (Chavan &

Shylaja, 2015), and the bully's persistence, power imbalance and visibility in the community (Dinakar, Reichart, & Lieberman, 2011).

In this study, we consider cyberbullying as an individual or group's repeated intentional aggressive behavior towards other peer(s) with the intent of harming them by sending offensive content or engaging in other forms of social aggression through the use of digital technologies (Belsey, 2006; Hindua & Patchin, 2009; Willard, 2005). Accordingly, characteristics such as repetition, intentionality to harm others, the occurrence among peers, and hostile language will be considered as key factors which the current literature in the field often misrepresents, and which could help identify the phenomenon during online interactions among adolescents.

In line with the aims of this study, we propose to answer the following research questions:

Has cyberbullying been automatically detected in previous research according to the criteria that constitute its definition and characteristics?

Which guidelines for future research in automatic cyberbullying detection can be brought forward from the findings of the systematic review presented?

To answer the research questions of this study, we will perform a quantitative systematic review of the related work in the area of automatic cyberbullying detection focusing on the following key aspects: a) criteria used to classify cyberbullying; b) methodological concerns which provide validity to the results (e.g., inter-rater reliability), and ensure privacy of data (i.e., users' consent); c) performance metrics evaluation used to develop classifiers of automatic detection. With these research questions, this study contributes towards a better representation of the phenomenon under study, and, consequently, to the development of more accurate classifiers of automatic cyberbullying detection, which can be used to prevent and intervene with adolescents in digital contexts.

## 2. Issues regarding automatic cyberbullying detection: a systematic review approach

One of the main issues in the process of automatic cyberbullying detection refers to the correct operationalization of cyberbullying, considering the main criteria provided by the literature in the field, in order to accomplish the goal of automatic detection systems, which is to accurately identify cyberbullying events. However, capturing the complexity of the phenomenon requires well-defined criteria to develop appropriate digital tools which integrate automatic detection features.

To provide a better understanding of the aspects which have been considered in automatic detection of cyberbullying, this study presents a systematic review of existing research with a focus on four main criteria, according to the aforementioned definitions of cyberbullying, namely, a) the use of aggressive or hostile language, b) intentionality to harm other(s), c) repetition of behavior, and d) incidence among peers. Furthermore, this study also focuses on methodological issues which provide validity to the classifiers used to build models of detection, such as, inter-rater reliability (i.e., whether the rate of agreement between coders was considered or not), the coders' expertise (i.e., whether coders were experts in the field of cyberbullying or not), peer interaction (i.e., whether users were peers in the universe from which data was extracted or not), and users' consent (i.e., whether users were given knowledge regarding the extraction of data).

Specifically concerning machine learning (ML) approaches, models based on methods from complexity science may be used in a wide range of social dynamics to prevent for instance, disasters, epidemic diseases, war, terrorism and crime (Helbing et al., 2014). Although ML and Natural Language Processing (NLP) techniques have been successful in a variety of text-based tasks (e.g., sentiment analysis, topic detection, machine translation, text summarization), their application to cyberbullying detection has encountered many challenges, maintaining it an unsolved issue. Cyberbullying as a classification task is fairly

"recent". Reynolds and colleagues (2011) for instance, reported how through the development of a simple language-specific method, they recorded the percentage of curse and insult words in a post, achieving a recall = 0.785 in cyberbullying identification on a small Formspring dataset. Bayzick and colleagues (2011) developed a program (i.e., BullyTracer) where they identified a "cyberbullying window" 85.3% of the time (recall) and an "innocent window" 51.9% of the time in MySpace posts. More recently, the most common approach to cyberbullying detection has been through feature engineering, which has expanded the common bag-of-words representation of text by creating additional features/dimensions that use domain knowledge of linguistic cues in cyberbullying to attempt to improve a given classical classifier's performance (e.g., Support Vector Machines - SVM, Logistic Regression). Frequent features relate to the use of profanity and how often it occurs in text (Al-Garadi et al., 2016; Dadvar et al., 2012; Dadvar et al., 2013; Zhao, Zhou, & Mao, 2016). Characteristics such as user age, gender and their network of friends/followers has also been taken into account (Singh, Huang, & Atrey, 2016). Some research has also used sentiment analysis as an added feature (Dinakar, Jones, Havasi, Lieberman, & Picard, 2012; Sugandhi, Pande, Agrawal, & Bhagat, 2016; Van Hee et al., 2015). Recently, other techniques such as deep learning networks have also started to be applied to this task (Authors, 2018b; Zhang et al., 2017), as well as Fuzzy Fingerprints [Authors, 2018a] which creates a k-sized ordered "fingerprint" containing frequent/relevant keywords for the cyberbullying phenomenon.

In the following sub-sections, we provide detail on which and why specific studies concerning automatic cyberbullying detection were selected. Firstly, we analyze the presence of criteria used to identify cyberbullying, and the methodological demands which were contemplated in these studies (sub-section 2.1). Subsequently, we detail which datasets are commonly used and their characteristics (sub-section 2.2), we present the best

performing result for each article (sub-section 2.3), and, finally, we take a closer look at which type of features were handcrafted to achieve optimal performance (sub-section 2.4).

### 2.1. Identifying cyberbullying definition criteria and other methodological demands of previous research

We mentioned that cyberbullying would be the main focus of the research and therefore, it would be the object of automatic textual detection to determine which and why specific studies were selected for analysis in the current study. Automatic cyberbullying detection oriented research is often deconstructed into related, but ultimately different constructs, namely, aggression detection, cyberbullying role classification, aggression intensity determination, racism detection, etc. Therefore, the chosen articles have been selected firstly on the criteria that the authors claimed to be performing cyberbullying detection. Moreover, the studies for analysis were also chosen on the basis that the authors used robust techniques to conduct their research, such as, cross-validation, text pre-processing (data cleaning), state-of-the-art classifiers and theoretically grounded feature engineering. The chosen studies were thus obtained by querying the several online tools available for researchers (e.g., Google Scholar, Research Gate, ACM Digital Library, Arxiv, Scopus, Mendeley). Of 71 studies originally found by these search engines, two postgraduate data scientists with expertise in machine learning deemed that 22 studies fit the criteria and therefore were viable for the systematic review we have proposed.

Most of the selected studies did not provide sufficient information regarding how datasets were built, especially considering which cyberbullying criteria were followed, along with other methodological concerns (see Table 1). With the exception of five studies (Bayzick et al., 2011; Hosseinmardi, Rafiq, Han, Lv, & Mishra, 2016; Ptaszynski et al., 2016; Sugandhi et al., 2016; Van Hee et al., 2015), the remaining studies did not provide details of the instructions given to annotators to label the data samples provided. Inter-rater

reliability was also rarely a reported metric, as well as the annotators' expertise. Referring to this last aspect, we also observed that those responsible for the annotation process varied from random people at the Amazon Mechanical Turk (2018) platform (i.e., Authors, 2018a; Authors, 2018b; Reynolds et al., 2011; Zhang et al., 2017) to students (Bayzick et al., 2011; Dadvar et al., 2012; Sugandhi et al., 2016). Thus, the requirement of experts in the field of cyberbullying as annotators was not reported in the majority of the studies, or otherwise was not contemplated, which may compromise the validity of the annotation process. Another key aspect is that when information about data extraction was reported, data was mostly extracted via web crawling [i.e., either directly from a website, or with usage of a public Application Programming Interface (API) provided by a social network], which means that the users extracted for the dataset were unaware of this specific use of their data, despite this being a common and legal practice. Additionally, data extraction was done randomly, which could not assure that the users from which cyberbullying was labelled were peers, with the exception of Ptaszynki and colleagues' (2016) study which extracted the dataset from school forums and discussion groups. Finally, out of the four key criteria to define cyberbullying (i.e., aggressive language; repetitiveness; intentionality; and behavior amongst peers), none of the chosen studies mentioned that all these criteria were met during the annotation process (Table 1), either because they lacked this information in the published studies, or by only considering some of the criteria or even none. More specifically, the most common single instruction given to annotators regards the use of aggressive language, and the criterion of occurrence between peers was missing in all of the studies. Therefore, it is unlikely that the same construct has been measured and integrated in the classification process. Moreover, it seems that only isolated aspects of cyberbullying have been captured by classifiers, which may not allow systems to accurately detect cyberbullying events, as previously mentioned.

### 2.2. Commonly used datasets

There are no standard datasets used for cyberbullying detection (Table 2). Although most studies recur to the same social networks in order to obtain data (e.g., Twitter, YouTube), the datasets are independently created by using a publicly available API or scrapping the website for samples. Therefore, the data cannot be compared. One frequently used dataset is Formspring, however it has been subject to updates throughout the years. When Formspring was first created, it had nearly 4000 samples (Reynolds et al., 2011), but it has since tripled in size [Authors, 2018a; Authors, 2018b]. The only repeating datasets amongst different authors have been from Kongregate, Slashdot and MySpace, (Fundación Barcelona Media, 2009) available in different iterations throughout the literature.

Table 2 shows that the original datasets are considerably unbalanced, with most articles working with datasets where less than 20% of the available samples have been categorized as cyberbullying. This imbalance is a challenge, since it has been widely documented to affect the predictive capabilities of machine learning classifiers (Chawla, 2009; Chawla, Japkowicz, & Drive, 2004). Some studies (Al-Garadi et al., 2016; Huang et al., 2014; Singh et al., 2016; Zhang et al., 2017) have used synthetic oversampling or undersampling techniques in order to achieve a more balanced dataset, which is reported to result in a better classifying performance due to the fact that cyberbullying is a naturally unbalanced phenomenon in terms of occurrence. Specifically, non-normal distributions are likely in cyberbullying research (Bauman et al., 2013). This imbalance is reflected in the data by a scarce number of cyberbullying samples in social networks, in contrast with everything else that people commonly post.

Another important aspect worth mentioning is that the vast majority of the datasets presented in Table 2 are labeled for cyberbullying with a single message/post available. As mentioned before, cyberbullying is by definition a repetitive act, therefore, it is unlikely to

assert the presence of a cyberbullying event from a single text message. Thus, we believe that in order to properly categorize cyberbullying, it is required that a history of repeatedly aggressive posts towards someone is identified. Despite being a serious issue, isolated cases of aggression cannot be considered as cyberbullying due to the repetitive nature of this phenomenon (Patchin & Hinduja, 2006; Smith et al., 2008). While a few studies acknowledge this difference (Authors, 2018a; Chavan & Shylaja, 2015; Hosseinmardi et al., 2016; Mangaonkar, Hayrapetian, & Raje, 2015; Nahar, Li, Pang, & Zhang, 2013; Van Hee et al., 2015), they keep the task labeled as cyberbullying detection. To the best of our knowledge, only Nahar and colleagues (2013) attempted to capture the repetitiveness of aggression by detecting cyberbullying in sessions consisting of streams with several messages. Chatzakou and colleagues (2017) also used a similar approach, by grouping batches of messages based on their timestamp, but in this instance, the task was cyberbullying role detection (i.e., bully, victim, bystander). Also, Zhao and colleagues (2016) labeled samples for what the authors described as “bullying traces”, which are defined as response to the bullying experience (i.e., it includes bullying samples but also texts of users talking about bullying in a broad sense).

To sum up, we put forth that the currently available datasets undermine the overall research done in this area and, thus, it is urgent to perform a paradigm shift on how the data reflects cyberbullying, so that more thorough research can take place in the following years.

### 2.3. Performance metrics

Table 3 shows the results of the experiments conducted by the 22 studies this article reviews. In the cases where the experiments were performed on more than one dataset or with more than one classifier/approach, we reported on the best performing classifier-dataset pair. Namely, in Nahar and colleagues’ (2014) study, we reported on the results from the Kongregate dataset and in the study of Zhao and Mao (2016) we reported on the results from

the Twitter dataset. In contrast, a combination of multiple datasets was used in other projects where more than one dataset is mentioned (i.e., Dinakar et al., 2012; Nahar et al., 2013; Sugandhi et al., 2016; Zhao & Mao, 2016).

The main aspect to consider in Table 3 and its sparseness of values, is the lack of consensus about how to evaluate cyberbullying detection systems. While the majority of studies reported on precision, recall and f1-score, in some cases, researchers reported on a classifier's binary performance (i.e., cyberbullying class "CB" and non-cyberbullying class "nCB"), such as the studies of Al-Garadi and colleagues (2016), Chavan and Shylaja (2015), Dadvar and colleagues (2013), Mangaonkar and colleagues (2015), Singh and colleagues (2016), Sugandhi and colleagues (2016), Zhao and colleagues (2016), Zhao and Mao (2016), and Zhang and colleagues (2017). Other studies reported on the cyberbullying class (CB) performance alone (e.g., Authors, 2015; Authors, 2018b; Dadvar et al., 2012; Huang et al., 2014; Nahar et al., 2013; Nahar et al., 2014; Ptaszynski et al., 2016; Reynolds et al., 2011; Singh et al., 2016; Van Hee et al., 2015). As a consequence, either the reported metrics were weighted macro-averages that proportionally impacted the results, or the supervised approaches were more proficient in detecting non-cyberbullying, because of the overwhelming majority of non-cyberbullying training samples.

In addition, other studies argued that metrics such as Area Under the Curve (AUC) and Receiver Operating Characteristic curve (ROC) are better predictors of a classifier's performance (e.g., Al-Garadi et al., 2016; Chavan & Shylaja, 2015; Huang et al., 2014), while a few chose to highlight recall as a more relevant metric (Authors, 2018a; Reynolds et al., 2011). Previous research has shown that in the specific case of cyberbullying detection, there are few advantages in reporting values that are macro-averaged from two classes [Authors, 2018a]. Not only is the nCB class (i.e., every other post in the world) impossible to fully represent in a training set, but it is also impractical to have a system that is very

good at predicting nCB events at the expense of true CB samples. Furthermore, while we argue that recall is theoretically more important, because we want to make sure no False Negative cyberbullying events are missed by the system, we cannot follow this approach at the expense of precision (i.e., too many False Positives that make the system inaccurate). An imprecise system prevents the possibility of working in almost realtime due to the need to check and filter the False Positives manually.

In terms of the reported results systems that focus on cyberbullying as a binary task often report better performance, because those systems will typically be better at detecting the nCB class [Authors, 2018a; Authors, 2018b]. As previously mentioned, this happens mostly due to the imbalance of the dataset [Authors, 2018a]. With the exception of Nahar and colleagues (2014) and Hosseinmardi and colleagues (2016) (this study mixes image and text since it is based on Instagram), all of the studies that report on the CB class alone share the common conclusion that their system is not yet ready for real world application, due to low values of performance, generally below f1-score = 0.650. From this analysis, we argue that automatic cyberbullying detection systems should only report the CB category results. . This is something that has been already argued in recent research [Authors, 2018a].

As a final note, in Table 3, the entry of Dinakar and colleagues' (2011, 2012) work is empty because their classifiers were not for "pure" cyberbullying prediction. Instead, they trained and tested their model on its ability to distinguish amongst three categories of potential cyberbullying: one's sexuality, race and intelligence. Nonetheless, we consider their work an advancement on cyberbullying detection and hence chose to mention it.

### 2.4. Types of features

As mentioned earlier, one of the most common approaches to improve cyberbullying detection is to perform feature engineering. From the various studies presented thus far, we divided the type of features into five categories:

a) Textual Features are features that relate to statistical input text dependent features. This includes things such as n-grams, skip grams, the length of the text, count/ratio of "emoticons", count/ratio of profanity, number of pronouns, parts-of-speech tagging, etc.

b) Social Features are features that extract information from the network of messages and/or friendships of the users involved within a given text input, namely the number of friends, the number of followers, the number of liked posts and several centrality measures that could be extracted from a graph representation of such networks (e.g., betweenness, eigenvector, Katz). Despite the fact that a few studies have included these extra dimensions, in most cases, the available datasets do not enclose this type of information.

c) User Features are features that relate to information regarding the posting user (e.g., age, gender).

d) Sentiment Features provide information regarding the sentiment of the input text or the individual words and/or expressions it contains. This is usually done through a well-established classifier purposefully trained for the task of sentiment analysis, or the use of a dictionary of words with sentiment related information about those words (e.g., valence, arousal).

e) Word Embeddings are a N-sized distributed representation for words, that were trained in an unsupervised fashion, therefore capturing their semantic "value" (Bengio, Ducharme, Vincent, & Jauvin, 2003; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). When used as extra features that extend the bag-of-words vector, the average of the vectors for all the document words is commonly used.

Table 4 shows what type of features each study uses in order to improve a standard classifier's performance. Please note that by stating that a given work used textual features, it does not mean it used all of the features mentioned in the above description. It simply pertains to a subset of those added dimensions that can vary from work to work.

Half of the studies selected for this work use textual features, making it the predominant category of features most commonly used to improve performance. This likely happens because it includes a very specific kind of features that most articles acknowledge to be important to the phenomenon (e.g., profanity). It is however important to remember that the plain existence of curse words is not enough to detect cyberbullying. As previously mentioned, two friends may message each other using derogatory terms, but no offense is taken because that is the nature of their friendship. This is one of the many challenges that is rarely addressed in the current literature, with the exception of Salawu, He, and Lumsden (2017).

In contrast, very few studies used either social features or user features. This is explained by the fact that many social networks protect user information (e.g., age, gender) from public extraction methods, to protect their users' data from being abused for marketing purposes. Additionally, building a network representation of a set of users' friends and followers is laborious, and will often include extra data extraction, something that the majority of datasets opted to not have.

Only three research teams proposed the use of sentiment analysis as a tool in cyberbullying detection (Dinakar et al., 2012; Sugandhi et al., 2016; Van Hee et al., 2015). Given the negative emotion that an act of cyberbullying may entail, it would be expected that more studies would have considered this option. Nonetheless, because sentiment analysis is in itself a complicated classification task, continuously subject to improvements

and new challenges, it is also reasonable to consider that few researches would choose to include something that can induce a classifier in error.

Finally, word embeddings as a dense representation of the input are a recent trend in NLP tasks and also in cyberbullying detection (Authors, 2018b; Zhao & Mao, 2016). In the study of Zhao and Mao (2016), embeddings were used to build what the authors' called "bullying features" based on examples of insults. Authors (2018b) used three different types of word embeddings that were tested as input (i.e., Google, Twitter and Formspring pre-trained) and, when coupled with different deep learning architectures, achieved better results than a baseline TF-IDF inputted SVM. In contrast, another study (Zhang et al., 2017) discovered that a phoneme-based representation for a convolutional neural network (CNN) outperformed a word embeddings representation.

## 3. Experimental setup

### 3.1. Material and method

In this section, we present the experimental setup in order to assess the current practices detailed previously, more specifically by focusing on the use of feature engineering as an attempt to improve cyberbullying automatic detection. We designed an experiment using some of the most commonly reported classifiers in Table 3 (i.e., SVM, Logistic Regression and Random Forests) and the most common features reported in Table 4 in order to further analyze its strengths and weaknesses (see sub-section 3.1.2). Furthermore, we also intended to test the use of psycholinguistic features, considering that this approach was explored to a lesser extent in the domain of feature engineering.

#### 3.1.1. Datasets

As mentioned in sub-section 2.2, there are no benchmark datasets. Based on the availability, we chose two datasets to perform our experiment on: i) the latest version

available of the Formspring dataset (i.e., Formspring4 from Table 2 available at https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection); and ii) the latest Bullying Traces V3.0 dataset (http://research.cs.wisc.edu/bullying/data.html) as proposed by Zhao and colleagues (2016). These datasets only define cyberbullying relying on a single message/post which, as we have previously shown in sub-section 2.1, is insufficient to capture the key aspects of this phenomenon.

The Formspring dataset consists of 13160 texts labeled through Amazon's Mechanical Turk (2018) by three annotators. Out of this total, 2205 texts were deemed by at least one annotator to contain cyberbullying, while 10955 show no evidence of the aforementioned phenomenon. The vocabulary consists of 17846 different words, averaging 6.56 characters *per* word. While 6.56 characters seems to be a high value, this can be explained by the presence of long nonsensical tokens (e.g., "spamspamspamspamspamspamspamspam").

The Bullying Traces dataset contains 2999 tweets, extracted from Twitter in August 2011, and manually labeled for the presence of "bullying traces" (as previously explained in sub-section 2.1). It contained 1246 tweets with said traces and 1753 tweets without. The vocabulary consists of 8920 different words, averaging 7.10 characters *per* word. This high value may be explained similarly to the Formspring dataset.

#### 3.1.2. Features

Based on the various types of dimensions added by all of the 22 works explored in this literature review, we handcrafted several features to concatenate the typical TF-IDF representation of documents. These are described below:

a) 10 Textual Features: ratio and count of nouns, verbs, adjectives, pronouns and adverbs in text; ratio and count of "bad words", i.e., swear words and/or adult

language extracted from a public list submitted online by users (available dataset at https://www.noswearing.com/dictionary).

b) 21 Sentiment Features: the polarity (sentiment) and subjectivity of each text, via TextBlob (https://textblob.readthedocs.io/en/dev/) ; the positive, negative, neutral and overall sentiment score of each text, via NLTK's (https://www.nltk.org/) VADER (Hutto & Gilbert, 2014); the sum, average, minimum, maximum and difference (i.e., maximum minus minimum) values of words/expressions in the input text based from their valence (i.e., the pleasantness of the stimulus), arousal (i.e., the intensity of emotion provoked by the stimulus) and dominance (i.e., the degree of control exerted by the stimulus), as created by Warriner, Kuperman, and Brysbaert (2013).

c) Word Embeddings: a word vector representing the document, with the average of the word embeddings model. The model was trained for either 100, 300 or 500 dimensions, in unsupervised fashion from the totality of the vocabulary of each dataset, via the genism (https://radimrehurek.com/gensim/models/word2vec.html) default implementation of word2vec.

Due to the fact that social and user features were not available in these datasets, for the reasons explained in sub-section 2.3, we attempted to build novel features and were inspired by Al-Garadi and colleagues' study (2016). In that study, the authors used a dictionary of words used by neurotic users on Facebook to improve cyberbullying detection, under the logic that neuroticism is a common trait in aggressors. As a consequence, we propose to test several novel psycholinguistic-related features, namely:

a) 15 Personality Trait Features: based on the "Big Five" Personality traits, which classifies personality into five dimensions: extroversion (e.g., outgoing, talkative,

active), agreeableness (e.g., trusting, kind, generous), conscientiousness (e.g., self-controlled, responsible, thorough), neuroticism (e.g., anxious, depressive, touchy), and openness (e.g., intellectual, artistic, insightful). Schwartz and colleagues (2013) created a list of words/expressions commonly used by Facebook users for each one of the five traits, via the World Well-Being Project (http://www.wwbp.org) We calculated the count, ratio and presence of words for each of the 5 traits, in each input document.

b) 210 MRC Psycholinguistic Features: based on the MRC Psycholinguistic database (http://websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html), which is a dictionary containing 150837 words and provides information about 26 linguistic properties of different subsets of those words. These properties range from the number of letters and phonemes in the word, to the age of acquisition and its status (e.g., colloquial, alien, archaic, nonsense, rhetorical, specialized). For each of these 26 linguistic properties and their sub-properties, the sum, average, minimum, maximum, and difference (maximum minus minimum) of those words are calculated as belonging to the input text.

### 3.1.3. Testing scenarios

In order to better determine the impact of each feature, we designed 12 experiment scenarios with different combinations of features used as input representation for each dataset. These scenarios are detailed below:

- scenario A: TF-IDF;
- scenario B: TF-IDF + Textual Features;
- scenario C: TF-IDF + Sentiment Features;
- scenario D: TF-IDF + Word Embeddings;
- scenario E: TF-IDF + Personality Trait Features;

- scenario F: TF-IDF + Textual Features + Word Embeddings;
- scenario G: TF-IDF + Textual Features + Personality Trait Features;
- scenario H: TF-IDF + Personality Trait Features + Word Embeddings;
- scenario I: TF-IDF + MRC Psycholinguistic Features;
- scenario J: TF-IDF + Sentiment Features + Personality Trait Features + MRC Psycholinguistic Features;
- Scenario K: all features described in sub-section 3.2 and without TF-IDF;
- Scenario L: TF-IDF + all features described in sub-section 3.2.

For each of the above scenarios and the classifiers used, we performed some basic hyper-parameter optimization during training. Specifically, we performed a 5-fold grid parameter search to determine if stop words should be used, as well as unigrams, unigrams plus bigrams or just bigrams. In terms of the Word Embeddings, they were pre-trained to have either 100, 300 or 500 dimensions and these three possible N-sized vectors were also a part of the grid search.

### 3.2. Experiment results and discussion

In this section, we present and discuss the results from the experiment detailed above. Since we have argued that the correct approach to cyberbullying is to report solely on the CB class results, regardless of the imbalance of the dataset, all tables and images in this section follow that principle.

#### 3.2.1. Bullying Traces dataset results

Regarding the Bullying Traces dataset, Figures 1, 2 and 3 illustrate the results of the experiment for the Bullying Traces dataset and focus on f-measure, precision and recall, respectively. Considering the f-measure as the benchmark metric, Figure 1 shows that adding handcrafted features will rarely results in an improved performance. The best

performing algorithm is SVM with scenario A (pure TF-IDF) with an f1-score = 0.740 which is only matched for scenarios D and H (f1-score = 0.740). The common element in these scenarios is that they both use Word Embeddings. Nonetheless, it is important to note that the addition of these different features never outperforms the TF-IDF approach, which indicates that the effort to perform feature engineering is not rewarding. In some cases, such as the use of MRC Psycholinguistic Features (scenarios I and J), they induce an approximate 32% drop in performance for SVM and Logistic Regression when compared do scenario A. We theorize that so many (i.e., 210) features with different rationales end up creating "noise" for the classifiers, as opposed to adding distinctive information. In addition, the MRC Psycholinguistic features were constructed from several merged databases; all of which were built in the second half of the 20th century, and therefore, do not contain much information on words common to 21th century social networks. Overall, SVM and Logistic Regression will always perform comparably, despite a slight advantage from the former. The Random Forests algorithm always performs worse than the aforementioned methods, which we theorize is due to the small size of the feature subset used by the decision trees in this ensemble method. The default parameter value, which was not optimized, is the square root of the vocabulary size: in this case, it is 94 features.

Figures 2 and 3 also show the analogous results for recall and precision, as opposed to the f1-score, since we have argued that recall is an important metric in cyberbullying detection, as long as precision remains within acceptable values. In this case, recall and precision are balanced. However, overall, the same main conclusions can be drawn for the key indicator, recall: a) there is no improvement in using handcrafted features; b) SVM is, for the most part, the best performing algorithm; c) the MRC Psycholinguistic Features induce a drop in performance.

The results presented (i.e., f1-score = 0.740) are in line with Zhao and colleagues' (2016) original findings (f1-score = 0.780), as shown in Table 3. The difference in our findings can be explained by the use of slightly different features, and the fact that the previous research reports on the overall binary performance, as opposed to the CB class alone. In addition, our version of the dataset had approximately twice as many bullying instances, which provides a greater balance to the dataset (i.e., 41.50% of the samples are cyberbullying related) and assures better results. However, as mentioned in section 3, this dataset is labeled for bullying traces, which includes more than just the bullying events themselves. This seems a too broad an interpretation of cyberbullying and dilutes the main goal of the task, which is detecting cyberbullying specifically.

### 3.2.2. Formspring dataset results

Regarding the Formspring dataset, Figures 4 and 5 illustrate the results of the experiment for the Formspring dataset and focus on f-measure, precision and recall, respectively. Considering the f-measure as the benchmark metric, Figure 4 shows that there is one trend common to the Bullying Traces experiment: SVM was the best performing method and Random Forest was the worst. In contrast, the inclusion of some features provided a marginal improvement over the base scenario A (f1-score = 0.410). Scenarios B, F and G, which share the characteristic of using Textual Features, all marked a 2.4% improvement in absolute f1-score (0.420). Scenario D which uses TF-IDF and Word Embeddings, also achieved the same performance. The best performing scenario was H (f1-score = 0.450), which has a 7.1% improvement over base scenario A and added Personality Traits and Word Embeddings to the TF-IDF representation of text.

We believe that the reason behind this improvement in scenario H was two-fold. On the one hand, the Word Embeddings were pre-trained using all samples of the dataset they

were tested on, which made the document representation narrowly defined within the scope of this particular dataset (i.e., Google's Word Embeddings were likely to be a poor choice given the formal nature of the documents used to train them). On the other hand, the Formspring dataset was more purely devoted to cyberbullying detection (or, at least, cyber aggression), Personality Features were likely to have a bigger impact, as previous studies have shown that neurotics, for instance, are more likely to be aggressors (Al-Garadi et al., 2016).

In previous research, different approaches on the Formspring dataset have been tested, such as, deep learning architectures [Authors, 2018b] and Fuzzy Fingerprints [Authors, 2018a]. In those instances, the best reported results were f1-score = 0.444 on a hybrid C-LSTM approach using pre-trained Twitter Embeddings as word representation and f1-score = 0.425, respectively. Therefore, we conclude that the feature engineering approach to cyberbullying detection is, at best, comparable to other techniques available.

Figures 5 and 6 also show that these classifiers tend to have better precision than recall, a situation where Logistic Regression and Random Forests were able to outperform SVM. As far as the key indicator, recall, Figure 6 shows that the added features as described in sub-section 3.2 provided more impact, with scenarios C, D, F, H, I, J and L, all showing improvements when using SVM as a predictor. Although, the best performance was achieved by the Random Forests (recall = 0.460), when using all features to TF-IDF vector (scenario L). Nonetheless, in previous work with the same dataset and Fuzzy Fingerprints, a recall of 0.597 was achieved [Authors, 2018a].

## 4. Final Considerations

This study aimed to understand whether cyberbullying has been automatically detected according to the criteria that constitute its definition and characteristics. Thus, we

presented an in-depth analysis of research that focused on automatic cyberbullying detection through a quantitative systematic review approach. We also complemented this approach with an extensive experiment to assess current practices, by using feature engineering. In this section we provide guidelines for future research and suggest improvements to current datasets and classifiers in automatic cyberbullying detection in line with the findings from the presented systematic review.

In this work, we conducted an in-depth analysis of 22 studies on cyberbullying automatic detection systems. In these studies, even though different definitions have been adopted, there are common and shared aspects, as described in the beginning of this work, which refer to the repetitive use of aggressive language amongst peers with the intention to harm others through electronic media (Patchin & Hinduja, 2006; Salmivalli, 2010; Smith et al., 2008). However, we found that the key aspects of cyberbullying were not fully represented in these studies which therefore, may lead to a mischaracterization of the phenomenon. As a consequence, the most representative studies on automatic cyberbullying detection, published from 2011 onwards, have conducted isolated online aggression classification, as opposed to cyberbullying classification. Therefore, in order to develop accurate classifiers of cyberbullying, there is a need for future research to take into account its operationalization, for instance, by providing instructions to annotators on objective criteria regarding the key features of cyberbullying. This could contribute to a better representation of this phenomenon and its complexity, and subsequently, lead to improved classifiers for automatic cyberbullying detection. Furthermore, methodological concerns should also be contemplated in order to provide greater validity to the classification process, such as the level of agreement between annotators. Additionally, we have also observed a lack of report considering the existing privacy policies, namely the need to request users' permission to use their social networks information.

In a more micro view of the related work, we found that there is a lack of quality datasets, both from how they are built (i.e., single messages as opposed to a history of posts) to how they are annotated (i.e., no information regarding cyberbullying criteria). The reported results are also not benchmarkable, due to a lack of consensus on both the metrics and the acceptance of cyberbullying as a single class problem. We believe that, while recall is an important key metric in any cyberbullying detection system, the f1-score remains the most balanced way to evaluate said systems, especially because we can parameterize the f1-score to be a f2 or f3-score, that attributes more weight to recall as a part of the f-measure. To validate these findings, we designed an experiment using two publicly available datasets used in previous studies and found that the current practice of performing feature engineering to improve classification performance is, at best, marginally better. While feature engineering provides competitive performance in currently available "cyberbullying detection" datasets, it does so through lengthy manipulation and pre-processing of the data. In addition, whilst reporting exclusively on a predictors' ability to detect cyberbullying, the results are insufficient for real-world applications. Systems with f1-score $< 0.80$ and that do not follow the key principle we have argued about reporting results solely on the CB class, may be incapable of completing their task.

Considering our first concern of understanding if cyberbullying has been automatically detected according to the criteria that constitute its definition and characteristics, we conclude that past research was unable to appropriately integrate the main aspects included in the definition of cyberbullying, thus, current detection systems seem to be misrepresenting this phenomenon. Considering this, we claim that automatic cyberbullying detection remains an unsolved task which was confirmed by the poor performance found during our experiment (f1-score = 0.450 and recall = 0.460 in the Formspring experiment).

To sum up, and according to our second objective, we suggest that future research should take into account and fully report a set of necessary information to increase the quality of future datasets and to improve classifiers' performance. It is of utmost importance to provide proper instructions to annotators according to the criteria that represents the definition of cyberbullying (i.e., intentionality, repetition, aggressiveness and behavior amongst peers), and also ensure that the annotators are experts in the field of cyberbullying. Additionally, users' data extraction should be obtained from peers, and closer attention should be given to users' privacy during this process. Also, mechanisms should be developed to attempt to capture the context and nature of the relationship of the participants in a cyberbullying event, as it is a key component to identify intentionality to harm and repetitive aggressions amongst peers.

### 4.1. Limitations and future work

We argue that a shift should occur aiming a more comprehensive and extensive classification of cyberbullying in future research, along with more rigorous and consensual metrics leading to quality datasets, considering the findings of the present study. In order to achieve a higher quality and accuracy of cyberbullying classification, we plan to build a dataset grounded on an annotation process developed by a team of psychologists specialized in the area of cyberbullying. With this future resource output, we hope to contribute with valuable research and tools regarding automatic cyberbullying detection, and also concerning cyberbullying digital interventions with the aim of reducing the incidence of this phenomenon in online contexts.

Some limitations of the present study need to be addressed. We could have considered performing a meta-analysis, however, the studies that were reviewed did not provide the necessary values (e.g., effects) that would enable this type of analysis.

Moreover, we could not perform a deeper analysis with regards to users' characteristics, also due to the fact that the studies which were reviewed did not provide this information. Despite these limitations, we consider that the present work contributes to shed some light on future efforts to improve the existing datasets and classifiers of automatic cyberbullying detection.

### 4.2. Practical implications for intervention in cyberbullying

The findings of this study highlight the need to improve the overall quality and accuracy of cyberbullying detection systems, presenting important implications for prevention and intervention in cyberbullying. Considering that classifiers can be integrated in digital tools to prevent and intervene in the context of cyberbullying, they need to be as precise as possible to be effective. For instance, automatic cyberbullying detection can be used to prevent individuals from receiving harmful online content in social networks, particularly among adolescents, thus, it may help to reduce the incidence of cyberbullying. Digital tools such as applications, games and websites need to be developed and integrated in social networks to prevent cyberbullying, as well as other forms of cyber harassment, considering the reported negative effects of this phenomenon on adolescents' mental health and well-being (Fridh et al., 2015; Miller, 2017; Nixon, 2014). More specifically, automatic monitoring can be used to detect cyberbullying early, thus, preventing harmful behavior from reaching its target. At the same time, it can be complemented by reflective interfaces (e.g., notifications, action delays) to promote users' self-reflection and more pro-social online behaviors, as well as positive online interactions. This type of cyberbullying digital interventions have been recently developed by other researchers (e.g., Dinakar et al., 2012; Van Cleemput, 2015; Van Royen et al., 2017). Thus, it is vital that these digital tools are able to correctly identify cyberbullying, distinguishing it from other online situations such as the use of curse language amongst peers in a playful context. One of the risks of an incorrect

detection of cyberbullying is that it can cause a decrease in the responsiveness to these tools, leading users to not adhere to them. This aspect is particularly relevant considering that adolescents' perceptions about automatic monitoring can be rather negative if they believe their freedom of expression is being jeopardized (Van Royen, Poels, & Vandebosch, 2016). Another risk of incorrectly detecting cyberbullying is designing ineffective digital tools which are unable to prevent and intervene in actual cyberbullying events, thus, failing to protect users from harmful situations. Therefore, the quality of future classifiers of automatic cyberbullying detection need to be improved in order to meet these challenges.

## References

Al-Garadi, M., Varathan, K. D., & 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. doi:10.1016/j.chb.2016.05.051

Amazon Mechanical Turk (2018). Retrieved from https://www.mturk.com

Anderson, T., & Sturm, B. (2007). Cyberbullying: From playground to computer. *Young Adult Library Services, 5*(2), 24-27. Retrieved from https://taraandersongold.files.wordpress.com/2015/11/cyberbullying-from-playground-to-computer.pdf

[Authors 2015]

[Authors 2017]

[Authors 2018a]

[Authors 2018b]

[Authors 2018c]

Bauman, S., Cross, D., & Walker, J. (2013). *Principles of cyberbullying research: Definitions, measures, and methodology*. New York, US: Routledge/Taylor & Francis Group. doi:10.4324/9780203084601

Bayzick, J., Kontostathis, A., & Edwards, L. (2011, June). Detecting the presence of cyberbullying using computer software. In Proceedings of the *3rd Annual ACM Web Science Conference* (WebSci '11). Retrieved from http://www.websci11.org/www.websci11.org/fileadmin/websci/Posters/63_paper.pdf

Belsey, B. (2006). *Cyberbullying: An emerging threat to the "Always On" generation*. Retrieved from http://www.cyberbullying.ca/pdf/Cyberbullying_Article_by_Bill_Belsey.pdf

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research, 3*, 1137-1155. Retrieved from http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017, June). Mean birds: Detecting aggression and bullying on twitter. In Proceedings of the *2017 ACM on web science conference* (pp. 13-22). doi:10.1145/3091478.3091487

Chavan, V. S., & Shylaja, S. S. (2015, August). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In Proceedings of the *International Conference on Advances in computing, communications and informatics* (ICACCI; pp. 2354-2358). doi:10.1109/ICACCI.2015.7275970

Chawla, N. V. (2009) Data mining for imbalanced datasets: An overview. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 875–886). Boston, MA: Springer. doi:10.1007/978-0-387-09823-4_45

Chawla, N. V., Japkowicz, N., & Drive, P. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter, 6*(1), 1-6. doi:10.1145/1007730.1007733

Dadvar, M., de Jong, F., Ordelman, R. J. F., & Trieschnigg, R. B. (2012, February). Improved cyberbullying detection using gender information. In Proceedings of the *12th Dutch-Belgian information retrieval workshop* (DIR 2012; pp.23-25). Retrieved from http://eprints.eemcs.utwente.nl/21608/

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013, March). Improving cyberbullying detection with user context. In Proceedings of the *European Conference on Information Retrieval* (pp. 693-696). Berlin: Springer. doi:10.1007/978-3-642-36973-5_62

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems, 2*(3), 1-30. doi:10.1145/2362394.2362400

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of Textual Cyberbullying. *The Social Mobile Web, 11*(2), 11-17. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/3841/4384

Fundación Barcelona Media. (2009). Retrieved from http://caw2.barcelonamedia.org/

Fridh, M., Lindström, M., & Rosvall, M. (2015). Subjective health complaints in adolescent victims of cyber harassment: moderation through support from parents/friends - a Swedish population-based study. *BMC Public Health, 15*(1). doi:10.1186/s12889-015-2239-7

Glover, D., Gough, G., Johnson, M., & Cartwright, N. (2000). Bullying in 25 secondary schools: incidence, impact and intervention. *Educational Research, 42*(2), 141-156. doi:10.1080/001318800363782

Helbing, D., Brockmann, D., Chadefaux, T., Donnay, K., Blanke, U., Woolley-Meza, O., … Perc, M. (2015). Saving human lives: What complexity science and information systems can contribute. *Journal of Statistical Physics*, *158*(3), 735–781. doi:10.1007/s10955-014-1024-9

Hinduja, S., & Patchin, J. (2009). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Thousand Oaks, CA: Sage Publications. ISBN: 978-141-29-668-9

Hosseinmardi, H., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2016, August). 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* (pp. 186-192). IEEE Press. doi:10.1109/asonam.2016.7752233

Huang, Q., Singh, V. K., & Atrey, P. K. (2014, November). Cyber bullying detection using social and textual analysis. In Proceedings of the *3rd International Workshop on Socially-Aware Multimedia - SAM '14*. doi:10.1145/2661126.2661133

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the *8th International Conference on*

*Weblogs and Social Media* (ICWSM-14; 216-225). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109%5Cnhttp://comp.social.gate

Li, Q. (2006). Cyberbullying in schools. *School Psychology International, 27*(2), 157-170. doi:10.1177/0143034306064547

Mangaonkar, A., Hayrapetian, A., & Raje, R. (2015, May). Collaborative detection of cyberbullying behavior in Twitter data. In Proceedings of *International Conference on the Electro/Information Technology* (EIT; pp.611-616). doi:10.1109/eit.2015.7293405

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the *26th international conference on neural information processing systems* (pp. 3111-3119). Retrieved from http://dl.acm.org/citation.cfm?id=2999792.2999959

Miller, K. (2017). *Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limited available redress*. Retrieved from https://gould.usc.edu/why/students/orgs/ilj/assets/docs/26-2-Miller.pdf on July 20

Nahar, V., Li, X., & Pang, C. (2014). An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering, 3*(5), 238-247. Retrieved from: http://www.academicpub.org/cisme/paperInfo.aspx?paperid=13552

Nahar, V., Li, X., Pang, C., & Zhang, Y. (2013, January). Cyberbullying detection based on text-stream classification. In Proceedings of the *11th Australasian Data Mining*

*Conference* (pp. 49-58). Retrieved from http://crpit.com/confpapers/CRPITV146Nahar.pdf

Nixon, C. (2014). Current perspectives: the impact of cyberbullying on adolescent health. *Adolescent Health, Medicine and Therapeutics, 5*, 143-158. doi:10.2147/ahmt.s36456

Olweus, D. (1993). *Bullying in school: What we know and what we can do*. Oxford: Blackwell. ISBN: 978-063-11-9241-1

Olweus, D. (2012). Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology, 9*(5), 520-538. doi:10.1080/17405629.2012.682358

Patchin, J. & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyber bullying. *Youth Violence and Juvenile Justice, 4*(2), 148-169. doi:10.1177/1541204006286288

Ptaszynski, M., Masui, F., Nitta, T., Hatakeyama, S., Kimura, Y., Rzepka, R., & Araki, K. (2016). Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization. *International Journal of Child-Computer Interaction, 8*, 15-30. doi:10.1016/j.ijcci.2016.07.002

Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In Proceedings of the *10th International Conference on Machine Learning and Applications and Workshops* (ICMLA 2011; pp. 241-244). doi:10.1109/ICMLA.2011.152

Salawu, S., He, Y., & Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. *Transactions on Affective Computing*. doi:10.1109/taffc.2017.2761757

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior, 15*(2), 112-120. doi:10.1016/j.avb.2009.08.007

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8*(9), e73791. doi:10.1371/journal.pone.0073791

Singh, V. K., Huang, Q., & Atrey, P. K. (2016, August). Cyberbullying detection using probabilistic socio-textual information fusion. In Proceedings of the *International Conference on Advances in Social Networks Analysis and Mining* (ASONAM; pp. 884-887). doi:10.1109/asonam.2016.7752342

Smith, P. K., & Brain, P. (2000). Bullying in schools: Lessons from two decades of research. *Aggressive Behavior, 26*(1), 1-9. doi:10.1002/(SICI)1098-2337(2000) 26:1<1::AID-AB1>3.0.CO;2-7

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, 49*(4), 376–385. doi:10.1111/j.1469-7610.2007.01846.x

Spears, B., Slee, P., Owens, L., & Johnson, B. (2009). Behind the scenes and screens: Insights into the human dimension of covert and cyberbullying. *Zeitschrift Für Psychologie/Journal of Psychology, 217*(4), 189-196. doi:10.1027/0044-3409.217.4.189

Sugandhi, R., Pande, A., Agrawal, A., & Bhagat, H. (2016). Automatic monitoring and prevention of cyberbullying. *International Journal of Computer Applications, 8*, 17-

19. Retrieved from https://pdfs.semanticscholar.org/eb09/e30150f3adbe00cb3e384d45fdd7e7df70af.pdf

Van Cleemput, K., Vandebosch, H., Poels, K., Bastiaensens, S., DeSmet, A., & De Bourdeaudhuij, I. (2015). The development a serious game on cyberbullying: a concept test. In T. Vollink, F. Dehue, & C. McGuckin (Eds.), *Cyberbullying: from theory to intervention* (pp. 106-124). Abingdon, UK: Psychology Press. ISBN: 978-1848723382

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., ... Hoste, V. (2015, September). Detection and fine-grained classification of cyberbullying events. In Proceedings of the *International Conference Recent Advances in Natural Language Processing* (RANLP; pp. 672-680). Retrieved from https://biblio.ugent.be/publication/6969774/file/6969839.pdf

Van Royen, K., Poels, K., Vandebosch, H. (2016). Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites. *Children and Youth Services Review*, *64*, 35-41. doi:10.1016/j.childyouth.2016.02.024

Van Royen, K., Poels, K., Vandebosch, H., & Adam, P. (2017). "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior, 66*, 345-352. doi:10.1016/J.CHB.2016.09.040

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207. doi:10.3758/s13428-012-0314-x

Willard, N. (2005). *Educator's guide to cyberbullying addressing the harm caused by online social cruelty*. Retrieved from http://cyberbully.org

Wright, M. F. (2017). Cyberbullying in cultural context. J*ournal of Cross-Cultural Psychology, 48*(8), 1136–1137. doi:10.1177/0022022117723107

Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., ... Dillon, E. (2017). Cyberbullying detection with a pronunciation based convolutional neural network. In Proceedings of the *15th International Conference on Machine Learning and Applications* (ICMLA; pp. 740-745). IEEE. doi:10.1109/ICMLA.2011.152

Zhao, R., & Mao, K. (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing, 8*(3), 328-339. doi:10.1109/taffc.2016.2531682

Zhao, R., Zhou, A., & Mao, K. (2016, January). Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the *17th international conference on distributed computing and networking* (pp. 43-48). New York, USA: ACM Press. doi:10.1145/2833312.2849567

Table 1

*Methodological demands*

| Works | Annotators' Expertise | Inter-rater Reliability | Users' Permission | Peer Oriented | Cyberbullying Criteria |
|---|---|---|---|---|---|
| Bayzick et al., 2011 | 3 Undergrad. Research assistants | unknown | Crawl | No | 3 |
| Reynolds et al., 2011 | 3 Amazon Mechanical Turk | unknown | Crawl | No | 0 |
| Dinakar et al., 2011 | unknown | unknown | Crawl | No | 0 |
| Dadvar et al., 2012 | 3 students | unknown | Crawl | No | 1 |
| Dinakar et al., 2012 | unknown | unknown | Crawl | No | 0 |
| Nahar et al., 2013 | unknown | unknown | Crawl | No | 0 |
| Dadvar et al., 2013 | unknown | unknown | Crawl | No | 0 |
| Huang et al., 2014 | unknown | 0.93 | Crawl | No | 0 |
| Nahar et al., 2014 | 3 | unknown | Crawl | No | 0 |
| Chavan & Shylaja, 2015 | 2 | 0.69 | unknown | unknown | 0 |
| Mangaonkar et al., 2015 | unknown | unknown | Crawl | No | 0 |
| Van Hee et al., 2015 | unknown | unknown | Crawl | No | 1 |
| Ptaszynski et al., 2016 | Unknown experts (Internet Patrol) | unknown | Crawl | Yes | 1 |
| Singh et al., 2016 | unknown | unknown | Crawl | No | 0 |
| Al-Garadi et al., 2016 | 3 | unknown | Crawl | No | 0 |
| Zhao et al., 2016 | unknown | unknown | Crawl | No | 0 |
| Zhao & Mao, 2016 | 3 | unknown | Crawl | No | 0 |
| Sugandhi et al., 2016 | 3 Undergrad. Research assistants | unknown | Crawl | No | 3 |
| Hosseinmardi et al., 2016 | 5 | unknown | Crawl | No | 3 |
| Zhang et al., 2017 | 3 Amazon Mechanical Turk | unknown | Crawl | No | 0 |
| Authors, 2018b | 3 Amazon Mechanical Turk | unknown | Crawl | No | 0 |
| Authors, 2018a | 3 Amazon Mechanical Turk | unknown | Crawl | No | 0 |

*Note: "Users' permission" refers to whether or not the data was obtained with users' consent; "Peer Oriented" refers to whether the users from which data is extracted exist in a universe where they are peers; "Cyberbullying Criteria" relates to how many of the key factors of cyberbullying (repetition, aggression, intentionality and between peers occurrence) were used to instruct annotators.*

Table 2

*Cyberbulling dataset*

| Works | Dataset | Language | Dataset Size | Balancing |
|---|---|---|---|---|
| Bayzick et al., 2011 | MySpace1 | English | unspecified | unspecified |
| Reynolds et al., 2011 | Formspring1 | English | 3915 | .142 |
| Dinakar et al., 2011 | YouTube1 | English | 4500 | — |
| Dadvar et al., 2012 | MySpace2 | English | 2200 | unspecified |
| Dinakar et al., 2012 | Youtube1; Formspring2 | English | unspecified | unspecified |
| Nahar et al., 2013 | Twitter1, MySpace2, Kongregate and Slashdot (CAW 2.0) | English | 1570000 | unspecified |
| Dadvar et al., 2013 | YouTube2 | English | 4626 | .097 |
| Huang et al., 2014 | Twitter1 (CAW2.0) | English | 4865 | .019 |
| Nahar et al., 2014 | Kongregate; Slashdot; MySpace2 (CAW 2.0) | English | unspecified | unspecified |
| Chavan & Shylaja, 2015 | Kaggle (unspecified) | English | 2647 | .272 |
| Mangaonkar et al., 2015 | Twitter2 | English | 1340 | .152 |
| Van Hee et al., 2015 | AskFM | Dutch | 85485 | .067 |
| Ptaszynski et al., 2016 | Schoolboard Bulletins (BBS) | Japanese | 2222 | .128 |
| Singh et al., 2016 | Twitter1 subset (CAW 2.0) | English | 4865 | .186 |
| Al-Garadi et al., 2016 | Twitter3 | English | 10007 | .060 |
| Zhao et al., 2016 | Twitter4 | English | 1762 | .388 |
| Zhao & Mao, 2016 | Twitter5; MySpace1 | English | 7321 | .210 |
| Sugandhi et al., 2016 | Train (Formspring and MySpace); Test (Twitter) | English | 3279 | .120 |
| Hosseinmardi et al., 2016 | Instagram | English | 1954 | .290 |
| Zhang et al., 2017 | Formspring3 | English | 13000 | .066 |
| Authors, 2018b | Formspring4 | English | 13160 | .194 |
| Authors, 2018a | Formspring4 | English | 13160 | .194 |

*Note: Report on the characteristics and origins of the datasets used in the state-of-the-art. When explicitly described by the authors, we provide the final size of the dataset and the ratio of cyberbullying instances it contains (i.e., "Dataset Size" and "Balancing" columns).*

Table 3

*Cyberbullying systems performance*

| Studies | Best Classifier | Class | Accuracy | Precision | Recall | F1-Score | ROC | AUC |
|---|---|---|---|---|---|---|---|---|
| Bayzick et al., 2011 | Handmade Dictionary | nCB | | | .519 | | | |
| | Based Rules | CB | | | .853 | | | |
| | | Total | | | .586 | | | |
| Reynolds et al., 2011 | J48 | nCB | | | | | | |
| | | CB | | | .785 | | | |
| | | Total | | | | | | |
| Dinakar et al., 2011 | Jrip, LinearSVM | nCB | | | | | | |
| | | CB | | | | | | |
| | | Total | | | | | | |
| Dadvar et al., 2012 | SVM | nCB | | | | | | |
| | | CB | | .310 | .150 | .200 | | |
| | | Total | | | | | | |
| Dinakar et al., 2012 | SVM | nCB | | | | | | |
| | | CB | | | | | | |
| | | Total | | | | | | |
| Nahar et al., 2013 | Ensemble Classifier | nCB | | | | | | |
| | | CB | | | | | .400 | |
| | | Total | | | | | | |
| Dadvar et al., 2013 | SVM | nCB | | | | | | |
| | | CB | | | | | | |
| | | Total | | .770 | .550 | .640 | | |
| Huang et al., 2014 | Dagging | nCB | | | | | | |
| | | CB | | .763 | | | .755 | |
| | | Total | | | | | | |
| Nahar et al., 2014 | SVM | nCB | | | | | | |
| | | CB | | .870 | .970 | .920 | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Chavan & Shylaja, 2015 | Logistic Regression | Total<br>nCB<br>CB | | | | | |
| Mangaonkar et al., 2015 | Collaborative Paradigm | Total<br>nCB<br>CB | | .769 | .710 | | .869 |
| Van Hee et al., 2015 | SVM | Total<br>nCB<br>CB | .900 | .880 | .580 | | |
| Ptaszynski et al., 2016 | Proposed Method | Total<br>nCB<br>CB | | .500 | .554<br>.100 | | |
| Singh et al., 2016 | Proposed Method | Total<br>nCB<br>CB | | .820 | .530 | .640 | |
| Al-Garadi et al., 2016 | RandomForest | Total<br>nCB<br>CB | | | .890 | | |
| Zhao et al., 2016 | SVM | Total<br>nCB<br>CB | | .941 | .939 | .936 | .943 |
| Zhao & Mao, 2016 | smSDA | Total<br>nCB<br>CB | | .768 | .794 | .780 | |
| Sugandhi et al., 2016 | SVM | Total<br>nCB<br>CB | .849 | | | .719 | |
| Hosseinmardi et al., 2016 | LinearSVM | Total<br>nCB<br>CB | .913 | .910 | .910 | .900 | |
| | | Total | | .750 | .710 | .790 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Zhang et al., 2017 | CNN | nCB | | | | |
| | | CB | | | | |
| | | Total | .968 | .740 | .453 | .562 |
| Authors, 2018b | C-LSTM | nCB | | | | |
| | | CB | | .448 | .445 | .444 |
| | | Total | | | | |
| Authors, 2018a | Fuzzy Fingerprints | nCB | | | | |
| | | CB | | .355 | .597 | .425 |
| | | Total | | | | |

Table 4

*Cyberbullying feature engineering*

| Works | Representation | Embeddings | Textual Feat. | Social Feat. | User Feat. | Sentiment Feat. |
|---|---|---|---|---|---|---|
| Bayzick et al., 2011 | Dict. of keywords | | | | | |
| Reynolds et al., 2011 | num. & density of "bad" words | | 1 | | | |
| Dinakar et al., 2011 | TF-IDF | | 1 | | | |
| Dadvar et al., 2012 | TF-IDF | | 1 | | 1 | |
| Dinakar et al., 2012 | TF-IDF | | | | | 1 |
| Nahar et al., 2013 | TF-IDF | | | | | |
| Dadvar et al., 2013 | TF-IDF | | 1 | | 1 | |
| Huang et al., 2014 | | | 1 | 1 | | |
| Nahar et al., 2014 | Weighted TF-IDF | | | | | |
| Chavan & Shylaja, 2015 | TF-IDF | | 1 | | | |
| Mangaonkar et al., 2015 | TF-IDF | | | | | |
| Van Hee et al., 2015 | Binary Bag-of-Words | | 1 | | | 1 |
| Ptaszynski et al., 2016 | | | | | | |
| Singh et al., 2016 | Probabilistic | | 1 | 1 | | |
| Al-Garadi et al., 2016 | TF-IDF | | 1 | 1 | 1 | |
| Zhao et al., 2016 | TF-IDF + LSA | 1 | 1 | | | |
| Zhao & Mao, 2016 | Word Embeddings | 1 | | | | |
| Sugandhi et al., 2016 | TF-IDF | | | | | 1 |
| Hosseinmardi et al., 2016 | image + text (TF-IDF) | | 1 | | | |
| Zhang et al., 2017 | Phoneme-Based | 1 | | | | |
| Authors, 2018b | Word Embeddings | 1 | | | | |
| Authors, 2018a | Fuzzy Fingerprints | | | | | |

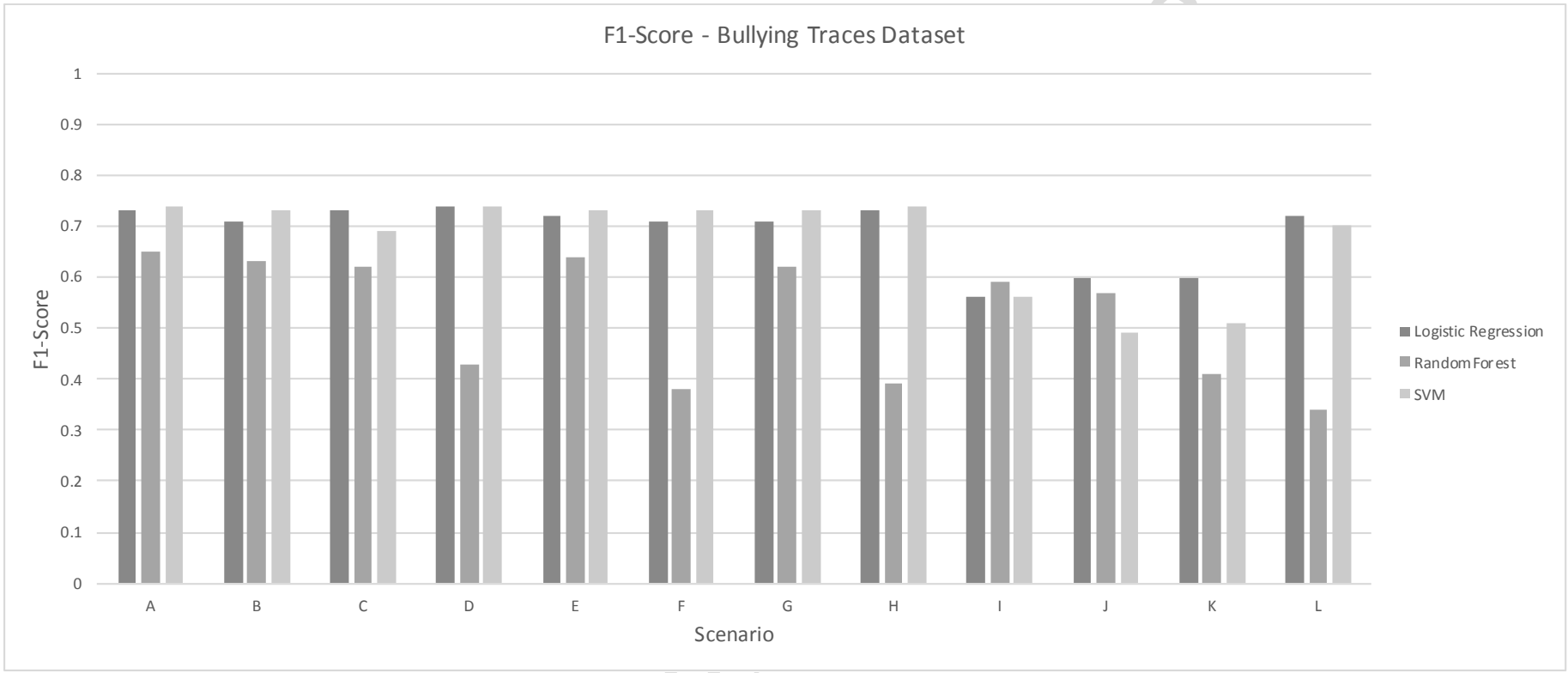



*Figure 1.* F1-Score for the Bullying Traces dataset, per scenario, per classifier.

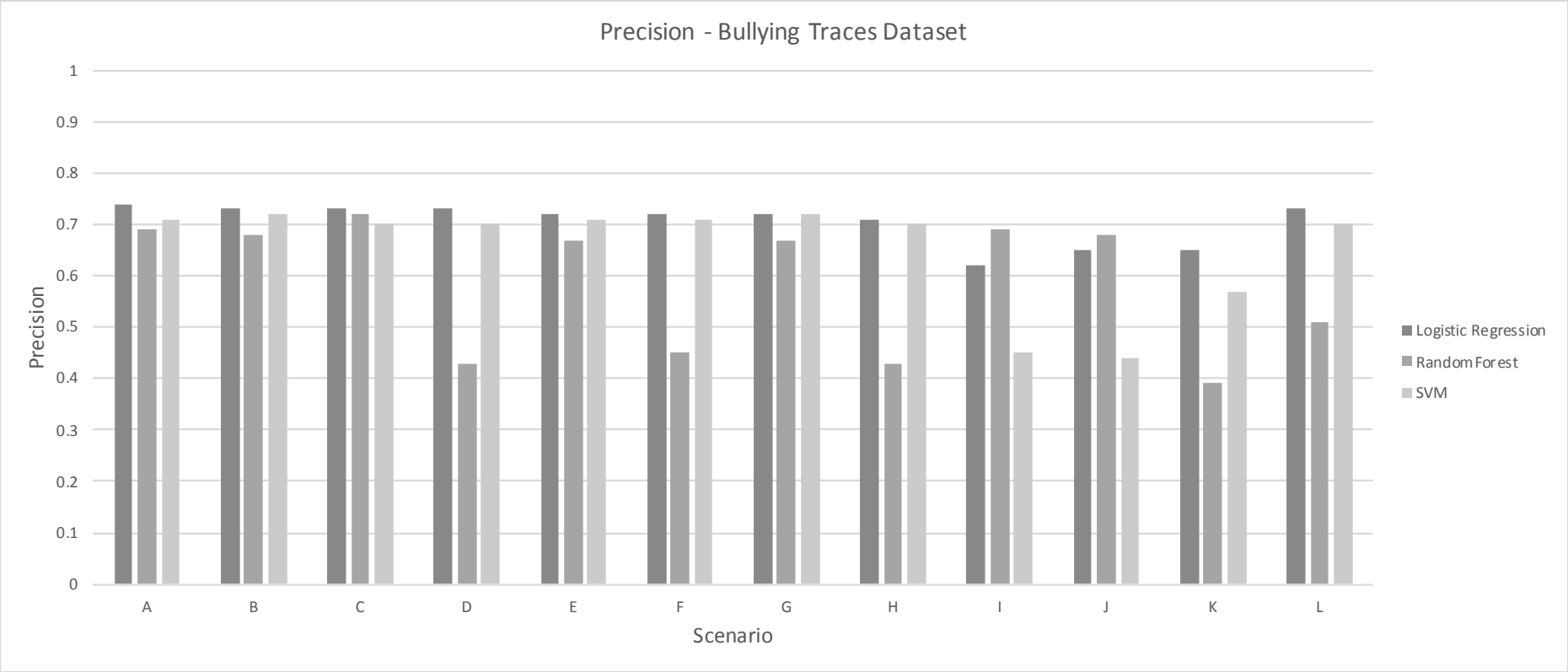



*Figure 2.* Precision for the Bullying Traces dataset, per scenario, per classifier.

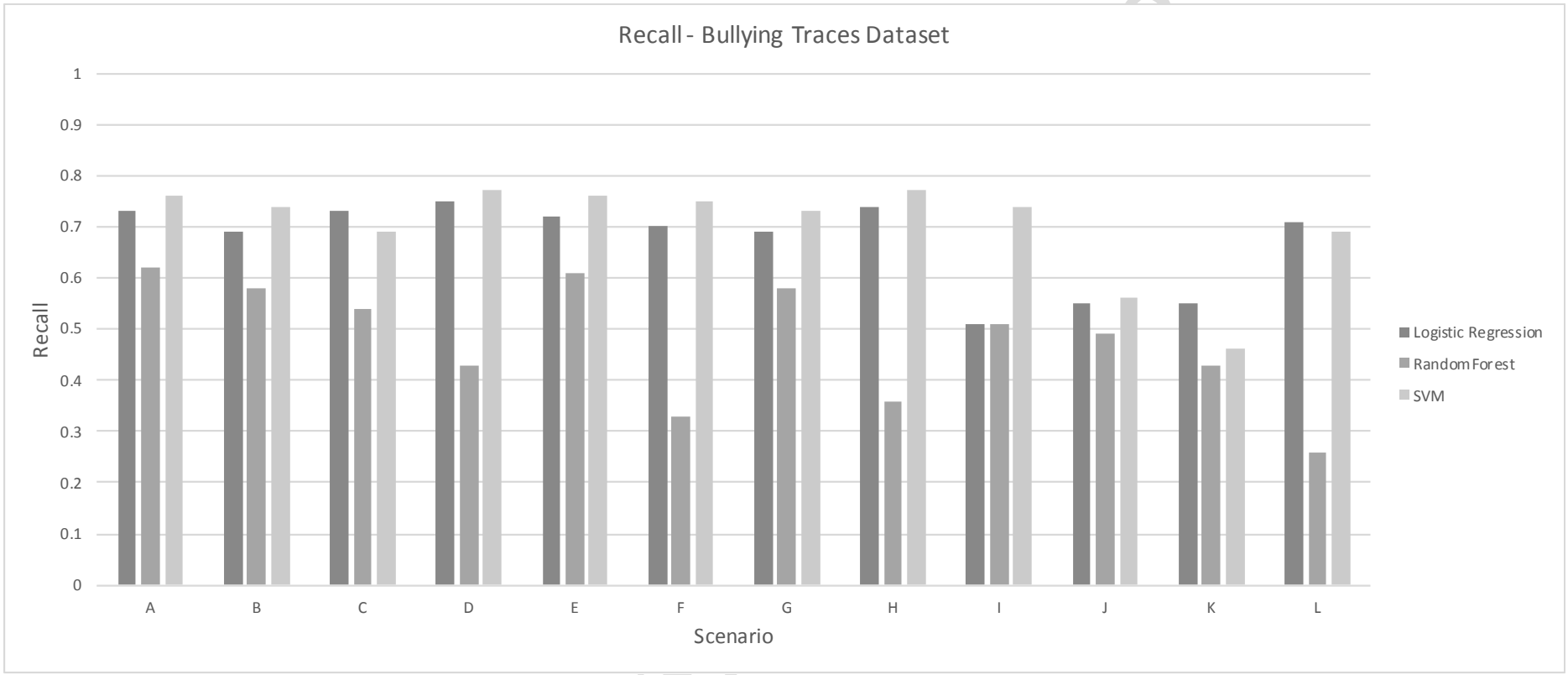



*Figure 3.* Recall for the Bullying Traces dataset, per scenario, per classifier.

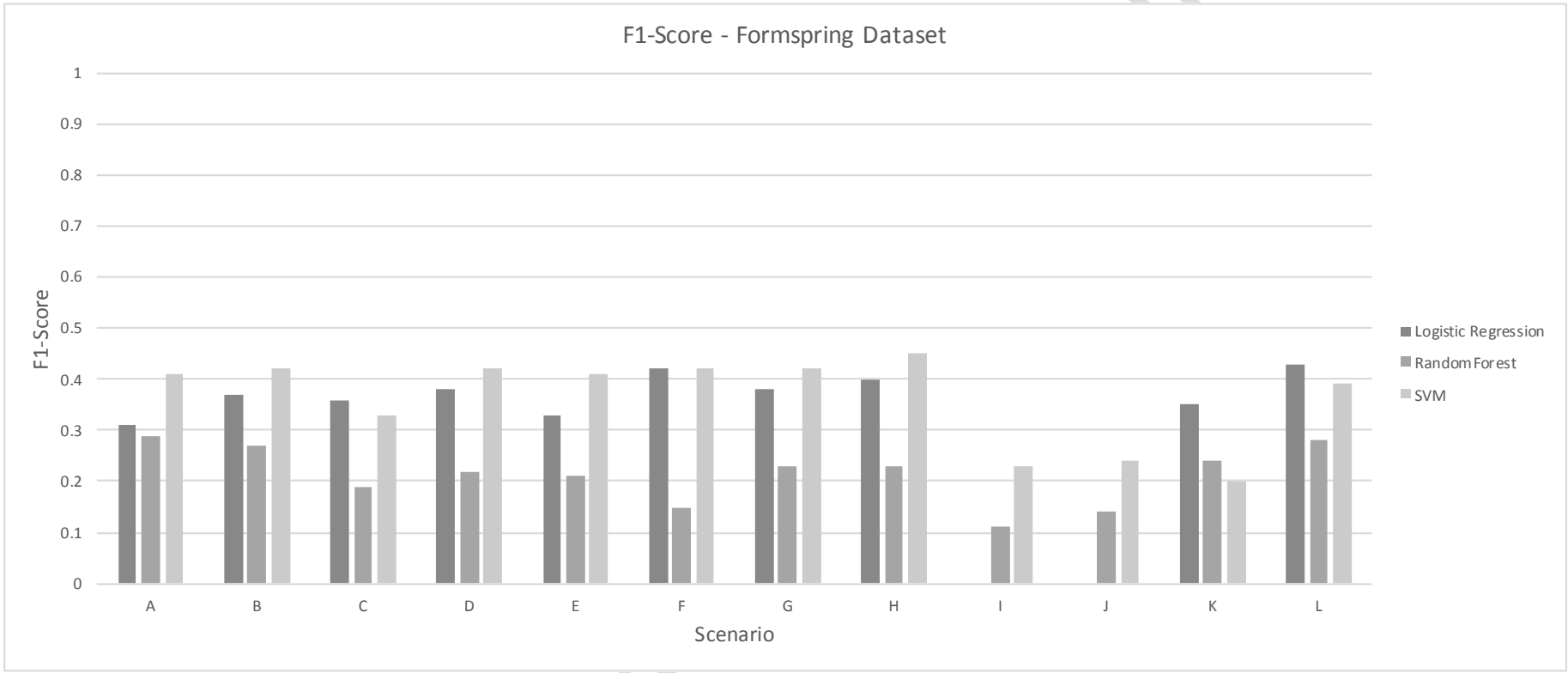



*Figure 4.* F1-Score for the Formspring dataset, per scenario, per classifier.

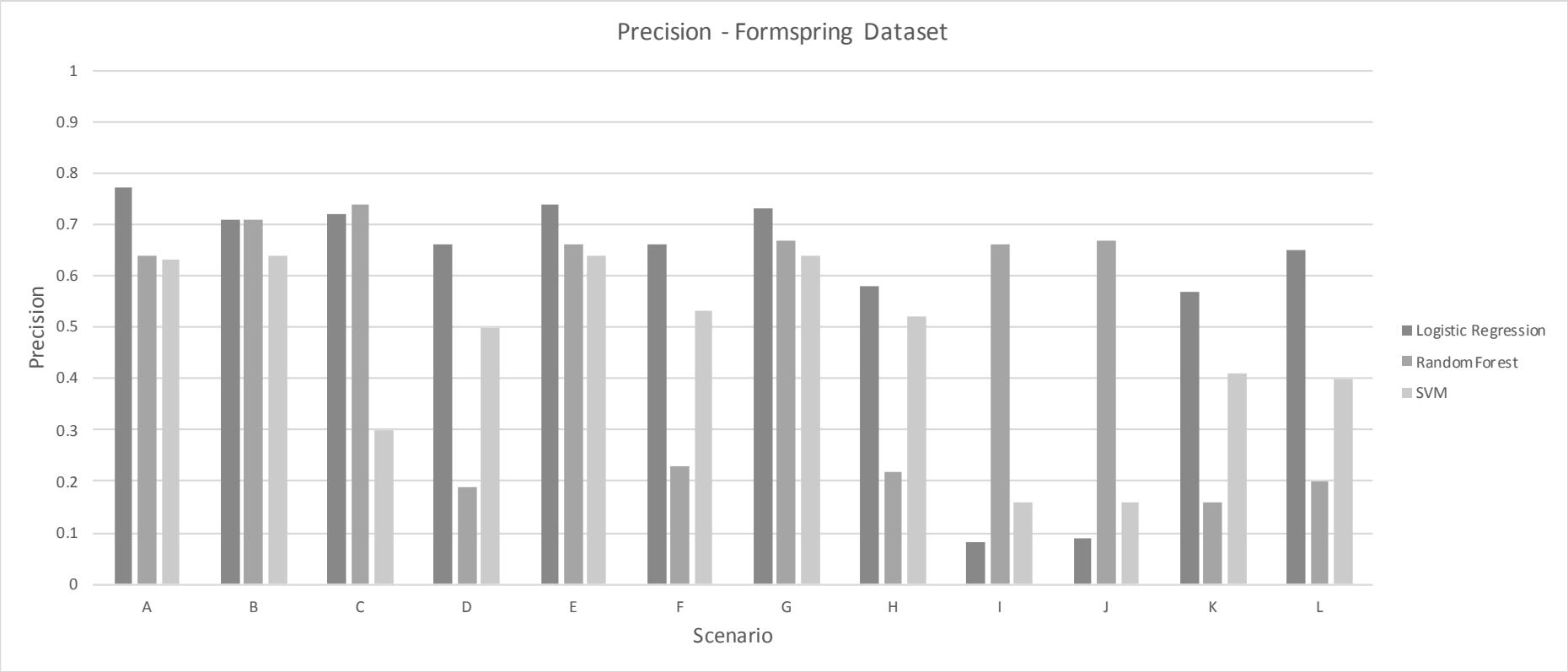



*Figure 5.* Precision for the Formspring dataset, per scenario, per classifier.

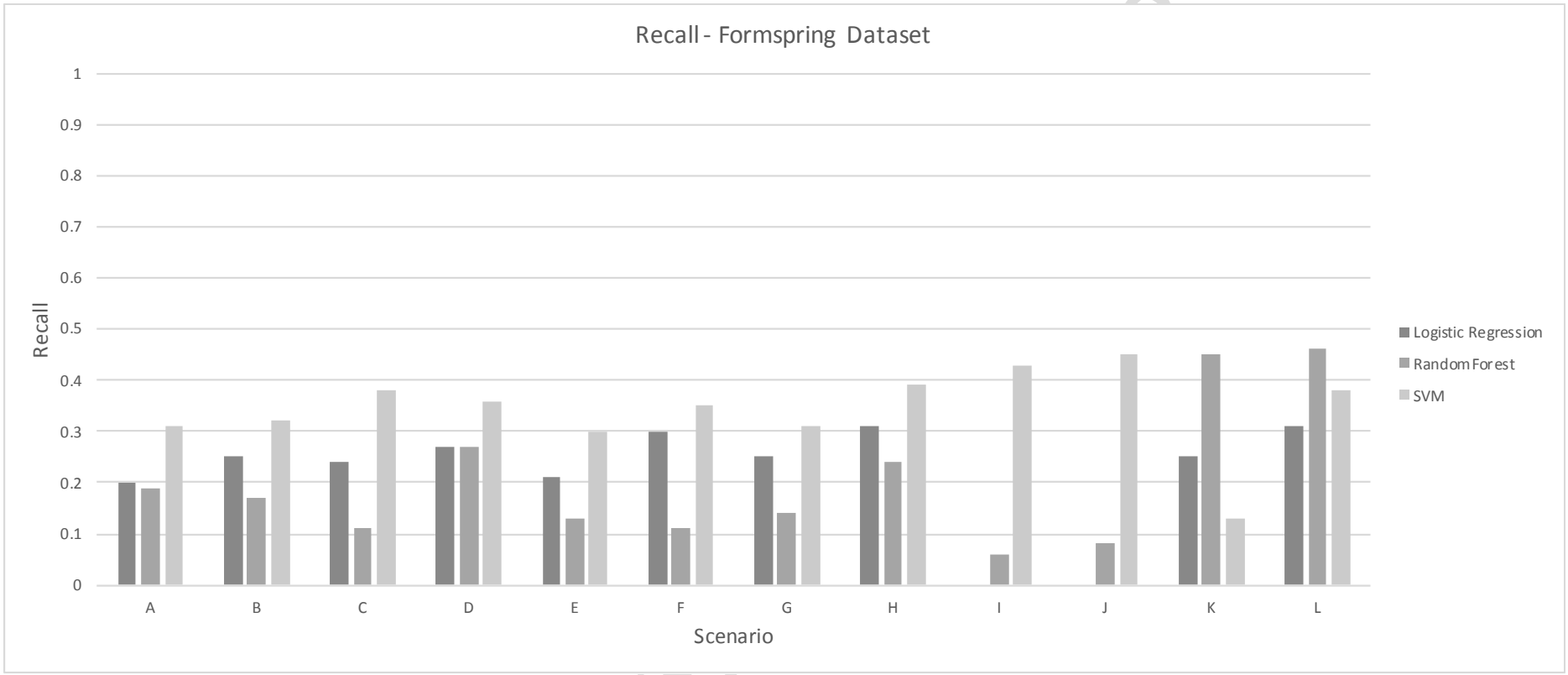



*Figure 6.* Recall for the Formspring dataset, per scenario, per classifier.

- Cyberbullying is often misrepresented in automatic detection state-of-the-art.
- Available systems do not capture all four key criteria of cyberbullying.
- Datasets used to train detection systems are incomplete.
- Feature engineering performance improvement is marginal.
- Cyberbullying detection systems seem not applicable to real world situations.