Inteligencia Artificial, 22(63) (2019), 81-100
doi: 10.4114/intartif.vol22iss63pp81-100

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

An experimental study on feature engineering and learning approaches for aggression detection in social media

Antonela Tommasel, Juan Manuel Rodriguez, Daniela Godoy
ISISTAN, CONICET-UNICEN, Argentina
antonela.tommasel@isistan.unicen.edu.ar
juanmanuel.rodriguez@isistan.unicen.edu.ar
daniela.godoy@isistan.unicen.edu.ar

Abstract With the widespread of modern technologies and social media networks, a new form of bullying occurring anytime and anywhere has emerged. This new phenomenon, known as cyberaggression or cyberbullying, refers to aggressive and intentional acts aiming at repeatedly causing harm to other person involving rude, insulting, offensive, teasing or demoralising comments through online social media. As these aggressions represent a threatening experience to Internet users, especially kids and teens who are still shaping their identities, social relations and well-being, it is crucial to understand how cyberbullying occurs to prevent it from escalating. Considering the massive information on the Web, the developing of intelligent techniques for automatically detecting harmful content is gaining importance, allowing the monitoring of large-scale social media and the early detection of unwanted and aggressive situations. Even though several approaches have been developed over the last few years based both on traditional and deep learning techniques, several concerns arise over the duplication of research and the difficulty of comparing results. Moreover, there is no agreement regarding neither which type of technique is better suited for the task, nor the type of features in which learning should be based. The goal of this work is to shed some light on the effects of learning paradigms and feature engineering approaches for detecting aggressions in social media texts. In this context, this work provides an evaluation of diverse traditional and deep learning techniques based on diverse sets of features, across multiple social media sites.

Resumen Con la difusión de nuevas tecnologías y los sitios de redes sociales surgió una nueva forma de acoso, que puede ocurrir en cualquier momento y lugar. Este nuevo fenómeno es denominado cyber agresión o acoso cibernético y hace referencia a actos agresivos e intencionales, cuyo objetivo es causar repetidamente daños a otras personas mediante comentarios insultantes, ofensivos, burlones o desmoralizadores a través de las redes sociales. Dado que estas agresiones representan una experiencia amenazadora para los usuarios de Internet, especialmente los niños y adolescentes, es crucial comprender cómo se produce el acoso cibernético para evitar que se intensifique. Teniendo en cuenta la gran cantidad de información que se comparte y distribuye en la Web, en los últimos tiempos ha cobrado importancia el desarrollo de técnicas inteligentes para la detección automática del contenido dañino. Esto potencialmente permite el monitoreo a gran escala de redes sociales, y la detección temprana de situaciones agresivas o no deseadas. A pesar de que en los últimos años se han desarrollado diversos enfoques basados tanto en técnicas tradicionales como en técnicas de aprendizaje profundo, diversas preocupaciones han surgido respecto a la duplicación de investigación y la dificultad para comparar resultados. Asimismo, no existe aún acuerdo respecto a qué tipo de técnica es mejor para la tarea, ni el tipo de características en las que se debe basar el aprendizaje. El objetivo de este trabajo es analizar el efecto de los diferentes paradigmas de aprendizaje y enfoques de ingeniería de características para la detección de agresión en redes sociales. En este contexto, este trabajo proporciona una

ISSN: 1137-3601 (print), 1988-3064 (on-line)

evaluación en múltiples redes sociales de diversas técnicas tradicionales y de aprendizaje profundo, basadas en diversos conjuntos de características.

Keywords: Cyberaggression, Social Media, Aggression detection, Feature Engineering, Machine Learning, Deep Learning
Palabras Clave: Cyber aggression, Medios sociales, Detección de agresión, Ingeniería de Características, Aprendizaje de Máquina, Aprendizaje Profundo

1 Introduction

People have fully embraced the Web and social media sites for socialising and communicating, and interact through different sites, such as *Facebook*, *Twitter*, *Instagram* and *YouTube* at the same time. These sites do not only allow users to create content, publish photos, comment on content other users have shared or tag content, but also foster the social connections between users. Nonetheless, alongside this vast exchange of information, ideas and friendships, undesirable phenomena and behaviours have appeared, this leads to the widespread dissemination of aggressive and potentially harmful content over the web. Even tough most of the time Internet use is safe and enjoyable, there are risks involving the online communications through social media. As the real world could be a dangerous place, social media sites are not the exception. Users might have to deal with threatening situations like cyberaggression, cyberbullying, suicidal behaviour or grooming [Whittaker and Kowalski, 2015].

Cyberbulling and cyberaggression are serious issues increasingly affecting Internet users. With the "help" of the widespread of social media networks, bullying once limited to particular places or times of the day (e.g. schools), can now occur anytime and anywhere [Chatzakou et al., 2017] and have a wider range of audience. Cyberaggression can be defined as aggressive online behaviour that intends to cause harm to another person [Hosseinmardi et al., 2015], involving rude, insulting, offensive, teasing or demoralising comments through online social media that target educational qualifications, gender, family or personal habits [Chavan and S, 2015]. Cyberbullying is one of the many forms of cyberaggression and is characterised by an act of online aggression, the existence of a power imbalance between the individuals involved (including diverse forms, such as physical, social, relational or psychological), and repetitions across time [Hosseinmardi et al., 2015]. This problem is aggravated by the persistence and durability of online materials, which gives these incidents an unprecedented power and influence to affect the lives of billions of people.

Links were found between experiences of cyberbullying and negative outcomes, such as decreased performance at school, dropping out and violent behaviour, in combination with devastating mental and psychological effects such as depression, low self-esteem, and even suicide [Hosseinmardi et al., 2015]. In recent years, there have been several high-profile cases involving teenagers taking their own lives in part for being harassed and mistreated over the internet. Additionally, cyberaggressive comments make their targets feel demoralised and frustrated, thus acting as a barrier for participation and socialisation. While these incidents are still isolated and do not represent the norm, their gravity demands deeper understanding [Hinduja and Patchin, 2010].

Considering the severity of the consequences that cyberaggression has on its victims and its rapid spread amongst internet users (specially kids and teens), there is an imperious need for research aiming at understanding how cyberbullying occurs, in order to prevent it or at least to decrease the harassing and bullying incidents in the cyberspace. Moreover, cyberaggression detection can be used to provide better support and advice for the victims as well as monitoring and tracking the bullies [Dadvar and de Jong, 2012]. Even though one incident cannot be a certain indication that the involved users are victims or bullies, following their behaviours after the incidents across different social networks during a period can provide a more accurate description of their profiles. Other important application of the detection of cyberaggression or aggressive content is the detection of cyberextremism, cybercrime and cyberhate propaganda [Agarwal and Sureka, 2015]. Most networking sites today prohibit the usage of offensive and insulting comments [Van Hee et al., 2015], which is partially being carried out and filtered to a limited extent. On the strategies to cope with aggressive behaviour online is to manually monitor and moderate the user-generated content. However, given the massive information overload on the Web and the pace at which new posts are being shared, it is unfeasible for human moderators to manually track and flag each insulting and offensive comments [Chavan and S, 2015]. Thereby, it is crucial to develop intelligent techniques to automatically detect harmful content, which would allow the large-scale social media monitoring and early detection of undesired situations.

Despite the seriousness of the problem, there are few successful efforts to detect abusive behaviour on social media data. This is related to the existence of several challenges [Chatzakou et al., 2017, Nobata et al., 2016], not only related to the nature of posts and the environment in which the aggression occurs, but also to technical limitations of the generalisation and comparability of the proposed approaches. One concern that arises from the approaches in the literature is related to the duplication of research and the difficulty of comparing results. Most works are evaluated over different datasets, which hinders the generalisation of their results. To advance towards solving this complex phenomenon, it is crucial to reach an agreed understanding of the different aspects of the problem, and the creation of standardised datasets [Kumar et al., 2018a], that would allow the comparison of approaches. In this context, this paper builds on a previous work [Tommasel et al., 2018] and focuses on the detection of aggressive content in the context of multiple and heterogeneous social media sites. In this context, the performance of several feature sets in the context of both traditional and deep learning techniques is evaluated in four publicly available datasets. Additionally, it is explored the feasibility of the selected feature sets and techniques for the identification of different types of accounts dedicated to distributing aggressive content. The goal is to both provide a comparison between different feature sets and techniques proposed in the literature over the same datasets to analyse the generalisation of results, and to shed some light on the usefulness or adequacy of the different techniques for the task, and the generalisation of models trained for a specific social media site to other sites.

The remainder of this paper is organised as follows. Section 2 describes recent related work regarding the detection of aggressive content. Section 3 describes the features considered in the analysis. Section 4 describes the experimental settings, including the selected datasets and implementation details. Section 5 analyses the obtained results for each of the selected datasets and their combinations. Finally, Section 6 presents the conclusions drawn from the study, and outlines future lines of research.

2 Related Work

Cyberaggression detection has captured the interest of researchers in the last years due to its proliferation across social media and its harmful effect on people. Consequently, automatic approaches for cyberbullying detection, mostly based on machine learning and natural language processing techniques, are being extensively developed. The detection of cyberbullying and online harassment is often formulated as a classification problem [Salawu et al., 2017], in which techniques traditionally used for document classification, topic detection, and sentiment analysis are used to detect electronic bullying based on the characteristics of messages, senders, and recipients. Thereby, feature engineering (i.e., the process of analysing and designing predictive features) becomes an important step in this process, as it allows to enhance the performance of techniques. Features can be broadly categorised into four groups [Salawu et al., 2017], namely content-based (e.g. bag-of-words, n-grams and part-of-speech tagged words), sentiment-based (e.g. positive and negative emotions), user-based (e.g. demographic information) and network-based features (e.g. number of friends and activity). Several sets of these four types of features have been used in the literature with both classical machine learning methods and methods in the deep learning paradigm. The former methods, such as Support Vector Machines (SVM), Naïve Bayes or Logistic Regression, rely on manually engineered features that are then used for training the models, whereas the latter methods employ neural networks to automatically learn features from raw data.

Most efforts related to cyberaggression detection focus on the detection of the actual aggressive events and their classification. In this regard, Van Hee et al. [2015] explored the identification and fine-grained classification of events into seven categories (non-aggressive, threat/blackmail, insult, curse, defamation, sexual talk, defence and encouragement to the harasser) based on two types of lexical features. First, bag-of-words features including unigrams, bigrams and character trigrams. Second, polarity features including the number of positive, negative and neutral lexicon words averaged over text length and the overall post polarity. In total, the authors created more than $300k$ features. Experimental evaluation was based on approximately $80k$ Dutch posts belonging to *Ask.fm*, which were manually labelled with a Kappa score of 0.69. Results achieved an average F-Measure of 0.55 and a minimum of 0.07 for the defamation class. The authors hypothesised that the discrepancy of results arose from the differences in post lexicality. For example, insults are generally highly lexicalised, whereas threats are often expressed in an implicit way.

Nobata et al. [2016] focused on the detection of hate speech on 2 million online comments from two domains (*Yahoo!* Finance and News) based on four types of features: n-grams, linguistic, syntactic and distributional semantics. Linguistic features explicitly look for inflammatory words and non-abusive language elements, such as the usage of politeness words or modal verbs. Distributional semantic features refer to features derived from word embeddings. The online comments were pre-processed by normalising numbers, replacing unknown words with the same token and replacing repeated punctuation. The best F-Measure results were obtained when combining all features. Interestingly, the individual sets of features obtaining the best results varied according to the dataset domain. For the finance dataset, the best individual results were obtained by n-grams, whilst for the news dataset, the best results were obtained by the embedded features. The authors hypothesised that the selected features could achieve good performance in other languages, although it remains to be evaluated.

Similarly to [Nobata et al., 2016], Chavan and S [2015] distinguished bullying and non-bullying comments by means of TF-IDF weighted n-grams, the presence of pronouns and skip-grams. Feature selection was then apply to select only the $3,000$ highest features according to χ^2. Experimental evaluation was based on approximately $6.5k$ comments from an unspecified site. Pre-processing was applied by removing non-word characters, hyphens and punctuation and applying a spell-checker. The best classification performance was achieved by selecting pronouns and skip-grams, with differences up to a 4.8% regarding the other features.

All previously described approaches are based on training traditional supervised models such as SVM, Naïve Bayes and Logistic Regression. Nonetheless, other approaches based on lexicons [Pérez et al., 2012, Bretschneider et al., 2014] and rules [Chen et al., 2012, Bretschneider et al., 2014] have been also proposed. For example, Pérez et al. [2012], Bretschneider et al. [2014] proposed using established patterns of aggression to detect bullying on lexically processed text. Particularly, Bretschneider et al. [2014] formulated rules to recognise word patterns indicating relationships between profane words and personal pronouns. Experimental evaluation was based on publicly available *Facebook* posts[1]. Chen et al. [2012] computed the offensiveness score of *YouTube* comments posted by over 2 million users based on a set of rules and syntactic features mined using the Stanford parser[2]. According to the authors, their rule-based approach was able to improve the results obtained with SVM and Naïve Bayes classifiers. Additionally, Fahrnberger et al. [2014] aimed at not only detecting the aggressive content in chat rooms, but also to replace it with alternatives suitable for minors extracted from WordNet[3].

The cyberbullying detection task does not only focus on the detection of the actual aggressive content, but also on the detection of users sharing it. In this context, Chatzakou et al. [2017] identified aggressive users by combining user, text, and network-based features. Experimental evaluation was based on the *#GamerGate* controversy in *Twitter*, including approximately $650k$ tweets, which were manually labelled into three categories (aggressor, bully and spammer). Tweets were pre-processed by removing numbers, stopwords, punctuations and converting the remaining words to lower case. According to the authors, their approach achieved an overall precision of 0.89, and precisions of 0.295 and 0.411 for the aggressive and bully classes, exposing the difficulties of the task. Results showed that features did not have the same relevance for the task. For example, considering session statistics, average emotional scores, hate score, average word embedding and community information added noise to the classification, whilst most text features did not contribute to the improvement of results. Conversely, the most effective features were the network-based ones. Moreover, the authors found that no statistical difference was found for concrete sentiment features (e.g. anger, disgust, fear, joy) amongst the abusive and normal posts.

Recently, the importance of detecting aggressive content motivated the development of competitions and shared tasks, like TRAC 2018 Shared Task on Aggression Identification[4] [Kumar et al., 2018a] and the MEX-A3T track at IberEval 2018[5] [Alvarez-Carmona et al., 2018]. The former aimed at discriminating between overtly aggressive, covertly aggressive (which resulted the most difficult class to predict), and non-aggressive texts. To that end, participants were provided with a training dataset of $15k$ aggression-

[1] `http://www.ub-web.de/research/index.html`

[2] `https://nlp.stanford.edu/software/lex-parser.shtml`

[3] `https://wordnet.princeton.edu/`

[4] `https://sites.google.com/view/trac1/shared-task`

[5] `https://sites.google.com/view/ibereval-2018`

annotated *Facebook* Posts and Comments in both Hindi and English, and two test sets. One of the text sets was extracted purely from *Facebook* and the other mixed posts from *Facebook* and other social media site. In turn, the MEX-A3T track was oriented to discriminate between aggressive and non-aggressive tweets in Mexican Spanish language. In this case, participants were provided with an annotated corpus of more than $11k$ tweets. Similar alternatives were proposed for both tasks, ranging from using traditional features for training SVM and random forest classifiers, to using more sophisticated deep neural network techniques.

For example, Samghabadi et al. [2018] combined lexical and semantic features. Lexical features included TF-IDF weighted word n-grams, char n-grams, and k-skip n-grams, whilst semantic features were extracted using vectors trained using Google News. Additional feature vectors included sentiment features, LIWC [Pennebaker et al., 2001] features, and the gender probabilities of each message. Ramiandrisoa and Mothe [2018] combined random forest and logistic regression classifiers. The former classifier was based on different set of features adapted from depression detection tasks and included common features like punctuation or emotions and others tending to analyse the readability and comprehensibility of texts. On the other hand, the latter classifier was based on document vectorization with doc2vec [Le and Mikolov, 2014].

As regards the deep learning based techniques, Aroyehun and Gelbukh [2018] leveraged on pre-trained word embeddings including word2vec[Mikolov et al., 2013], GloVe [Pennington et al., 2014], SSWE and fastText [Joulin et al., 2017], which exhibited the highest vocabulary coverage. Galery and Charitos [2018] combined pre-trained English and Hindi fastText word embeddings by means of pre-computed SVD matrices to join the representations from both languages into a single space. Roy et al. [2018] proposed an ensemble of classifiers that included a SVM classifier trained with TF-IDF vectors of unigrams, and a Convolution Neural Network (CNN) with pre-trained GloVe vectors. Risch and Krestel [2018], Frenda and Banerjee [2018], Gómez-Adorno et al. [2018] explored the usage of traditional features (e.g. unigrams, bigrams, and character n-grams, POS tags, named entities and sentiment polarity) for training deep learning models. Particularly, Risch and Krestel [2018] focused on logistic regression and bidirectional short-term memory (LSTM) networks learning. Frenda and Banerjee [2018] also included features related to style and writing density as well as the extraction of syntactic patterns, lists of aggressive words and affective features, and Gómez-Adorno et al. [2018] included morphological features.

The reviewed approaches in most cases show several of the challenges that have yet to be tacked by the selected features and techniques. For example, the lack of grammar and syntactic structure of social media posts, which hinders the usage of natural language processing tools. For example, the intentional obfuscation of words and phrases to evade checks. On the other hand, abusive content might span over multiple sentences. Second, the limited context provided by each individual post, causing that an individual post might be deemed as normal text, whilst it might be aggressive in the context of a series of posts. Third, the fact that aggression could occur in multiple forms, besides the obvious abusive language. For example, the usage of irony and sarcasm. Fourth, the difficulty of tracking racial and minority insults, which might be unacceptable to one group, but acceptable to another. Moreover, the difficulty for detecting aggression might depend on the language in which the aggression is made.

In summary, the main concern of the summarized works was to improve classification results through the application of both classical machine learning algorithms as well as deep learning approaches. However, a conclusion regarding which approach is more adequate to the task at hand has not yet been reached, as the performance of the different approaches did not seem to differ much. For example, if features are carefully selected, then classifiers like SVM and even random forest and logistic regression perform at par with deep neural networks. Similarly, if deep learning approaches are not carefully designed they might perform poorly.

From the reviewed approaches, concerns over the duplication of research and the difficulty of comparing results arise. Most works are evaluated over different datasets, which hinders the generalisation of results. To advance towards solving this complex phenomenon, it is crucial to reach an agreed understanding of the different aspects of the problem and the creation of standardised datasets [Kumar et al., 2018a], that would allow the comparison of approaches. In this context, this paper builds on a previous article [Tommasel et al., 2018] to provide a more extensive evaluation of both traditional and deep learning techniques over not only four publicly available datasets, but also on their cross-combinations.

3 Characterising Aggression

Many studies about aggression or bullying detection have assessed the capabilities of content, sentiment, user, network-based features or a combination of them. However, there is still no consensus regarding which features perform better for characterising and detecting acts of aggression. This situation is evidenced by the variability of results and the diversity of social media sites, which have their own intrinsic characteristics. The main goal of this work is to study different feature sets and their performance for detecting aggression in different social media sites. In particular, this work focuses on four type of features: character-based, word-based, sentiment-features, and irony-features. Additionally, this work also aims at shedding some light on the generalisation of features and models trained for a specific social media site to other sites, i.e. for cross social media cyberaggression detection.

Character-, syntactic- and word-based features have been traditionally used in most text related tasks. Character-based features usually include the number and ratio of punctuation marks (e.g. question marks, exclamation marks, period, commas and ellipses), the number and ratio of upper case letters, and the number and ratio of emoticons/emojis. Syntactic features are associated to the part-of-the-speech (POS) tagging of text, and usually include the number and ratio of nouns, verbs, adverbs and adjectives, or the selection of only a particular type of word. For example, only words tagged as nouns, verbs, adjectives and adverbs could be selected. Word-based features can include stemming, lemmatisation, name entity recognition, average word length, number of synonyms base on *WordNet*[6], commonness of words based on the American National Corpus[7] and frequency of rarest word. Word Embeddings can also be considered be used to define features.

Since cyberbullying has been associated to negative emotions, such as anger, irritation, disgust and depression, sentiment-based features might be useful for characterising aggression and cyberbullying. Sentiment detection could refer either to identifying the overall polarity of texts, i.e. whether the text yields a positive, negative or neutral polarity, or to identifying specific emotions, such as anger, joy, love or hate, amongst others. For the purpose of the aggression detection task, social media post were characterised according to their overall polarity. Polarity is associated to the diverse syntactic structures of posts, the polarity of their individual words, the number and ratio of curse words. Two pre-trained sentiment models are considered: *StandordNLP* and *SentiWordNet*[8]. In addition to words, emoticons and emojis are an integral part of social media language and they are considered to deliver sentiment information. In this context, the analysis includes the average sentiment polarity of emojis in posts based on the Emoji Sentiment Ranking [Kralj Novak et al., 2015][9].

Finally, irony based features might be helpful for detecting cyberbullying and cyberaggression. As irony is meant to communicate the opposite of the literal interpretation of the expressions, it might be used to masked aggression. For example, a congratulation might be actually a mocking about a bad outcome. Ironic statements can elicit affective reactions [Hernańdez Farías et al., 2016] For example, ironic criticism has been recognised as offensive and associated with particular negative affective states, which could enhance negative emotions such as anger, irritation or disgust. In this context, the feature sets defined in [Barbieri and Saggion, 2014, Hernańdez Farías et al., 2016] are also considered. Such feature sets focus on both character- and word-based features, and emotive word lists and lexicons (e.g. AFFIN[10], the lexicon created by [Hu and Liu, 2004] and the Whissell's Dictionary of Affect in Language[11]).

4 Experimental Settings

This section describes the experimental settings considered for evaluating the capabilities of of the selected features for cyberaggression detection in social media, and is organised as follows. Section 4.1 outlines the data collections used. Then, Section 4.2 describes the process for extracting the features and creating the posts representations. Finally, Section 4.3 describes implementation details.

[6] https://wordnet.princeton.edu/

[7] http://www.anc.org/

[8] http://sentiwordnet.isti.cnr.it/

[9] http://kt.ijs.si/data/Emoji_sentiment_ranking/

[10] http://neuro.imm.dtu.dk/wiki/AFINN

[11] https://www.god-helmet.com/wp/whissel-dictionary-of-affect/index.htm

4.1 Data Collections Used

The performance of the aggression detection was evaluated considering four data collections gathered from diverse social media sites. Table 1 summarises the general characteristics of the selected collections. Unless data collections were already separated into training and test set, they were randomly split 70% training and 30% test sets.

Kumar et al. [Kumar et al., 2018b] It was made public as part of the challenge of the TRAC 2018 Shared Task on Aggression Identification[12] and comprises more than $17k$ posts extracted from *Twitter* and *Facebook*. Posts were collected from Hindi pages related to news, forums, political parties, student's organisations, and groups in support or in opposition to recent incidents. Most of the posts are in English, some contains Hindi word or expressions, and a minority are completely in Hindi. Human annotators assigned posts to either one of three classes, overtly aggressive, covertly aggressive and non aggressive. According to the authors, the best classification achieved a F-Measure of 0.7. However, the authors did not specify which features were used. The collection is divided into four different sets of posts. The first two only includ $15k$ *Facebook* posts, and are intended as training and validation datasets. The other two collections were intended as testing datasets and comprise posts from different social media sites. The first one includes 916 *Facebook* posts, whilst the other $1,257$ *Twitter* posts.

Davidson et al. [Davidson et al., 2017] It comprises approximately $25k$ tweets containing terms compiled by *Hatebase*[13]. Tweets were assigned to one of three classes (hate speech, offensive but not hate speech and neither hate speech nor offensive) by human annotators. The agreement of the labelling was 92%. Interestingly, only 5% of the tweets were coded as hate speech by the majority of coders, showing the limitations of the *Hatebase* lexicon. According to the authors, the best classification achieved an F-Measure of 0.9, when considering n-grams and POS information.

Reynolds et al. [Reynolds et al., 2011] It comprises approximately $3k$ questions and answers extracted from the ask and answer site *FormSpring.me*[14]. Posts are manually labelled into three categories (strongly aggressive, weakly aggressive and non-aggressive). Each post was labelled by three different taggers to improve the labelling quality. According to the authors, the best classification achieved an overall accuracy of 81%, when considering features related to the number and intesity of curse words.

Chatzakou et al. [Chatzakou et al., 2017] it comprises tweets gathered between June and August 2016, in relation to the *GamerGate* controversy, which is one of the most well documented and mature, large-scale instances of aggressive behaviour. It focuses on the classification of users instead of posts. Collection started with the "#GamerGate" hashtag and continued with co-ocurring hashtags. Additionally, a random set of tweets was crawled, as it was assumed less likely to contain offensive behaviour. Unlike the other selected collections, this one focused on classifying users according to their behaviour instead of classifying each independent post.

4.2 Feature Extraction and Post Representation

For the purpose of feature extraction, posts were first pre-processed by removing all non-standard characters, such as non-printable and control characters. Syntactic-based features, such as character- or word-based features, required text tokenisation, which was carried out using two tools, one specifically designed for social media (*twokeniser*[15]), and one of general purpose (*StanfordNLP* library[16]) English stopwords were also removed. The set of extracted features and their combinations in presented in Table 2.

Once the tokens were obtained, post representations were built according to the defined feature sets. Two strategies were followed for describing posts. The first strategy mapped posts to feature vectors. This strategy represents posts considering all character-, syntactic- and sentiment-based features, and most of the word-based features. In this case, each feature represented a dimension of the vector. In case features represented actual words in posts, they were weighted by means of TF-IDF. This kind of representation considers global characteristics of the post, such as the terms in each post and their

[12] First Workshop on Trolling, Aggression and Cyberbullying: `https://sites.google.com/view/trac1/home`

[13] `https://www.hatebase.org/`

[14] `https://spring.me/`

[15] `http://www.cs.cmu.edu/~ark/TweetNLP/`

[16] `https://stanfordnlp.github.io/CoreNLP/`

	Kumar et al.	*Davidson et al.*	*Reynolds et al.*	*Chatzakou et al.*
# of classes	non aggressive, overtly aggressive, covertly aggressive	hate, offensive, neither	strongly, weakly, non aggressive	normal, aggressor, bully, spammer
# of posts	14,984: 6283, 3417, 5284	24,783: 1430, 19190, 4163	12,773: 799, 1224, 10,750	4954: 3562, 59, 24, 1300
average number of words per post	27: 23.83, 32.26, 27.40	16.84: 16.06 16.76 17.49	33.20: 34.29, 32.10, 33.25	17.55: 17.87, 16.42, 16.125, 15.95
average number of nouns per class	4.19, 5.57, 7.75	3.92, 3.90, 3.97	7.95, 7.11, 6.51	5.07, 4.33, 4.58, 3.89
average number of verbs per class	1.15, 1.46, 1.41	0.73, 0.76, 0.62	1.66, 1.44, 1.54	0.45, 0.37, 0.41, 0.30
average number of adverbs per class	1.06, 1.57, 1.46	0.62, 0.72, 0.69	1.47, 1.34, 1.57	0.31, 0.28, 0.25, 0.20
average number of adjectives per class	1.53, 2.24, 1.77	1.03, 0.82, 0.96	1.65, 1.36, 1.42	0.65, 0.64, 0.75, 0.48
average number of punctuation per class	1.29, 1.77, 1.51	0.72, 0.61, 0.84	1.95, 1.80, 1.97	0.56, 0.57, 0.66, 0.39
average number of emoticons-emojis per class	0.05, 0.01, 0.02	0.01, 0.01, 0.02	0.59, 0.73, 0.72	0.07, 0.08, 0.12, 0.02

Table 1: Data Collection Characteristics

TF-IDF	Tokenisation, stopword removal and TF-IDF weighting.	*Stanford Sentiment*	Overall sentiment of the post and sentiment of each detected syntactic structure.
Char	The defined char-based features.	*word2vec*	Matrix representation based on word2vec.
Lemma	Only the lemma of the tokenised terms are kept.	*GloVe*	Matrix representation based on GloVe.
NER	Only the recognised types of entities are kept.	*Barbieri*	Irony detection features based on Barbieri and Saggion [2014].
POS-NVAA	Only noun, verbs, adjectives and adverbs are kept.	*Hernandez*	Irony detection features based on Hernańdez Farías et al. [2016].
POS Tags	Instead of considering the actual terms, it considers their POS tags.	*TF-IDF + SentiWordNet*	TF-IDF + sentiment polarity of the post extracted with SentiWordnet.
POS-NVAA + POS-Frequencies	POS-NVAA + frequency of the different POS tags.		

(a) Simple Feature Sets

TF-IDF + SentiWordNet + Emoji	*TF-IDF + Hernandez*	*TF-IDF + Stemmer + Barbieri*
TF-IDF + Stemmer + Hernandez	*TF-IDF + Barbieri*	word2vec + GloVe
TF-IDF + Stemmer + Char	*TF-IDF + POS Tags*	*TF-IDF + Stemmer + POS Tags*
TF-IDF + Stemmer	*TF-IDF + Char*	*TF-IDF + Char + POS Tags*

(b) Combined Feature Sets

Table 2: Summary of the Evaluated Feature Sets

frequency, but losses local information, such as word order. This type of representation is suitable for any traditional classification technique.

The second strategy involves representing posts in the form of matrices as a sequence of vectors each representing a term according to the selected word embedding model. This kind of representation preserves local information, such as which adjective is modifying which noun. In this case, posts were represented considering the average number of words per post in the collection. For example, for *Kumar et al.*, each post was represented by their last 23 words. Then, each word was replaced by its corresponding 300-dimension vector (as suggested in [Mikolov et al., 2013]), resulting in a matrix representation of posts of dimensionality 23 × 300. The word-vector mapping was performed using a pre-trained word embedding model. Particularly, two commonly used models are considered: word2vec [Mikolov et al., 2013] (trained with Google News data) and GloVe[17] (trained with *Twitter* data). Additionally, the matrix representation also considered the sentiment of words. Each word was associated to the corresponding *WordNet* synset. For each sense associated to the synset, it was retrieved its negative, positive and neutral polarity. Finally, each word was represented by its positive, negative and neutral average polarity and standard deviation.

4.3 Implementation Details

Two experimental methodologies were followed, which were closely related to the feature extraction strategy used. For the vector representation of posts, evaluation was performed using different configurations of three traditional classification algorithms. First, two variations of the SVM, one with a poly kernel and the other with a RBF kernel, both setting $\gamma = 0.1$. Second, Random Forest using 10 and 20 estimators, and third Naïve Bayes. In each case, the implementation provided by Sklearn[18] was used. Finally, the performance of multi-layer perceptrons was also evaluated. The number of hidden layers ranged between 0 and 2 hidden layers. Note that having 0 hidden layers corresponds to performing a logistic regression. Training was performed by means of rmsprop, and loss was analysed in terms of the categorical cross entropy. Hidden layers were activated with the RELU function and had $features/(\#layer + 1)$ neurons. Three normalisation alternatives were applied: no normalisation, feature scaling (minimum and maximum values were computed from the training set) and standardisation.

On the other hand, for the matrix representation, classification was based on recurrent neural networks. Two neural network architectures were evaluated. First, a stacked LSTM network including: a dropout layer with a probability of 0.5, two LSTM layers with 150 and 50 neurons, a RELU layer with $10 \times \#classes$ neurons and finally a softmax activated layer. Second, a hybrid architecture that concatenated the results obtained for word2vec, GloVe and *SentiWordnet* in combination with the first architecture. After concatenation, four layers were added: a RELU layer with $10 \times \#classes$ neurons, dropout with a probability of 0.5, another RELU layer with $10 \times \#classes$ neurons, and finally a softmax layer with $\#classes$ neurons. Neural networks were implemented with Keras[19], using a Theano[20] backend. In all cases, performance was assessed considering the traditional precision and recall metrics, summarised by means of F-Measure.

5 Experimental Results

This section presents the empirical evaluation results of the capabilities of the selected feature sets and their combinations for the task of cyberbullying and cyberaggression detection. This section is organised as follows. Section 5.1 discusses the results obtained for each individual data collection. This means that each evaluation pertains to a single social media site, i.e. to one particular data collection. Then, Section 5.2 presents a cross-social site evaluation. The main goal of this section is to evaluate whether a model trained for a particular social media site is suitable for detecting aggression in another one.

For both empirical evaluations, the statistical significance of performance differences was assessed based on the Wilcoxon test [Corder and Foreman, 2009] for related samples. The test was performed over

[17] https://nlp.stanford.edu/projects/glove/

[18] http://scikit-learn.org/

[19] https://keras.io/

[20] http://deeplearning.net/software/theano/

the results observed for the different feature sets, where samples corresponded to the results obtained for each classification alternative. Two hypothesis were defined. The null hypothesis stated that no difference existed amongst the results of the different samples, i.e. every evaluated feature set performed similarly across the different classification techniques. On the contrary, the alternative hypothesis stated that the differences amongst the results obtained for each feature set were significant and non-incidental.

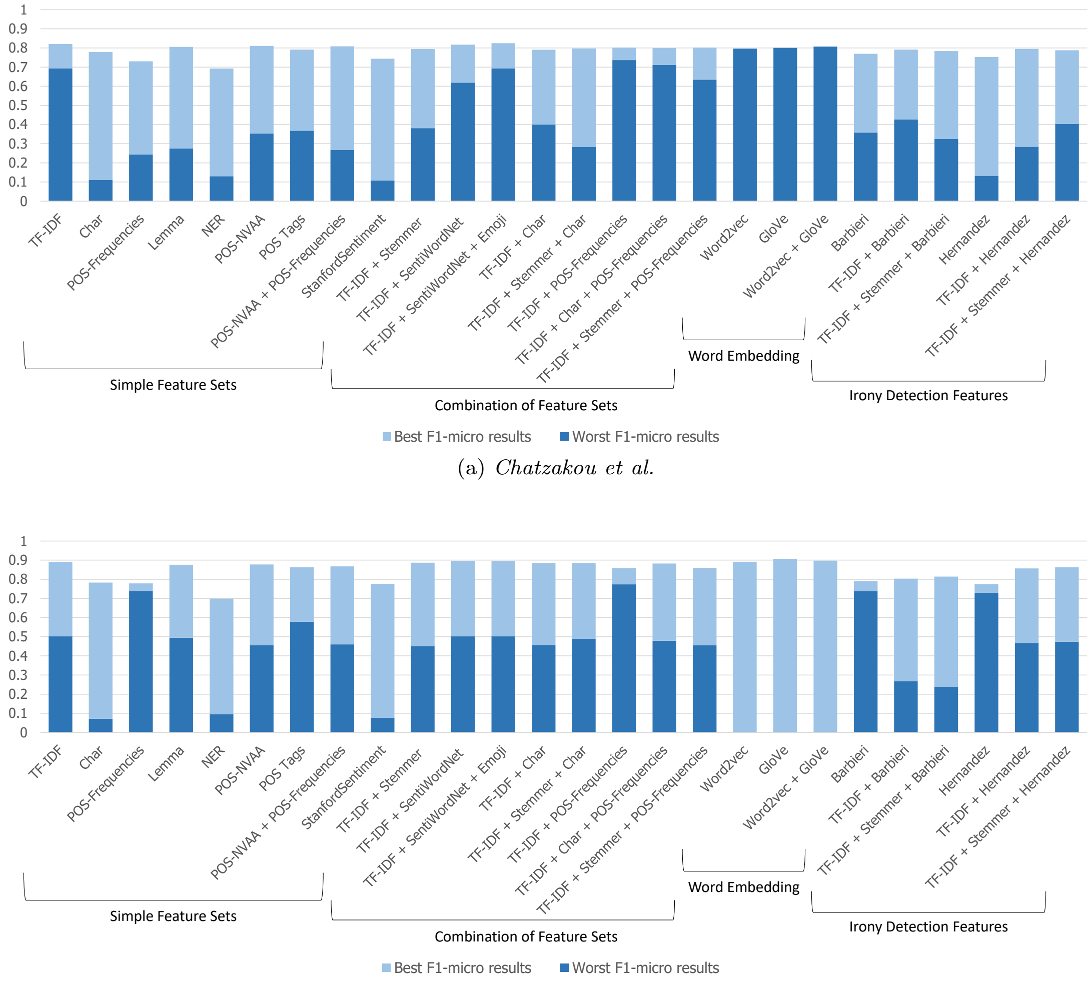
5.1 Single Social Media Site Evaluation

Empirical evaluation consisted of training the classifiers described in Section 4.3 with the selected feature sets described in Section 4.2. The evaluation was independently performed per each of the collections described in Section 4.1. For *Kumar et al.*, the training was performed using the train collection, while testing was performed using the validation dataset. It is worth noting that evaluations considering feature selection techniques were also performed. Particularly, Information Gain was used for retaining the 75% of the most important features. Nonetheless, although feature selection allowed improving results up to a 2%, such differences were statistically insignificant. Hence, those results are not reported.

Figure 1 presents the results obtained for each data collection. Each stacked bar reports the worst and best results obtained for the corresponding feature set. In most cases, the worst results were obtained for Naïve Bayes, followed by SVM with a polynomial kernel, regardless of the data collection under analysis. On the other hand, the best results were mostly obtained with either SVM with a RBF kernel or a neural network with 0 hidden layers. An exception was observed for *Chatzakou et al*, for which the Random Forest technique outperformed the others. This might be related with the fact that, during data collection, authors filtered tweets based on a specific hashtag known to be related to an aggressive and violent controversy. As a result, there was a high predominance of hashtags in the text (more than the 6% of terms were actually hashtags), which could have worked as labels [Huang et al., 2018], despite not being completely accurate. As a result, they could have introduced bias to the classification techniques, especially to decision trees. Interestingly, even though neural networks with hidden layers were the most computationally complex techniques, they did not achieve the best results, probably due to overfitting.

As it can be observed, the results for *Kumar et al.* are lower than those observed for the other collections, regardless of the evaluated feature set. Moreover, in many cases, the worst results observed for *Chatzakou et al., Davidson et al.,* and *Reynolds et al.* are higher than the best results observed for *Kumar et al.* It is worth noting that the results obtained for *Reynolds et al.* and *Chatzakou et al.* are higher than those originally reported in [Reynolds et al., 2011, Chatzakou et al., 2017], reaching the same range of scores than *Davidson et al.* [Davidson et al., 2017]. This highlights the complexity of detecting aggression, and how prediction quality depends not only on the selected features, but also on the intrinsic characteristics of data. For example, despite being written in English, posts in *Kumar et al.* were gathered from Hindi sites. Thereby, they could encompass idiomatic expressions that could differ from those used by Occidental users, or with those presenting a more colloquial usage of English. Additionally, due to cultural differences and that the concept of aggression is subjective, the criteria for defining what is and what is not an aggression could differ, hence it could also occur that posts might have a hidden sense that might not be captured by the English language. Furthermore, word embeddings or corpus-based techniques for extracting features might be biased by the origin of the training data, or by how such training corpus was created.

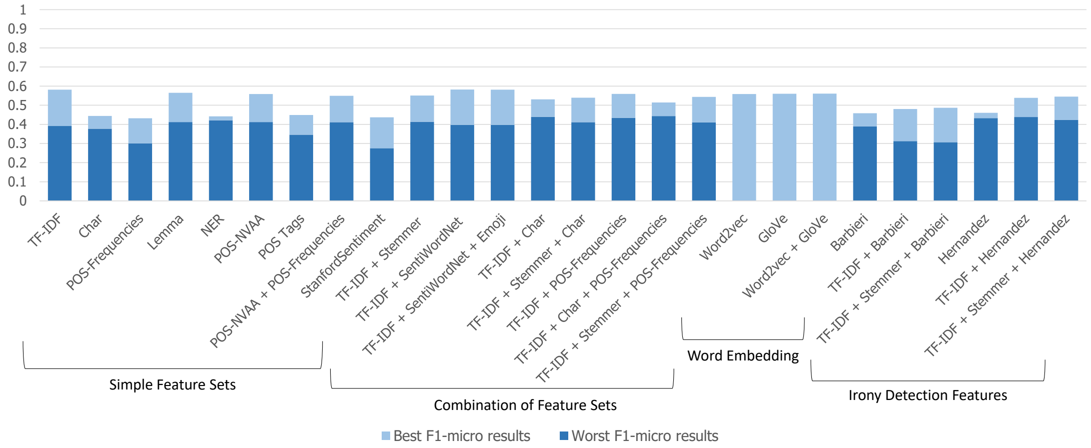
For both *Reynolds et al.* and *Kumar et al*, the best results were obtained when considering *TF-IDF + SentiWordNet*. In the case of *Chatzakou et al.*, the best results were observed for *TF-IDF + SentiWordNet+ Emoji*, whilst for *Davidson et al.*, they were observed for *GloVe*. *StanfordSentiment* consistently obtained the worst results for *Reynolds et al.*, *Chatzakou et al.* and *Davidson et al.* Conversely, in the case of *Kumar et al.*, the worst results were obtained when considering *POS Tags*. It is worth noting that the best and worst results differed at most in a 3%, 13%, 20% and 34% for *Reynolds et al, Chatzakou et al., Davidson et al., and Kumar et al.* respectively. This evidences that features are not the only variable that affects the performance of classifiers. In the case of *Reynolds et al.*, the implications of these observations might be two-fold. First, results could indicate that there is a clear differentiation between the post types, implying that the different feature sets could correctly classify most posts. Nonetheless, as neither precision nor recall were perfect, there also exists a set of posts that is similar to posts belonging to the other categories, thus misguiding the classifier. Second, as results are similar for most of the evaluated



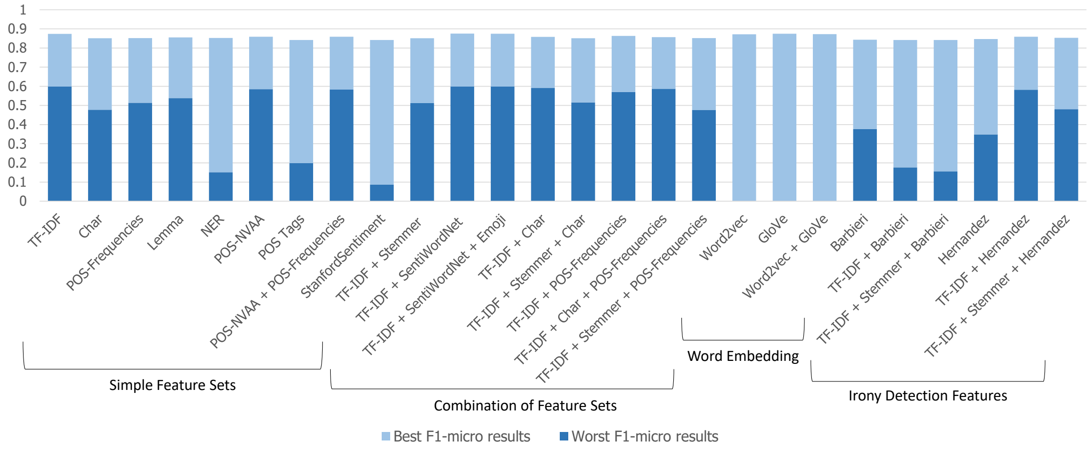
(a) *Chatzakou et al.*

(b) *Davidson et al.*

Figure 1: Aggression Detection Results



(a) *Kumar et al.*



(b) *Reynolds et al.*

Figure 1: Aggression Detection Results (cont.)

feature sets, it might seem that for this data collection, despite providing different characterisations of posts, the diverse feature types might not contribute with new information. Conversely, in the cases of *Chatzakou et al., Davidson et al.,* and *Kumar et al.*, the high variability of results might indicate the difficulties for differentiating similar posts belonging to different classes, and the fact that the different combinations of feature sets provide complementary posts characterisation. For example, the result obtained for *Kumar et al.* when combining *TF-IDF* with sentiment features are higher that those obtained for the individual *TF-IDF* and sentiment features.

As regards the different types of features, their behaviour was similar for all data collections. For example, simply considering the textual features achieved high results in every collection. Feature sets including the *POS* tags or their frequencies did not achieve high results. Similarly, applying lemmatisation did not improve results of applying stemming. Interestingly, representing posts by their word embeddings only improved the results of simply considering the content of posts in one case, and for a marginal difference. Moreover, some features seemed to misguide the classifier, as exposed by the results of *TF-IDF + SentiWordNet + Emoji* that were slightly lower than those observed for *TF-IDF + SentiWordNet.* These results show that adding more features does not necessarily lead to a quality improvement of classifications. Finally, the feature sets for identifying irony were amongst the worst performing ones, which could imply that aggression does not convey irony.

The performed statistical analysis showed in the case of *Kumar et al.* with a confidence of 0.01 that the differences between most pairs of features sets were not statistically different. Nonetheless, statistically significant differences were observed for *Barbieri* and *StanfordSentiment*, which were shown to be statistically lower than feature sets involving *TF-IDF*. Similarly, for *Chatzakou et al.*, there were no statistical significant differences for most cases, with the exception of *TD-IDF* and *StanfordSentiment,* and *TF-IDF + SentiWordNet + Emoji* and *StanfordSentiment.* On the other hand, in the cases of *Davidson et al.* and *Reynolds et al.*, no statistical differences were observed for any of the feature sets. These results imply that more evaluations are needed to truly assess the descriptive power of features, and thus to improve the quality of results.

5.2 Cross Social Media Site Evaluation

Another evaluated scenario was the capabilities of models trained with data belonging to a particular social media site to detect aggression in other social media sites. In particular, the models were trained using a combination of the *Kumar et al.* training and validation collections. For the sake of brevity, this data collection will be called *Kumar et al. - Training.* As described above, this collection was built using English posts gathered from Indian *Facebook* pages. To assess the capabilities of the trained models, three testing collections were used. The first one was *Kumar et al. - Facebook testing.* This collection has similar characteristics to the training one, as it also comprises English posts gathered from Indian *Facebook* pages. The main goal of this evaluation is to act as a baseline for the cross social media aggression detection. The second dataset was *Kumar et al. - Twitter testing.* This collection comprises English post of Indian users gathered from *Twitter.* Note that the cultural context of the published posts in both collections is the same, but the social media and its interaction mechanisms are different. Finally, the third testing collection was *Reynolds et al.*, which comprised posts extracted from the American social network *FormSprin.* As a result, this collection differs from the training one in both cultural context and social network.

As regards the ground truth, all *Kumar et al.* datasets were labelled following the same conventions. As previously described, the classes in this dataset are non-aggressive, covertly aggressive, and overtly aggressive. The main difference between covertly and overtly aggressive is that in the former, aggression's are masked, e.g by irony, whilst the latter has openly aggressive or violent terms. On the other hand, *Reynolds et al.* was labelled following a different strategy. Posts were independently tagged by three Amazon Mechanical Turk workers. In this case, posts were classified as non-aggressive, weakly aggressive, and strongly aggressive. Therefore, there is nodirect correspondence between *Kumar et al.* and *Reynolds et al.* classes. For the purpose of the evaluation, a mapping between the classes in both collections was made, by which the weakly aggressive class was mapped to the covertly aggressive one (CAG), and the strongly aggressive class was mapped to the overtly aggressive one (OAG). Table 3 depicts the class distributions of data collections. In addition to the number of posts per class, the table also presents the

Dataset	*NAG*	*CAG*	*OAG*	*KL divergence*
Kumar et al. - Training	$6,284$	$5,297$	$3,419$	N/A
Kumar et al. - *Facebook* Testing	630	142	144	0.0672
Kumar et al. - *Twitter* Testing	483	413	361	0.0041
Raynolds et al.	$10,837$	$1,160$	776	0.1715

Table 3: Data Collections Class Distribution

Dataset	*Posts*	*Word Stems*	*Shared Word Stems*
Kumar et al. - Training	$15,000$	$17,636$	N/A
Kumar et al. - *Facebook* Testing	916	$3,406$	2694 (79.09%)
Kumar et al. - *Twiteeter* Testing	$1,257$	$2,532$	$1,725$ (68.12%)
Raynolds et al.	$12,773$	$11,416$	$3,966$ (34.74%)

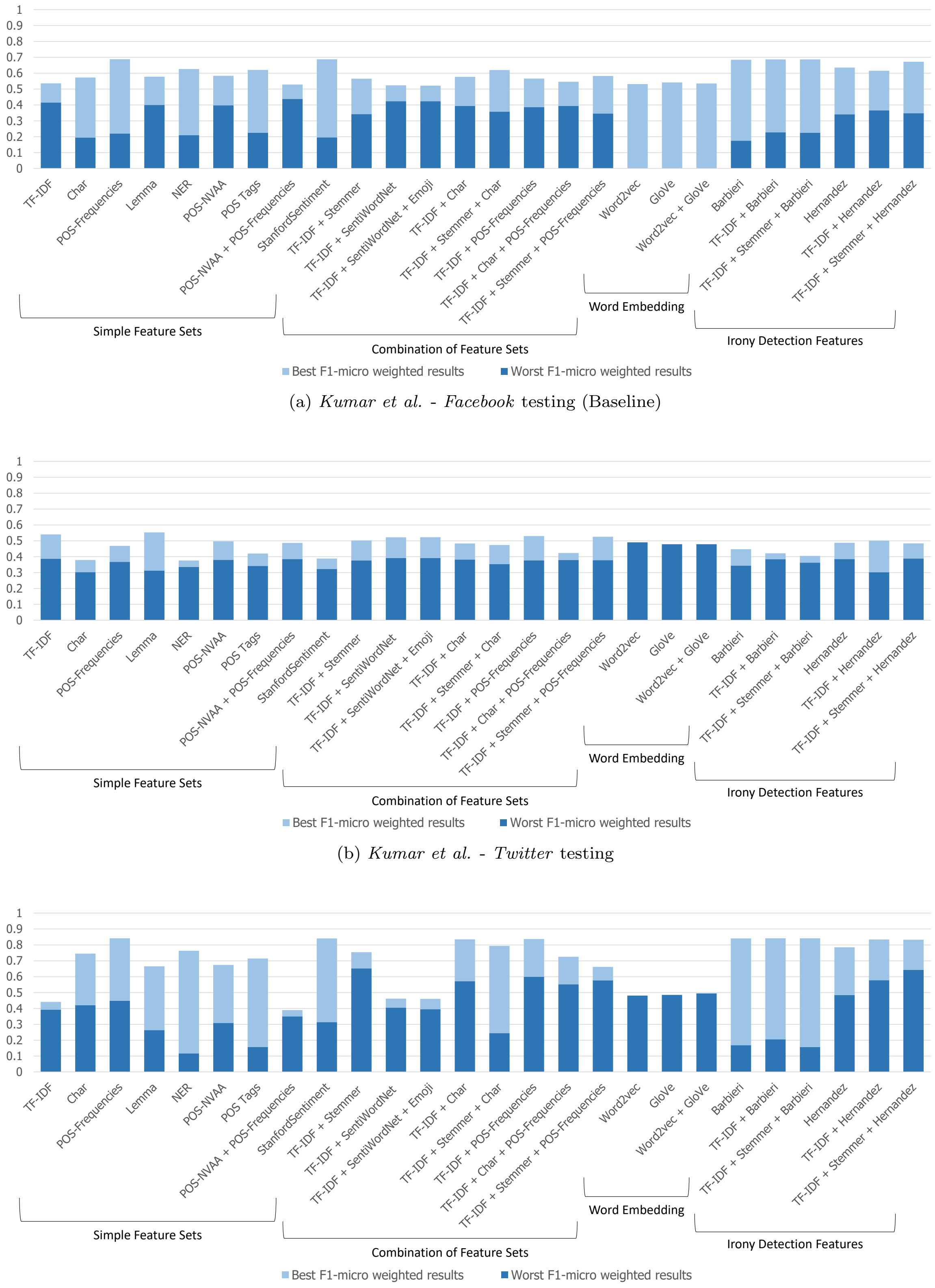
Table 4: Dataset Word Stems

Kullback-Leibler (KL) divergence (i.e. entropy difference) between the class distribution in the test sets against the train set. As it can be observed, *Kumar et al.* - *Twitter* testing is the most similar collection to the training dataset, followed by *Kumar et al.* - *Facebook* testing and *Reynolds et al.*

In this evaluation, models were trained considering the same feature sets previously described. One potential issue of using training and testing collections belonging to different data sources is that they might not share the same terms. Table 4 presents the number of word stems in each collection and how many of them were shared with the training collection. As it can be observed, approximately 79% of the word stems in *Kumar et al.* - *Facebook testing* are in the *Kumar et al.* - *Training.* When considering *Kumar et al.* - *Twitter testing*, only 68% of the word stems are shared with the training dataset. This might be related to the difference on the usage of language in both social media sites. However, further study on this topic is required. Finally, when considering *Reynolds et al.* dataset, even though this collection comprises approximately the same number of posts and terms than *Kumar et al.* - *Training*, the overlapping between them reached only the 35% of word stems.

Figure 2 shows the weighted F-measure according to the inverse class frequency of each instance. The stacked bars represent the worst and best results observed for each feature set. Figure 2a presents the results for *Kumar et al.* - *Facebook testing*, for which the best results were relatively stable across feature sets. Figure 2b depicts the results for *Kumar et al.* - *Twitter testing*, for which the best results obtained for this collection were slightly lower than the ones for *Kumar et al. Facebook* - *testing.* However, the worst results showed a higher estability and spanned over a similar range than the best ones. This indicates that the selected classification technique might have a low effect over the classification results. Finally, Figure 2c presents the results for *Reynolds et al.* In this case, the best and worst results were very dissimilar across feature sets. Nonetheless, in most cases, the best results were higher than the ones observed for *Kumar et al.* - *Facebook testing.*

In the case of *Kumar el al.* - *Facebook* testing, the best results were obtained for *POS-Frequencies*, followed by *StanfordSentiment*, whilst the worst ones were observed for both *TF-IDF + SentiWordNet* and *TF-IDF + SentiWordNet + Emoji.* For *Reynolds et al.* the best results were observed for *TF-IDF + Stemmer + Barbieri* and *TF-IDF + Barbieri,* highlighting the relevance of irony related features for this dataset. On the other hand, the worst results were observed for both *POS-NVAA + POS-Frequencies* and *TF-IDF.* Note that even though combinations including *TF-IDF* achieved the highest results, considering *TF-IDF* alone did not achieve high results, which reinforces the importance of analysing how the different feature types interact. This could be related to the differences in language usage between both datasets. For example, Indians commonly use British spelling, hence it would be more likely to find the spelling "travelled" rather than "traveled". On the other hand, as *FormString* was mostly an American social site, the spelling "traveled" would be more likely to occur. In this context, simply considering the terms in post might not find many matchings, thus affecting the quality of predictions. This mismatching might be overcome by applying stemming or lemmatisation techniques. In the example, both "travelled" and "traveled" are mapped to the same stem "travel", which could help to improve the quality of classifications



(a) *Kumar et al.* - *Facebook* testing (Baseline)

(b) *Kumar et al.* - *Twitter* testing

(c) *Reynolds et al.*

Figure 2: Cross Social Media Site Evaluation Results

Dataset	*Best combination*	*Max. difference best results*	*Avg. Best - Worst*
Kumar et al. - Facebook testing	0.6879	0.1667	0.2807 (±0.1367)
Kumar et al. - Twitter testing	0.5529	0.1767	0.1089 (±0.0490)
Reynolds et al.	0.8416	0.4519	0.3295 (±0.2154)

Table 5: Summary of Results

as shown by the superiority of results of *TF-IDF + Stemmer* over *TF-IDF*. Finally, for *Kumar el al. - Twitter testing*, the best results were observed for both *Lemma* and *TF-IDF,* whilst the worst ones were observed for *NER* and *Char*.

For *Kumar et al. - Facebook testing* results were slightly higher (on average 14%) than the ones obtained for *Kumar et al.* in the individual evaluation. On the other hand, results for *Kumar et al. - Twitter testing* where slightly lower with differences up to a 9%. Despite this performance difference, the best results observed for each feature set were fairly stable, with standard deviations of 9.06% and 10.48% regarding the best results for *Kumar et al. - Facebook* and *Twitter testing* respectively. Similarly, the standard deviation for the previous evaluation was 9.45%. These results might have multiple implications. First, they could indicate that the feature sets have a similar capability for characterising aggression in both training and testing sets. Conversely, they could indicate that the selected feature sets are not able to grasp on the characteristics of data. Finally, the results might imply that even though feature sets aim at describing aggressions from different points of view, feature sets might not actually represent complementary points of view. This situation leads to the necessity of continuing the exploration and analysis of not only the characteristics and capabilities of the selected feature sets, but also the characteristics of the texts being analysed. Unlike the previous evaluation, results observed for *Reynolds et al.* have a greater variance across features. Whilst for the previous evaluation, considering both training and test sets from *Reynolds et al.* yielded a standard deviation of the best results of 0.01 (representing a 1.27% of the best average results), when training with *Kumar et al.* and testing with *Reynolds et al.*, the standard deviation was 0.153 (representing a 22.15% of the best average results). These observations might indicate that feature sets have different capabilities for characterising aggressions across different social media sites.

Another discrepancy with previous results was found in those observed for the different classification techniques. Whilst for *Kumar et al.* the multilayer perceptron with 0 hidden layers and standardised features performed statistically better than the majority of classifiers, for *Reynolds et al.* the Naïve Bayes classifier achieved statistically significant better results than the majority of classifiers. Particularly, Naïve Bayes achieved the best results for some of the low dimensional feature sets, such as *StanfordSentiment* (51 features), *POS Tags* (45 features), or high dimensional feature sets that integrate low dimensional feature sets, such as *TF-IDF + Char + POS-Frequencies* (29,439 for *TF-IDF*, 30 for *char* and 7 for *POS-Frequencies*) or *TF-IDF + Stemmer + Hernandez* (17,636 for *TF-IDF + Stemmer* and 14 for *Hernandez*). Regarding the combinations of low and high dimensional feature sets, the high dimensional feature sets that are mainly based on textual features (such as *TF-IDF + Stemmer*) are highly sparse. For instance, when mapping the *TF-IDF + Stemmer* features defined for *Kumar et al.*, only the 33% of them could be mapped to features in *Reynolds et al.* On the other hand, for feature sets like *Hernandez*, the sparseness was lower, and all features appeared in at least one instance. As a result, for the *TF-IDF + Stemmer + Hernandez* feature set, the *TF-IDF + Stemmer* features could be acting as noise, thus having a low impact on Naïve Bayes [El Hindi, 2014] models, which would be guided by the *Hernandez* features. Conversely, for the *Reynolds et al.* individual evaluation, Naïve Bayes did not perform well for these feature sets. This differences might be caused by changes in the balanced or unbalanced nature of datasets [Frank and Bouckaert, 2006].

6 Conclusions

Cyberbulling and cyberaggression are serious and widespread issues increasingly affecting Internet users. With the "help" of the widespread of social media networks, bullying once limited to particular places or times of the day (e.g. schools), can now occur anytime and anywhere and have a wider range of audience. Cyberaggression can be defined as aggressive online behaviour that intends to cause harm to another person, involving rude, insulting, offensive, teasing or demoralising comments through online social media that target educational qualifications, gender, family or personal habits. This problem is aggravated by the persistence and durability of online materials, which gives these incidents an unprecedented power and influence to affect the lives of billions of people. In this context, cyberaggressions can have deeper and longer-lasting effects in comparison to physical bullying.

Links were found between experiences of cyberbullying and negative outcomes, such as decreased performance at school, dropping out and violent behaviour, in combination with devastating mental and psychological effects such as depression, low self-esteem, and even suicide. Considering the severity of the consequences that cyberaggression has on its victims, there is an imperious need for research aiming at understanding how cyberbullying occurs, in order to prevent it from escalating. Moreover, cyberaggression detection can be used to provide better support and advice for the victims as well as monitoring and tracking the bullies. Other important application of the detection of aggressive content is the detection of cyberextremism, cybercrime and cyberhate propaganda. Given the massive information overload on the Web and the pace at which new posts are being shared, it is unfeasible for human moderators to manually track and flag each insulting and offensive comment. Thereby, it is crucial to develop intelligent techniques to automatically detect harmful content, which would allow the large-scale social media monitoring and early detection of undesired situations.

Despite the seriousness of the problem, there are few successful efforts to detect abusive behaviour on social media data, due to the existence of several challenges, not only related to the nature of posts and the environment in which the aggression occurs, but also to technical limitations of the generalisation and comparability of the proposed approaches. This paper focused on the detection of aggressive content in the context of multiple and heterogeneous social media sites and analysed the capabilities of diverse feature sets for such task. Additionally, it was explored the feasibility of the selected feature sets and techniques for the identification of different types of accounts dedicated to the distribution of aggressive content. Feature sets included char, word and emotional-based features, features used for detecting irony and word-embeddings. The goal was to both provide a comparison between different feature sets and techniques proposed in the literature over the same datasets to analyse the generalisation of results, and to shed some light on the usefulness or adequacy of the different techniques for the task, and the generalisation of models trained for a specific social media site to other sites. Experimental evaluation conducted on four real-world social media dataset showed the difficulties for accurately detecting aggression in social media posts. Moreover, results exposed the limitations of the selected features in relation to the characteristics of the social media sites.

This work also studied the feasibility of using models trained for a specific social media site for classifying the content generated in other social media sites. In this regard, results showed the dependence of the trained models on the characteristics of the users sharing content (e.g. their culture and language usage), and hence the difficulty of generalising models and results to different social media sites. Moreover, another difficulty arose from the fact that there was no pre-defined agreement on the tagging of datasets. This worsens when considering datasets pertaining users belonging to different communities or cultures. This hints the necessity of studying how different demographic groups use, define and express aggressions through language in social media.

Considering the observed results, there still are open issues and challenges that could be tackled in future work. First, given the effect that the intrinsic characteristics of social media sites have on the performance of the detection task, it could be studied how such characteristics impact on each selected feature set. In this regard, it could be also studied how the information belonging to multiple and diverse social media sites could be integrated into a unified model to overcome the observed difficulties. Second, considering that nowadays social media sites are not limited to only textual posts, additional features such as images and videos could also be explored. In addition, it would be interesting to analyse the importance of considering the neighbourhood of users and how they behave to detect aggressions. Finally, given the unbalanced distribution of aggressive and non-aggressive posts, the performance of semi-supervised learning techniques could be studied.

References

S. Agarwal and A. Sureka. Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. *arXiv preprint arXiv:1511.06858*, 2015. 1

M. A. Alvarez-Carmona, E. Guzmán-Falcón, M. Montes y Gómez, H. J. Escalante, L. Villaseñor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican spanish tweets. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 74–96, Sevilla, Spain, 2018. 2

S. Taofeek Aroyehun and A. Gelbukh. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97, Santa Fe, USA, 2018. 2

F. Barbieri and H. Saggion. Modelling irony in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–64, 2014. 3, 2a

U. Bretschneider, T. Wöhner, and R. Peters. Detecting online harassment in social networks. In *International Conference on Information Systems - Building a Better World through Information Systems, ICIS 2014*, 2014. 2

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 (WebSci '17)*, pages 13–22, New York, NY, USA, 2017. ISBN 978-1-4503-4896-6. 1, 2, 4.1, 5.1

V. S. Chavan and S. S S. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2354–2358, Aug 2015. 1, 2

Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80, Sept 2012. 2

G. W. Corder and D. I. Foreman. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach.* John Wiley & Sons, Inc., 2009. 5

M. Dadvar and Franciska M.G. de Jong. Cyberbullying detection; a step toward a safer internet yard. In *Proceedings of the 21st International World Wide Web Conference, WWW 2012 - PhD-Symposium*, pages 121–125, United States, 4 2012. Association for Computing Machinery. ISBN 978-1-4503-1323-0. doi: 10.1145/2187980.2187995. 1

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*, 2017. 4.1, 5.1

Khalil El Hindi. A noise tolerant fine tuning algorithm for the naïve bayesian learning algorithm. *Journal of King Saud University-Computer and Information Sciences*, 26(2):237–246, 2014. 5.2

G. Fahrnberger, D. Nayak, V. S. Martha, and S. Ramaswamy. Safechat: A tool to shield children's communication from explicit messages. In *2014 14th International Conference on Innovations for Community Services (I4CS)*, pages 80–86, June 2014. 2

Eibe Frank and Remco R Bouckaert. Naive bayes for text classification with unbalanced classes. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 503–510. Springer, 2006. 5.2

S. Frenda and S. Banerjee. Deep analysis in aggressive mexican tweets. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 108–113, Sevilla, Spain, 2018. 2

T. Galery and E. Charitos. Aggression identification and multi-lingual word embeddings. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 74–79, Santa Fe, USA, 2018. 2

H. Gómez-Adorno, G. Bel-Enguix, G. Sierra, O. Sánchez, and D. Quezada. A machine learning approach for detecting aggressive tweets in Spanish. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 102–107, Sevilla, Spain, 2018. 2

D. I. Hernańdez Farías, V. Patti, and P. Rosso. Irony detection in Twitter: The role of affective content. *ACM Transaction of Internet Technology*, 16(3):19:1–19:24, July 2016. ISSN 1533-5399. doi: 10.1145/2930663. 3, 2a

S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14(3): 206–221, 2010. 1

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 Tie-Yan Liu, Christie Napa Scollon, and Wenwu Zhu, editors, *Social Informatics*, pages 49–66, Cham, 2015. Springer International Publishing. ISBN 978-3-319-27433-1. 1

M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pages 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. doi: 10.1145/1014052.1014073. 3

H.-H. Huang, C.-C. Chen, and H.-H. Chen. Disambiguating false-alarm hashtag usages in tweets for irony detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 771–777, 2018. 5.1

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 (EACL 2017)*, volume 2, pages 427–431, Valencia, Spain, 2017. 2

P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič. Sentiment of emojis. *PLoS ONE*, 10(12): e0144296, 2015. 3

R. Kumar, A. Kr. Ojha, S. Malmasi, and M. Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, USA, 2018a. 1, 2

R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*, 2018b. 4.1

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*, ICML'14, pages II–1188–II–1196. JMLR.org, 2014. 2

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2, 4.2

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*, pages 145–153, 2016. ISBN 978-1-4503-4143-1. 1, 2

J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001. 2

J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. 2

P. J. C. Pérez, C. J. L. Valdez, M. Ortiz, J. P. S. Barrera, and P. F. Pérez. MISAAC: Instant messaging tool for ciberbullying detection. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, page 1, 2012. 2

F. Ramiandrisoa and J. Mothe. IRIT at TRAC 2018. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 19–27, Santa Fe, USA, 2018. 2

K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In *Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops*, volume 2, pages 241–244, Dec 2011. 4.1, 5.1

J. Risch and R. Krestel. Aggression identification using deep learning and data augmentation. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 150–158, Santa Fe, USA, 2018. 2

A. Roy, P. Kapil, K. Basak, and A. Ekbal. An ensemble approach for aggression identification in English and Hindi text. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 66–73, Santa Fe, USA, 2018. 2

S. Salawu, Y. He, and J. Lumsden. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2017. ISSN 1949-3045. 2

N. Safi Samghabadi, D. Mave, S. Kar, and T. Solorio. RiTUAL-UH at TRAC 2018 shared task: Aggression identification. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 12–18, Santa Fe, USA, 2018. 2

A. Tommasel, J. M. Rodriguez, and D. L. Godoy. Features for detecting aggression in social media: An exploratory study. In *XIX Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 47 (CABA, 2018)*, 2018. 1, 2

C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection and prevention of cyberbullying. In *Proceedings of the International Conference on Human and Social Analytics*, pages 13–18. IARIA, 2015. ISBN 978-1-61208-447-3. 1, 2

E. Whittaker and R. M. Kowalski. Cyberbullying via social media. *Journal of School Violence*, 14(1): 11–29, 2015. 1