Cyberbullies in Twitter: A focused review

Nicolas Tsapatsoulis* and Vasiliki Anastasopoulou†
*Dept. of Communication and Internet Studies, Cyprus University of Technology, Cyprus
Email: nicolas.tsapatsoulis@cut.ac.cy
†Dept. of Psychology, National and Kapodistrian University of Athens, Greece

***Abstract*—This literature review focuses on recent techniques and studies regarding cyberbullying on Twitter. The purpose of these studies is to find smart and sophisticated ways and methods to detect these cyberbullying incidents and in case that it is not possible to fully eliminate them to provide the means to vastly reduce them. We present the theoretical roots of the term cyberbullying and we discuss thoroughly some influential studies to motivate new researchers that work in the general area of cybersecurity and privacy.**

***Index Terms*—cyberbullying, Twitter, online media, hashtags, cybersecurity, privacy**

I. Introduction

Bullying existed before the onset of the term 'cyber'. Cyberbullying and cyber aggression mirror the actions of bullying and aggression accordingly through the Internet. Aggression in bottom line is the oral statement or action to hurt someone just once while cyber-aggression is the use of electronic means towards a person or a group who perceive such behaviour as offensive, derogatory, harmful or unwanted. By extension cyberbullying is the intentional aggressive behaviour repeated over and over involving an imbalance of power. Basically, bullying is about dominating and our ancestors were big into dominance hierarchy. According to Boehm [1] "any species that has a social dominant hierarchy, like apes or monkeys or wild dogs or lions, has bullies". Additionally, he states that bullying is adaptive for many species, to anyone in any way "because you get better food or mating opportunities In primates, studies have shown that the top bullies have more offspring and therefore their genes proliferate". Thus, it seems that there is a payoff to it, the more you bully, the higher you will rise in social ranks. Although this seems to be something that was happening eons ago, the anonymity of the Internet and the ability to create plethora of fake accounts on the contemporary social media, boosted a new form of bullying, known as cyberbullying, to expand the last decade. A typical profile of a cyberbully refers to a person that might pass threats of violence, involved in sexual harassment, stalking or hate crimes or posting online sexually-explicit photos or messages [2].

Cyberbullying explode as a side effect of the widest use of the Internet by everyday people. Anonymous users began to take their roles in the stage by expressing themselves without taking into account the consequences to the other users on the Internet, based on the "safety" that the victims will never come to them. One of the earliest incidents of cyberbullying was recorded when a middle school boy in 1998 threatened his algebra teacher and the school principal, through a website which he created on his own[1]. Megan Meier, an American teenager, committed suicide by hanging herself and her suicide was attributed to cyberbullying through the social networking website MySpace. The mother of a friend of Meier was indicated on the matter but she was eventually acquitted.

The increasing frequency of cyberbullying incidents caused some actions of prevention and repercussions such as the anti-bullying statute and laws for schools to have strict policies for cyberbullying. Software tools for parents that support their role to prevent cyberbullying by monitoring electronic devices and Internet usage were also developed. However, cyberbullying is not an issue involving only teenagers, either as bullies or victims, it is much more involving people in every age and class. While the anti-bullying measures taken to protect minors are vey welcome and critical they caused, at the same time, a disorientation of the problem and underestimation of the dangers of cyberbullying on the rest of the human society.

II. Background

According to several researchers, the scourge of cyberbullying was started with #GamerGate[2]; a campaign that was organised to harass in online platforms. This hashtag was placed on a tweet by Adam Baldwin when he saw a couple of videos without any intention "of creating a hashtag of movement or anything like that", as he declared in an interview with Every Joe[3]. An ex-boyfriend got upset with his girlfriend, Zoe Quinn who is a game developer of "Depression Quest", and he wrote a disparaging post in a blog. The users of this hashtag accused Quinn falsely for unethical relationship with the journalist Nathan Grayson, causing Quinn and her family to be exposed to a virulent and often misogynistic harassment campaign. The people behind this campaign initially referred to as "quinnspiracy, but adopted the Twitter hashtag "Gamergate".

Supporters of Gamergate targeted in August 2014 women in the video game industry. They were organised on online platforms like *4chan*[4], Twitter, Reddit[5] and Internet Relay Chat[6] anonymously or using pseudonymous. Even though this

978-1-7281-3634-9/19/$31.00

[1]http://howtoadult.com/history-cyberbullying-6643612.html
[2]https://en.m.wikipedia.org/wiki/Gamergate_controversy
[3]http://www.everyjoe.com/2014/10/06/news/interview-adam-baldwin-gamergate-politics-ranger/
[4]http://www.4chan.org/
[5]https://www.reddit.com/
[6]https://en.wikipedia.org/wiki/Internet_Relay_Chat

was five years ago and it still exists and whenever cyberbullies need support for their post they just put on that hashtag and everything go as they "planned". Even the authors of the research article entitled "Hate is not binary: Studying abusive behavior of #Gamegate on Twitter" have reported harassment after they published their results. Probably that is because of the nature of Twitter: "Twitter is a word game and its 'Words with Frenemies' theyre going to do their best to win the word game", Adam Baldwin said. Till today there have been found 370 hashtags of #GamerGate and 369 co-appeared ones.

The need to detect and measure cyberbullying and cyberaggression is prominent even with a simple look on the numbers of cyberbullying behavior in teenagers: 15% of high school students in grades 9-12 were electronically bullied in the past year, 9% of students in grades 6-12 experienced cyberbullying and rest of them they were passive viewers of bullying in their school[7][3]. In the cyberspace, is very hard to detect or measure bullying or aggression. In contrary to the physical space, such as the schools, where things are manifesting physically and be observed, in the cyberspace cyberbullying occurs in web and social media platforms that work in very different ways. In addition, cyberbullies and cyber-aggregators organise themselves in online transient user groups, for example with the aid of hashtags [4], that are coming and going. Finally, the perception of what is cyberbullying or cyberaggression varies from country to country depending on cultural differences. Thus, among the main questions in studies concerning cyberbullying detection are the following:

- Q_1: What distinguishes abusers from regular social media users?,
- Q_2: Is it possible to develop practical methods to automatically detect abusing behavior on contemporary social media?
- Q_3: Could be collective human assessment through crowdsourcing platforms considered as a fair benchmark for those methods?'

The literature review that is presented next is organised across the axes defined with the three previous questions.

III. Abusers vs regular users in Twitter

A. Platforms

Online abuse comes in many forms and it appears in several platforms. People use to exaggerate, reaching the other side of negative manners; discrimination bullying, hate, even threats and attacks in real-world, towards a group or individually. On the other hand, the popularity of social media sites and the ease of crawling data through them consist, nowadays. the main data source for social research. Twitter may not be the most popular social media platform, since Facebook, Instagram and WhatsApp have more monthly active users, but it surpasses all of the in terms of data availability and as a results is the favour platform of researchers. The uniqueness of Twitter, for which many people argue on that, is the infrastructure in sense of freedom on following other user by any user. Also, it provides almost 100% of its data through its Twitter APIs [5]. Thus, for the majority of the researchers, Twitter is a huge library which archives human behaviour activity in terms of written messages. Through the Twitter hashtags [6] and the related norms, effective search for collecting data, is facilitated while major incidents, news stories and events are organised.

B. The information processing flowchart

Although various different approaches have been proposed for the detection of abusers in Twitter the most common information processing chain is the one depicted in Figure 1 and consist of the following steps: First, a research question is defined and the ideal types of data, that are required to answer this question, are decided. Data crawling and cleaning is involved next and Ground Truth sets are created. Deployment of Ground Truth on social science related problems has been traditionally done through questionnaires and / or interviews but unfortunately these methods are clearly inappropriate for online data and Twitter users. During the last years a common way to create Ground Truth sets is via crowdsourcing platforms, like Amazon Mechanical Turk[8] and *Figure-eight*[9]. Crowdsourcing platforms allow poling the crowd: people provide their assessments on various aspects and get paid for their work. As in every collective intelligence application, proper design of the user interaction for the crowdsourcing task is of primary importance. Users' selection is also important but for the task of sentiment assessment of tweets, as well as in many other annotation / labelling tasks, no special skills are required from the crowd; this is, in fact, one of the reasons crowdsourcing annotation became so popular the last few years. Finally, contemporary crowdsourcing platforms handle quite effectively the reliability issue [7] of the annotators.

Once a significant amount of annotated data, in our case tweets, have been produced sanity checks, based on basic statistics, are applied. Through these tasks an initial evidence regarding the appropriateness of the specific data types for answering the research question is obtained. Failing to pass the sanity check would either lead to collecting more data or changing the data types. The previous steps are executed in a repeated loop until proper and enough data are collected; otherwise the research question must be altered.

The remaining steps, namely feature extraction and models learning, are discussed in more detail next.

C. Feature Extraction

Feature extraction for identifying abusers in the Twitter involves three types of features: tweet - based, user - based and network - based. The text based features help us classify individual tweets into various categories according to their sentiment. A recent summary and assessment of various types of features, for tweet classification, was conducted by Tsapatsoulis and Djouvas [8]. Their main point in that work is that unigram based indices (i.e., keywords), either automatically constructed or explicitly indicated by humans, show the best

[7]http://www.pacer.org/bullying/resources/stats.asp

[8]https://www.mturk.com/mturk/welcome

[9]https://www.crowdflower.com/

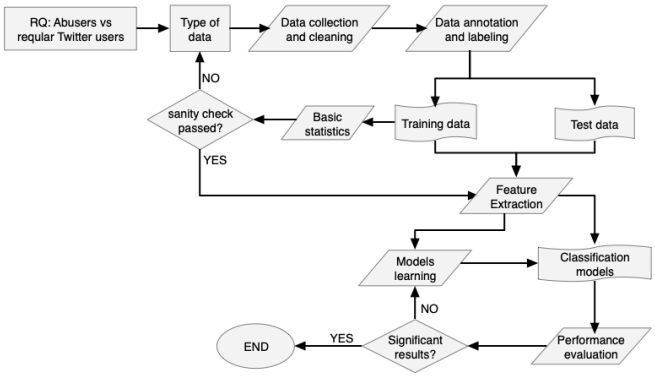

Fig. 1. A typical information processing chain for the detection of abusers in Twitter

performance among all compared feature sets, mainly due to the presence of hashtags, emoticons and slang words in tweets. Bigram (or generally n-gram) based features are totally inappropriate despite their usefulness in coping with negation. On the other hand, character based features and especially character n-grams of 4-6 characters length lead to tweet classification models with decent performance.

Word embeddings, which allow both semantic and syntactic correlation of words or tokens in general, gave a significant boost to tweet classification. In particular, sentence or short text implementation of the word embedding scheme led to two important variations, namely Doc2Vec [9] and *fasttext* [10], that are considered nowadays, in companion with deep learning methods, the state of the art.

Lexicon based tweet classification remains an option due to its simplicity and efficiency of implementation. Several techniques that are based on the *SentiStrength* [11] were proposed while a new trend is the crowd-derived lexicons [12]. As suggested by Tsapatsoulis and Djouvas [5], crowd lexicons combined with machine learning, achieve tweet classification performance close to that of *fasttext* and deep learning. Crowd lexicons can be mined form dedicated corpora, such as the *Hatebase*[10] database [12], or directly from filtered tweets [5].

There are some works that do not make use of text as the primary source of information for abuse detection. Singh *et al.* [13] report that there are some visual features that complement textual features in cyberbullying detection and can help improve predictive results. They used the Instagram as their data source platform [14] and through the Microsoft's Project Oxford (a computer vision API) extracted visual features from the Instagram images]. Those features included dominant colour, number of people present, adult content, age score, etc.

Network based features capture more permanent characteristics of social media users and they are commonly based on their egocentric [15] networks or approximations of those networks. According to Pieschl *et al.* [16], cyberbullies manipulate the emotional and behavioural state of victims by getting advantage of the Power Difference, i.e., the difference in power that a cyberbully has with respect to the users he/she mentions in his/her posts, in terms of their respective followers / friends ratio. Bullies are aggregated together and increasing their popularity by following each other, so their attack in a specific post has the form of a group assault. This observation lead us to the hypothesis that the form of the egocentric network of a cyberbully is a "core-periphery" one instead of a "connected component" which is the typical case of everyday social media users [15]. Typical centrality metrics, such as the hub and authority scores, the influence and closeness centrality, etc, can be used to discriminate between abusers and regular users but they require the full egocentric network to be available [17]. On the contrary, some measures such us the reciprocity, i.e., the extent to which a user reciprocates her / his follower connections she / he receives from other users, can be easily computed using only a rough approximation of the egocentric network of a social media user.

As correctly noticed by Zhao *et al.* [18], "many existing approaches in the literature are just normal text classification models without considering bullies' characteristics". User based features are about users profile and his/her activity within social media. While a user profile is mainly composed from static characteristics such as number of posts, likes, group participation, language, location, etc, the social media activity is highly dynamic. This activity is usually recorded through sessionization [19], i.e., by grouping tweets (or social media posts in general) which are close each other in time. The number of posts in each (fixed-time) session to filter out inactive or partially active users. Estimation of the optimal session duration and the inactivity threshold can be obtained through thorough analysis of the crowdsourced annotated training data. Feature extraction from those sessions, such percentage of abusive posts, targeted users, hashtags used, retweets and many others, provide important characteristics that can be used to classify a social media user as abuser or not. Chatzakou *et al*. [20] suggested three types of non-regular Twitter users: bullies, aggressors and spammers. Aggressors are the users who tweet or retweet at least one post and they have the intention to harm or insult other users. A bully is a user who posts multiple times targeting receivers that may not be able to defend themselves while a spammer is an advertiser [21] or marketing person who post texts for what he/she represents or texts that encompassing any other nature of phishing attempt.

D. Models Learning

There is no doubt that deep learning is the state of the art for most problems that can be approached through machine learning techniques. This trend was also reflected, the last few years, in cyberbullying detection as well. However, the emphasis is given on the sentiment analysis of tweets borrowing the initial work of Xu *et al.* [22] for text representation through deep learning, instead of the detection of cyberbullies as discussed above. Thus, traditional approaches that separate feature extraction from models' learning are in extended use

[10]https://hatebase.org/

because the majority of user based and network based features are high level ones [23] and semantically rich [24].

While there is no clear evidence on the appropriateness of specific machine learning algorithms for abuser models' learning it seems that tree based algorithms do work well and frequently reported in relevant works [25], [26]. For instance, Chatzakou *et al*. [26] experimented with various tree classifiers, such as J48, LADTree, LMT, NBTree, Random Forest, and Functional Tree in a four-class classification problem (bully, aggressive, spam, and normal users) and in a three-class classification problem (bully, aggressive, and normal users) and in both case reported Random Forest as the best models' learning algorithm. Al-garadi *et al.* [25], on the other hand compared Naive Bayes, SVM, Random Forest and KNN for cyberbullying detection and concluded that Random Forest achieves the second best performance behind Naive Bayes at the basic setting while achieve the best performance in case oversampling is used to account for unbalanced sets.

The WEKA tool [27] is a useful partner for many researchers for both data cleaning and feature selection as well as for models' learning. Another popular choice is the scientific toolkit library[11] of Python programming language.

E. Big Data and Scalability

The complexity of user modelling and categorisation based on the enormous amount of social media content that is exchanged everyday requires scalable solutions [28]. Most of the steps presented in Figure 1 can be parallelised. There is a flexibility of using different modelling algorithms and processing platforms dependent whether the data are batch form or are coming in a streaming fashion. Nevertheless, some of the steps, like the crowdsourcing annotations can be periodically executed. On the other hand, pipeline programming allows regular updates of the learned model benefitting a lot the performance, accuracy and extensibility of abuser detection applications. Additionally, new features can be plugged-in and different components can be updated or extended with new technologies for better data cleaning, feature extraction and modelling. A practical example on parallelisation for scalability on multiple machines using a Map-Reduce script for computing the top N keywords list from a huge amount of tweets is reported in Chatzakou *et al*. [20].

F. Evaluation Metrics

As in most detection (non ranking) problems, abuser detection performance is evaluated using standard machine learning performance metrics such as precision, recall, confusion matrix and the weighted area under the ROC curve (AUC). For examples on how all these measures are applied in practice see the work of Caruana and Niculescu - Mizil [29].

The statistical significance of results are usually assessed with typical t-tests, whenever two categories of users are compared, while ANOVA is employed when dealing with more than two categories of users. χ^2-tests in conjunction with a multiway ANOVA is applied whenever the influence of each feature type (see for instance Al-garadi *et al.* [25]) and/or machine learning algorithm, on the classification performance, is desired. A clear example of this can be seen in the work of Tsapatsoulis and Djouvas [5].

[11]https://scikit-learn.org/stable/

IV. Indicative Studies

In the following we analyse in depth some indicative studies that emphasise on abuser detection in Twitter. Detection of hate or aggressive speech in general is a problem that was extensively studied in previous works and reports and borrows techniques from the field of natural language processing, as nicely explained by Schmidt and Wiegand [30].

Al-garadi *et al.* [25], although mainly emphasised on the importance of various features they proposed for cyberbullying detection, followed very closely the architecture presented in Figure 1. They developed feature-based models for the classification of tweets into cyberbullying or non-cyberbullying ones using features extracted from the tweets' content but also from the users that post these tweets (user activity and characteristics) as well as user-network features, such as the number of friends following a user and the average number of followers to following. For tweet content analysis they used the number of vulgar words in the post, the presence of any of the 100 most commonly used words in social media that are positively correlated with neuroticism, the 100 most commonly used words in social media that are used by males, etc. Examples of features relate with user characteristics are the number of tweets, the second person pronouns and the number of mentions per tweet. The authors compared also three feature selection algorithms, namely χ^2, information gain, and Pearson correlation, to identify the most significant proposed features. Synthetic minority over-sampling and weights adjusting were used to balance the classes in the data set. According to the authors, the annotation of the 10606 tweets, they used in their study, was obtained with the aid of three experts, i.e., without the employment of a crowdsourcing platform.

Chatzakou *et al*. [20], [26] studied a dataset of 340k unique Twitter users that they had posted over than 1.6 million tweets within a three months period (June till August 2016). They have created two datasets with the aid of Twitter Streaming API: A baseline dataset, composed from one million tweets, corresponding to randomly selected tweets and a set of 659K tweets which probably express bad - inappropriate behaviour and named by the authors as *Gamegater* dataset. An interesting and innovative approach was adopted for the creation of the latter dataset. They started with tweets containing the hashtag *#GamerGate* as well as some other tokens indicated in the Hatebase database. In the second round they used the other hashtags identified in the previous set and the process continued with this snowball sampling procedure. By the end of the data collection period they reached 308 hashtags related with tweets with probably inappropriate content. The set of these hashtags created a first version of the *cyberbullying*

lexicon which was further expanded with common keywords identified with the aid of TF-IDF metric.

Each tweet in both the baseline and the *Gamegater* datasets was classified into one of three categories: aggressive, spam and normal, though the *Figure-eight* (previously known as *Crowdflower*) crowdsourcing platform. Every crowd-worker had to classify batches of 10 tweets into the above mentioned categories while every tweet was assessed by at least five different crowd-workers. The crowd-workers had to enter basic demographic information about themselves while their reliability was assessed by using typical inter-rater reliability measures on controlled cases, i.e., labeled by the job creator tweets.

During pre-processing stopwords, URLs, punctuation marks and numbers were removed from the tweets while a normalisation method was to eliminate repetitive characters. As far as the spam / marketing tweets (off topic tweets) is concerned the approach of Wang [31] was adopted. Wang considers two main Twitter spammer indicators: the use of large number of hashtags in tweets and posting a large number of tweets highly similar to each other. To find the optimal thresholds of these heuristics, the authors study the distribution of hashtags and the similarity of tweets. They found that Twitter users use between 0-17 hashtags per tweet and decided to remove tweets over five hashtags considering that those are stemming from spammers / marketeers. For the similarity filtering they remove the mentions and then compute the Levenshtein distance [32] between pairs of tweets. All users with an average intra-tweet similarity higher than 0.8 were excluded from the dataset.

The *cyberbullying lexicon*, mentioned above, was used as a binary index to characterise the individual tweets. User based features were extracted through sessionization of tweets, i.e., by grouping the tweet from the same user based on time clusters, and follows two dimensions spanning the emotional characteristics and the activity characteristics. The first refer to the assessed sentiment of tweets, the use of emotionally coloured or offensive works, and the use of emoticons and uppercase. Activity characteristics include the Twitter account age, how long the user is an active Twitter user, the number of posts, the participation in lists, as well as the number of favourites (likes), mentions, followers, friends, etc. Network based features were taken through a rough approximation of each Twitter user's egocentric network and included mainly centrality and modularity features as well as the reciprocity balance between ego and alters [17]. During the feature selection process, performed with the aid of Weka, some user based features were found non-informative and were excluded. Among those features is the account verification status, the default profile image, as well as main statistics on sessions, such as the average emotional and hate scores. Some network based features. such as closeness centrality and Louvain modularity were also found non-discriminative and were excluded as well.

For the classification of both tweets and users a variety of classifiers, including probabilistic ones (e.g., Nave Bayes), decision trees (e.g., J48), ensembles (e.g., Random Forests) as well as feed-forward neural networks, were adopted. According to the authors the tree based classifiers provide a good compromise between efficiency (time complexity) and effectiveness (accuracy of classification).

Chatzakou *et al*. [33] approached, aslo, the problem of cyberbullying on Twitter from a different perspective. In their work entitled "Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter" they analysed *GameGater* related (owned or mentioned by) accounts in terms of their status, i.e., active, deleted, suspended. Comparisons with normal Twitter accounts were also performed. They used the same or similar features as in their previous works [20], [26] and concluded that is that those features are meaningful in studying such user Twitter user behaviours, and probably useful in detecting what status a user should be given by Twitter. They adopted an unsupervised clustering approach to identify possible difference between the various types of accounts in terms of their clustering coefficient [34]. Among the various clustering techniques they used Random Forest achieved the best results in terms of both training efficiency and overfitting avoidance.

The identified differences between suspended and deleted users were quite interesting. Active GamerGaters express themselves more aggressively, they are, usually, repulsive and their posts have an intense notion of hate speech. On the other hand, their posts are assessed by the crowd-workers as being more joyful that of normal Twitter users. An overall 30% of active Gamegaters' posts are negative and the rest are positive. This positive percentage is higher than that of normal Twitter users! On the other hand, deleted GameGater accounts exhibit higher levels of anger but lower than the suspended GameGater accounts. They also express less joy but more sadness and fear, they type less in uppercase, compared to suspended accounts and normal Twitter users, but more than active Gamergaters. It is, therefore, very likely that deleted Gamergater accounts correspond to more emotionally introverted users and might be deleting their accounts to protect themselves from negative behaviours or attention.

Gamergaters, including suspended and deleted accounts, are more active than normal Twittr users, while deleted Gamergaters tend to have less friends and followers. On the contrary, the popularity of active and suspended Gamergaters is high; this is probably due to the traffic and interest they create in Twitter. The latter might be one of the reasons why their suspension was delayed. Suspended Gategaters used to post lot, had a lot of favourite lists. and they became popular in very short periods of time.

V. Conclusion

In this paper we attempted a focused review regarding the cyberbullying on Twitter. Our emphasis was given to the identification of Twitter abusers. Thus, we have clearly indicated all practical steps that are required for the development of effective applications for the detection of Cyberbullers. The current trends concerning the data annotation platforms, the feature types and machine learning models were detailed while

some indicative studies that made use of such tools were also presented. We believe, that this review will serve as a helpful starting point for young researchers that are planning to work on the field of Cyberbullying in Social Media as well as relevant topics such as privacy and cybersecurity.

ACKNOWLEDGMENT

The authors acknowledge research funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska - Curie ENCASE project, Grant Agreement No. 691025.

REFERENCES

[1] C. Boehm, *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York, USA: Basic Books, 2010.

[2] K. Goodboy, Alan and M. Martin, "The personality profile of a cyberbully," *Computers in Human Behavior*, vol. 49, no. C, pp. 1–4, Aug. 2015.

[3] A. Tsirtsis, N. Tsapatsoulis, M. Stamatelatos, K. Papadamou, and M. Sirivianos, "Cyber security risks for minors: A taxonomy and a software architecture," in *Proceedings of the 11th International Workshop on Semantic and Social Media Adaptation and Personalization*, ser. SMAP'16. IEEE, 2016, pp. 93–99.

[4] S. Giannoulakis and N. Tsapatsoulis, "Evaluating the descriptive power of instagram hashtags," *Journal of Innovation in Digital Ecosystems*, vol. 3, pp. 114 –129, 11 2016.

[5] N. Tsapatsoulis and C. Djouvas, "Opinion mining from social media short texts: Does collective intelligence beat deep learning?" *Frontiers in Robotics and AI*, vol. 5, p. 138, 2019.

[6] A. R. Daer, R. Hoffman, and S. Goodman, "Rhetorical functions of hashtag forms across social media applications," in *Proceedings of the 32Nd ACM International Conference on The Design of Communication CD-ROM*, ser. SIGDOC '14. ACM, 2014, pp. 16:1–16:3.

[7] J. Mao, K. Lu, G. Li, and M. Yi, "Profiling users with tag networks in diffusion-based personalized recommendation," *Journal of Information Science*, vol. 42, no. 5, pp. 711–722, 2016.

[8] N. Tsapatsoulis and C. Djouvas, "Feature extraction for tweet classification: Do the humans perform better?" in *Proceedings of the 12th International Workshop on Semantic and Social Media Adaptation and Personalization*, ser. SMAP'17. IEEE, 2017, pp. 53–58.

[9] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–1188–II–1196.

[10] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017, pp. 427–431.

[11] M. Thelwall, "The heart and soul of the web? sentiment strength detection in the social web with sentistrength," in *Cyberemotions: Collective Emotions in Cyberspace*, 10 2017, pp. 119–134.

[12] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Proceedings of the 12th International Conference on Web and Social Media*, ser. ICWSM'18, 2018, pp. 491–500.

[13] V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '17. ACM, 2017, pp. 2090–2099.

[14] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in *Social Informatics*. Springer International Publishing, 2015, pp. 49–66.

[15] S. P. Borgatti and D. S. Halgin, "On network theory," *Organization Science*, vol. 22, no. 5, pp. 1168–1181, Sep. 2011.

[16] S. Pieschl, T. Porsch, T. Kahl, and R. Klockenbusch, "Relevant dimensions of cyberbullying results from two experimental studies," *Journal of Applied Developmental Psychology*, vol. 34, p. 241252, 09 2013.

[17] S. Zenonos, A. Tsirtsis, and N. Tsapatsoulis, "Twitter influencers or cheated buyers?" in *Proceedings of the 3rd IEEE Cyber Science and Technology Congress*, ser. CyberSciTech 2018. IEEE, 2018, pp. 236–242.

[18] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th International Conference on Distributed Computing and Networking*, ser. ICDCN '16. ACM, 2016, pp. 43:1–43:6.

[19] M. Washha, A. Qaroush, and F. Sedes, "Leveraging time for spammers detection on twitter," in *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, ser. MEDES. ACM, 2016, pp. 109–116.

[20] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Measuring #gamergate: A tale of hate, sexism, and bullying," in *Proceedings of the 26th International Conference on World Wide Web Companion*, ser. WWW '17 Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 1285–1290.

[21] W. Shin and T. T.-C. Lin, "Who avoids location-based advertising and why? investigating the relationship between user perceptions and advertising avoidance," *Computers in Human Behavior*, vol. 63, pp. 444 – 452, 2016.

[22] Z. E. Xu, M. Chen, K. Q. Weinberger, and F. Sha, "From sbow to dcot marginalized encoders for text representation," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12. ACM, 2012, pp. 1879–1884.

[23] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 18:1–18:30, Sep. 2012.

[24] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Transactions on Affective Computing*, vol. 8, pp. 328–339, July-Sept 2017.

[25] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433 – 443, 2016.

[26] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in *Proceedings of the 2017 ACM on Web Science Conference*, ser. WebSci '17. ACM, 2017, pp. 13–22.

[27] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[28] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra, "Scalable and timely detection of cyberbullying in online social networks," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ser. SAC '18. ACM, 2018, pp. 1738–1747.

[29] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. ACM, 2006, pp. 161–168.

[30] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Apr. 2017, pp. 1–10.

[31] A. Wang, "Don't follow me: Spam detection in twitter," in *Proceedings of the 2010 International Conference on Security and Cryptography*, ser. SECRYPT'10. IEEE, 2010, pp. 1–10.

[32] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1091–1095, Jun. 2007.

[33] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Hate is not binary: Studying abusive behavior of #gamergate on twitter," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, ser. HT '17. ACM, 2017, pp. 65–74.

[34] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '07. ACM, 2007, pp. 29–42.