Features for Detecting Aggression in Social Media: An Exploratory Study

Antonela Tommasel, Juan Manuel Rodriguez, Daniela Godoy

ISISTAN, CONICET-UNICEN, Argentina

Cyberbullying 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, can now occur anytime and anywhere. Cyberaggression refers to aggressive online behaviour intending to cause harm to another person, involving rude, insulting, offensive, teasing or demoralising comments through online social media. Considering the gravity 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 from escalating. Given the massive information overload on the Web, 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. Considering the challenges posed by the characteristics of social media content and the cyberaggression task, this paper focuses on the detection of aggressive content in the context of multiple social media sites by exploring diverse types of features. Experimental evaluation conducted on two real-world social media dataset showed the difficulty of the task, confirming the limitations of traditionally used features.

1 Introduction

Social networking and micro-blogging sites have increased their popularity in recent years attracting millions of users, who spend on them an increasingly amount of time sharing personal information and interacting with other users. For example, sites like *Flickr*, *YouTube*, *Facebook* or *Twitter* allow users to create content, publish photos, comment on content other users have shared, tag content, and socially connect with other users. In conjunction with the recent growth of social media popularity, other undesirable phenomena and behaviours have appeared. Even though most of the time Internet use could be safe and enjoyable, there are risks involving the online communication 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 [21].

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 [4]. Cyberaggression can be

defined as aggressive online behaviour that intends to cause harm to another person [11], involving rude, insulting, offensive, teasing or demoralising comments through online social media that target educational qualifications, gender, family or personal habits [5]. Cyberbullying is one of the many forms of cyberaggression and is characterised by, an act of online aggression (accentuated by the permanent nature of online posts), the existence of a power imbalance between the individuals involved (including diverse forms, such as physical, social, relational or psychological), and repetitions across time [11].

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 psychological effects such as depression, low self-esteem, and even suicide [11]. 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. While these incidents are still isolated and do not represent the norm, their gravity demand deeper understanding [10]. Additionally, cyberaggressive comments make their targets feel demoralised and frustrated, thus acting as a barrier for participation and socialisation.

Considering the gravity of the consequences that cyberaggression has on its victims and its rapid spread amongst internet users (and specially kids and teens), there is an imperious need for research aiming at understanding how cyberbullying occurs, in order to prevent it from escalating. Other important application of the detection of cyberaggression or aggressive content is the detection of cyberextremism, cybercrime and cyberhate propaganda [1]. Most networking sites today prohibit the usage of offensive and insulting comments [20], which is partially being carried out and filtered to a limited extent. Given the massive information overload on the Web, it is unfeasible for human moderators to manually track and flag each insulting and offensive comments [5]. 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 [4, 16]. First, 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 all racial and minority insults, which might be unacceptable to one group, but acceptable to another one. Fifth,the difficulty of creating a corpus for evaluating the developed techniques.

Considering the existing challenges, this article focuses on the detection of aggressive content in the context of multiple and heterogeneous social media

sites. To that end, several features were extracted from each dataset and evaluated. The remainder of this article is organised as follows. Section 2 describes related work regarding the detection of both aggressive content and bullying accounts. Section 3 describes the information and features that were considered in the analysis. Section 4 describes the experimental settings, including the selected datasets and implementation details. Section 5 analyses the obtained results. Finally, Section 6 presents the conclusions derived from the study, and outlines future lines of research.

2 Related Work

Research into cyberaggression detection has increased in recent years due to its proliferation across social media, and its detrimental effect on people [19]. Cyberaggression or cyberbulling detection can comprise four different tasks [19]: identification of the individual aggressive messages in a social media data stream, assessing the severity of the aggression, identification of the roles of the involved individuals, and the classification of events that occur as a consequence of an aggression incident. Nonetheless, most approaches focus on the first [20, 16, 5] and third [4] tasks.

Regarding the identification of aggressive users, Chatzakou et al. [4] combined user, text, and network-based features in the context of *Twitter*. Experimental evaluation was based on a Twitter dataset related to the #GamerGate controversy 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. Results showed that features did not have the same importance for the task. For example, session statistics, average emotional scores, hate score, average word embedding and community information, amongst other features, were shown to add noise to the classification. According to the authors, their approach achieved an overall precision of 0.89, which decreased to 0.295 and 0.411 for the aggressive and bully classes, exposing the difficulties of the task. The best precision results were achieved for the normal class. Additionally, results also showed that the most effective features for classifying aggressive behaviour were the network-based ones, whilst most text features did not contribute to the improvement of results.

Van Hee et al. [20] explored both the detection of cyberbulling events, and the fine-grained classification of them. The authors considered two types of lexical features (bag-of-words and polarity), which resulted in approximately 300k features. Bag-of-words features included unigrams, bigrams and character trigrams. On the other hand, polarity features included the number of positive, negative and neutral lexicon words averaged over text length, and the overall post polarity. Experimental evaluation was based on approximately 80k Dutch posts belonging to *Ask.fm*. Data was manually labelled into 7 categories (non-aggressive, threat/blackmail, insult, curse, defamation, sexual talk, defence and encouragement to the harasser) with a Kappa score of 0.69. Results achieved

an overall F-Measure of 0.55, whilst the F-Measure for the defamation class was lower than 0.07. The authors hypothesised that the discrepancy of results was due to the extent to which posts in each category are lexicalised. For example, insults are generally highly lexicalised, whereas threats are often expressed in an implicit way.

Nobata et al. [16] aimed at detecting hate speech on 2 million online comments from two domains (Yahoo! Finance and News), which were manually labelled by Yahoo employees. Four types of features were considered: n-grams, linguistic, syntactic and distributional semantics. Linguistic features aim at explicitly look for inflammatory words and elements of non-abusive language such as the usage of politeness words or modal verbs. Distributional semantic features refer to embedding derived features. Pre-processing was applied to comments by normalising numbers, replacing unknown words with the same token, replacing repeated punctuation. Results showed that combining all features achieved the best F-Measure results. Regarding the individual features, the best performing ones were the n-grams for the finance dataset and the embedded features for the news one. The authors hypothesised that the set of selected techniques could achieve good performance in other languages, although it remains to be evaluated.

Similarly to [16], Chavan and S [5] aimed at distinguishing between bullying and non-bullying comments. The selected features included TF-IDF weighted n-grams, the presence of pronouns and skip-grams. Only the 3,000 highest features were selected according to χ^2. Experimental evaluation was based on approximately 6.5k comments from an unspecified site. Posts were pre-processed by removing non-word characters, hyphens and punctuation. Additionally, a spell-checker was applied to correct potential spelling mistakes. Results showed that the best performance was achieved by selecting the pronouns and skip-grams, with differences up to a 4.8% regarding the traditional features.

All previously described approaches are based on supervised models trained, mostly, with SVM, Naïve Bayes or Logistic Regression. Nonetheless, there are also approaches based on lexicons [17, 3] and rules [6, 3]. For example, [17, 3] proposed to detect bullying content by lexically processing text and comparing it against established patterns of aggression. Besides detecting the aggressive content in chat rooms, the goal of [8] was to replace such content by alternatives suitable for minors found in WordNet.

Regarding the rule-based approaches, Chen et al. [6] computed the offensiveness score of sentences based on a set of rules and syntactic features mined using the Stanford parser[1]. Experimental evaluation was based on *YouTube* comments posted by over 2 million of users. According to the authors, their rule-based approach was able to improve the results obtained with SVM and Naïve Bayes classifiers. Bretschneider et al. [3] formulated rules to recognise word patterns indicating relationships between profane words and personal pronouns. Experimental evaluation was based on publicly available *Facebook* posts[2].

[1] `https://nlp.stanford.edu/software/lex-parser.shtml`
[2] `http://www.ub-web.de/research/index.html`

In summary, most works have been based on content, sentiment, user, network-based features or a combination of them. Content-based features include the extractable lexical items of documents, such as keywords, profanity, pronouns, part-of-speech tagging and punctuations symbols. Additionally, lexicon-based features could also be included, such as lists of hate or aggressive words. Sentiment features refer to features that indicate the sentiment of emotion polarity of the content. Generally, texts include certain keywords, phrases or symbols than can be used to determine the sentiments expressed in a document. User-based features represent those characteristics on users' profiles that can be used to judge the role played by such user in a series of online communications. For example, age, gender or sexual orientation. Finally, network-based features refer to metrics that can be extracted from the social networks, including the number of friends, number of followers, frequency of posting and how many times posts were shared.

3 Characterising Aggression

Most approaches have analysed the performance of content, sentiment, user, network-based features or a combination of them for the task of aggression or bullying detection. Nonetheless, with the variability of results and the diversity of social media sites, which have their own intrinsic characteristics, there is still no general consensus regarding which features to select. This study explored different feature sets (and their combinations) including character, word, syntactic, emotion and irony-based features.

Character-based features included the number and ratio of punctuation marks (question marks, exclamation marks, period, commas, ellipses), the number and ratio of upper case letters, and the number and ratio of emoticons/emojis. Syntactic features required the part-of-the-speech (POS) tagging of text, and included the number and ratio of nouns, verbs, adverbs and adjectives. Additionally, other feature sets only considered those words tagged as nouns, verbs, adjectives and adverbs.

Word-based features considered stemming, lemmatisation, name entity recognition, average word length, number of synonyms (analysed considering *WordNet*[3]), commonness of words (analysed by means of the American National Corpus[4]) and frequency of rarest word. Word Embeddings were also considered as features. Particularly, two models were used: word2vec [15] (trained with Google News data) and GloVe[5] (trained with *Twitter* data).

Aggression and cyberbullying detection is difficult due to the subjective nature of bullying. In this regard, sentiment analysis features can contribute to the detection of offensive or abusive content. Sentiment detection includes two aspects. It could refer to either the overall polarity of texts, which are classified as positive, negative and neutral or to specific emotions such as anger,

[3] `https://wordnet.princeton.edu/`
[4] `http://www.anc.org/`
[5] `https://nlp.stanford.edu/projects/glove/`

joy, love or hate, amongst others. Particularly, cyberbulling has been associated to negative emotions, such as anger, irritation, disgust and depression. In this context, features have been defined to consider the overall polarity of posts, the polarity associated to the diverse syntactic structures of posts, the polarity of individual words, the number and ratio of curse words, intensity of posts (total, mean, map, gap). Features were extracted considering two trained models: *StandordNLP* and *SentiWordNet*[6]. Emoticons and, lately, emojis have been considered to convey important sentiment information. In this context, the Emoji Sentiment Ranking [13][7] was included in the analysis by computing the average sentiment polarity of the emojis in posts.

The function of irony is to communicate the opposite of the literal interpretation of the expressions. Moreover, ironic statements can elicit affective reactions [9]. For example, ironic criticism has been recognised as offensive and associated with particular negative affective states, which could enhance the anger, irritation or disgust. As a result, it could be stated that bullying behaviour might be disguised in ironic statements. In this context, the feature sets defined in [2, 9] were considered. Such feature sets focused on the character and word-based features, and emotive word lists and lexicons (AFFIN[8], the lexicon created by [12] and the Whissell's Dictionary of Affect in Language[9]).

4 Experimental Settings

This Section presents the experimental evaluation performed to assess the effectiveness of the selected features for aggression detection, and is organised as follows. Section 4.1 describes the data collection used. Then, Section 4.2 describes the process for extracting the features and creating the posts representations. Finally, Section 4.3 describes implementation details.

4.1 Data Collections Used

The performance of the aggression detection was evaluated considering two datasets that comprise posts belonging to different social media sites. Table 1 summarises the general characteristics of the selected datasets.

Kumar et al. It [14] comprises 15k 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 and opposition groups of recent incidents. Human annotators assigned the posts to one of three classes (overtly aggressive, covertly aggressive and non aggressive). The dataset includes other fine-grained classification, which was not included in the release. According to the authors, the best classification achieved a F-Measure of 0.7. However, the authors did not specify the used features.

[6] http://sentiwordnet.isti.cnr.it/
[7] http://kt.ijs.si/data/Emoji_sentiment_ranking/
[8] http://neuro.imm.dtu.dk/wiki/AFINN
[9] https://www.god-helmet.com/wp/whissel-dictionary-of-affect/index.htm

Reynols et al. It [18] comprises approximately 3k questions and answers extracted from *FormSpring.me*[10]. In such site, users openly invite others to ask and answer questions with the option of anonymity. Posts were manually labelled into three categories (strongly aggressive, weakly aggressive and non-aggressive). According to the authors, the best classification achieved an overall accuracy of 81%, when considering features related to the number of curse words and their intensity.

	Kumar et al.	*Reynols et al.*
# of classes	non aggressive, overtly aggressive, covertly aggressive	strongly, weakly, non aggresive
# of posts	14,984: 6283 - 3417 - 5284	12,773: 799 - 1224 - 10,750
average number of words per post	27: 23.83 - 32.26 - 27.40	33.20: 34.29 - 32.10 - 33.25
average number of nouns per class	4.19 - 5.57 - 7.75	7.95 - 7.11 - 6.51
average number of verbs per class	1.15 - 1.46 - 1.41	1.66 - 1.44 - 1.54
average number of adverbs per class	1.06 - 1.57 - 1.46	1.47 - 1.34 - 1.57
average number of adjectives per class	1.53 - 2.24 - 1.77	1.65 - 1.36 - 1.42
average number of punctuation per class	1.29 - 1.77 - 1.51	1.95 - 1.80 - 1.97
average number of emoticons-emojis per class	0.05 - 0.01 - 0.02	0.59 - 0.73 - 0.72

Table 1. Data Collection Characteristics

Unless datasets were already separated into training and test set, they were randomly split 70% training and 30% test sets.

4.2 Feature Extraction

Before feature extraction, posts were sanitised and pre-processed by removing all non-standard characters, such as non-printable and control characters. Character, word and syntactic-based features required the tokenisation of text, which was performed considering two tools: *twokeniser*[11] (specifically designed for social media) and the *StanfordNLP* library[12]. English stopwords were also removed. Then, considering the features described in Section 3, two strategies were followed for describing posts. The first strategy represent posts as vectors that can be used by any traditional classification technique. The second strategy represents post in the form of matrices, to be used with neural network techniques.

[10] `https://spring.me/`
[11] `http://www.cs.cmu.edu/~ark/TweetNLP/`
[12] `https://stanfordnlp.github.io/CoreNLP/`

The first strategy represents posts considering all character, syntactic and sentiment-based features, and most of the word-based features. In this case, each feature represents a dimension of the vector. In case features represented actual words in posts, they were weighted by means of TF-IDF. On the other hand, the second strategy involves representing posts as a sequence of vectors each representing a term according to the selected word embedding models. In this case, posts were represented considering the average number of words per post in the dataset. For example, for the *Kumar et al.* dataset, each post was represented by their last 23 words. Then, each word is transformed into a vector of dimensionality 300 (as suggested in [15]), resulting in a matrix representation of posts of dimensionality 23×300, for each embedding model. 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

According to the created post representations, two experimental methodologies were followed. For the vector representation of posts, evaluation was based on 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. Evaluation was based on the Sklearn[13] library for Python. Additionally, it was also analysed the performance of multi-layer perceptrons, comprising between 0 and 2 hidden layers. Training was performed by means of rmsprop and loss was analysed by means 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[14], using a Theano[15] backend. In all cases, performance was assessed considering the traditional precision and recall metrics, summarised by means of F-Measure.

[13] `http://scikit-learn.org/`
[14] `https://keras.io/`
[15] `http://deeplearning.net/software/theano/`

5 Experimental Results

Experimental evaluation considered several combination of the features described in Section 3, which are summarised in Table 2. Figure 1 presents the obtained results for both datasets. Each stacked bar reports the worst and best results obtained for the corresponding feature set, for each of the selected datasets. Although results were slightly higher (with differences up to a 2%) when performing feature selection by retaining the 75% of the most important features according to Information Gain, such difference was statistically insignificant. In most cases, the worst results were obtained for Naïve Bayes, followed by SVM with a polynomial kernel, regardless of the analysed dataset. 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. Interestingly, despite being the most computationally complex techniques, neural networks with hidden layers did not achieve the best results.

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 [2].
POS-NVAA	Only noun, verbs, adjectives and adverbs are kept.	*Hernandez*	Irony detection features based on [9].
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.		

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

Table 2. Summary of the Evaluated Feature Sets

As it can be observed, the results for *Kumar et al.* are lower than those for *Reynolds et al.* regardless of the evaluated feature set. Moreover, in most cases, the worst results observed for *Reynolds et al.* are higher than the best results observed for *Kumar et al.*. Interestingly, the results obtained for *Reynolds et al.*

are higher than those originally reported in [18]. This evidences the difficulty of the aggression detection task and how the quality of predictions does not only depend on the selection of features, but also on the intrinsic characteristics of the data under analysis. For example, even though *Kumar et al.* comprises content written in English, it was extracted from Hindi sites, as a result, it could encompass different idiomatic expressions that could differ from those used by Occidental users, or with those presenting a more colloquial usage of English. Additionally, given the cultural differences, the criteria for defining what is an aggression and what it is not could differ, hence it could also occur that posts might have a hidden sense that the English language might not be able to capture.

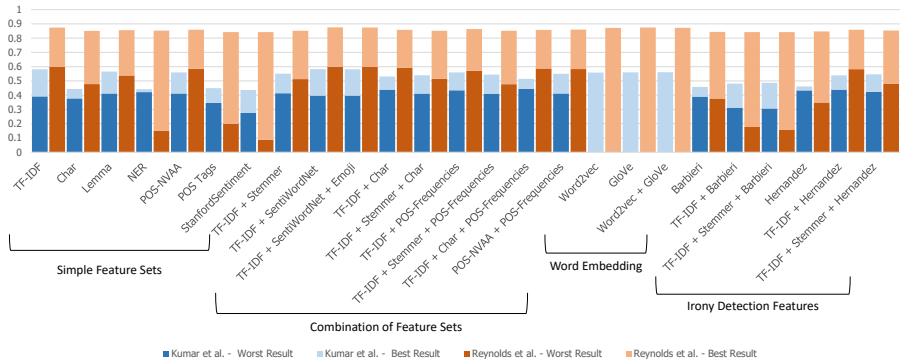

Figure 1. Aggression Detection Results

For both datasets, the best results were obtained when considering *TF-IDF + SentiWordNet*, whilst the worst results were observed for *Stanford Sentiment* and POS Tags, for *Reynolds et al.* and *Kumar et al.*, respectively. It is worth noting that for *Reynolds et al.* results differed at most in a 3%, whilst for *Kumar et al.*, results differed at most in a 34%. This is an interesting phenomenon as it shows the effect that the characteristics of the datasets have on the performance of the selected feature sets. For example, in the case of *Reynolds et al.*, the implications might be two-fold. First, results could indicate that there is a clear differentiation of 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 in the other categories, thus misguiding the classifier. Second, as results are similar for most of the evaluated feature sets, it might seem that for this dataset, despite providing different characterisation of posts, the diverse feature types do not contribute with new information. Conversely, in the case of *Kumar et al.*, the high variability of results might indicate the difficulties for differentiating similar posts belonging to different classes, and the fact the different feature sets provide complementary

characterisation of posts. For example, the result obtained when combining *TF-IDF* with sentiment features is higher that the results obtained for the individual *TF-IDF* and sentiment features.

As regards the different types of features, their behaviour was similar for both datasets. For example, simply considering the textual features achieved high results for both datasets. 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 did not improve the results of simply considering the content of posts. 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 implies a quality improvement of classifications. Finally, the feature sets for identifying irony were amongst the worst performing ones, which might imply that aggression is not implicitly expressed.

Finally, as data was shown not to be normal, a statistical analysis based on the Wilcoxon test [7] for related samples was performed over 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. On the contrary, the alternative hypothesis stated that the differences amongst the results obtained for each feature set were significant and non-incidental. In the case of *Kumar et al.*, for most pairs of feature sets no statistically significant differences were observed with a confidence of 0.01. Nonetheless, statistically significant differences were observed for *Barbieri* and *StanfordSentiment*, which were shown to be statistically lower than feature sets involving *TF-IDF*. On the other hand, in the case of *Reynolds et al.*, no statistical differences were observed for the different 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.

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. 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.

Considering the gravity 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 cyberbully-

ing occurs, in order to prevent it from escalating. Other important application of the detection of cyberaggression or aggressive content is the detection of cyberextremism, cybercrime and cyberhate propaganda. Given the massive information overload on the Web, 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.

This paper focused on the challenges posed by the characteristics of social media content and analysed the capabilities of diverse feature sets for detecting aggression. Feature sets included char, word and emotional-based features, features used for detecting irony and word-embeddings. Experimental evaluation conducted on two 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, as well as the characteristics of the users of thoe sites. In conclusion, results evidenced the necessity of continuing to explore the phenomenon and develop new and more efficient approaches for cyberaggression detection.

Considering the observed results, there still are open issues and challenges that could be tackled in future work . First, the integration of additional features (such as, images, content extracted from images) could be explored. Second, feature selection techniques could be explored to assess the relative importance of each feature, and feature transformation techniques could be explored to discover implicit relations between features, which could increase the descriptive power of the individual features, and thus the quality of classifications. Third, 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. 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. Finally, given the unbalanced distribution of aggressive and non-aggressive posts, semi-supervised learning techniques could be applied.

Bibliography

[1] 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.

[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] 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.

[4] 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. ACM. ISBN 978-1-4503-4896-6.

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

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

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

[8] 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.

[9] D. I. Hernańdez Farías, V. Patti, and P. Rosso. Irony detection in twitter: The role of affective content. *ACM Trans. Internet Technol.*, 16(3):19:1–19:24, July 2016. ISSN 1533-5399. doi: 10.1145/2930663.

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

[11] 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.

[12] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth 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.

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

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

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

[16] 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. International World Wide Web Conferences Steering Committee, 2016. ISBN 978-1-4503-4143-1.

[17] 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. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

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

[19] 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.

[20] 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 Pascal Lorenz and Christian Bourret, editors, *International Conference on Human and Social Analytics, Proceedings*, pages 13–18. IARIA, 2015. ISBN 978-1-61208-447-3.

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