A Comparison of Classification Models to Detect Cyberbullying in the Peruvian Spanish Language on Twitter

Ximena M. Cuzcano[1], Victor H. Ayma[2]
Systems Engineering Department, University of Lima, Lima 15023, Peru

***Abstract*—Cyberbullying is a social problem in which bullies' actions are more harmful than in traditional forms of bullying as they have the power to repeatedly humiliate the victim in front of an entire community through social media. Nowadays, multiple works aim at detecting acts of cyberbullying via the analysis of texts in social media publications written in one or more languages; however, few investigations target the cyberbullying detection in the Spanish language. In this work, we aim to compare four traditional supervised machine learning methods performances in detecting cyberbullying via the identification of four cyberbullying-related categories on Twitter posts written in the Peruvian Spanish language. Specifically, we trained and tested the Naive Bayes, Multinomial Logistic Regression, Support Vector Machines, and Random Forest classifiers upon a manually annotated dataset with the help of human participants. The results indicate that the best performing classifier for the cyberbullying detection task was the Support Vector Machine classifier.**

Keywords*—Cyberbullying detection; machine learning; natural language processing; feature extraction

I. Introduction

Harassment in social networking sites, better known as cyber bullying, has silently impacted many people in recent years. The most prominent acts of virtual harassment occur through rumors, insults, threats, humiliation, and sexual harassment [1]. A survey conducted in 28 countries across the world revealed that 17% of young people experience cyberbullying before the age of 25 [2]. In Europe, 13-15-year-olds are more likely to be bullied online [3]. On the other hand, the Asia-Pacific countries present around 53% of cyberbullying experiences on social networks, followed by the Middle East and Africa with 39% [2]. Regionally in America, 59% of United States adolescents have experienced some form of cyberbullying [4]. Meanwhile, Latin America experiences the highest amount (76%) of cyberbullying on social media platforms [2]. In Peru, a study revealed that at least 58% of kids between 8-12 years old are prone to online harassment [5].

Despite the efforts to prevent cyberbullying events and mitigate its effects [6, 7], the problem coexists with a generation that is always connected to different social media platforms through the Internet, using a computer or mobile phone, where they interact between groups [8]. Moreover, the use of popular social network platforms, such as Twitter, which offer tweet posting anonymity, encourage harassing behaviors with more frequency and cruelty [9], negatively affecting the self-esteem of the victims. Hence, automatic cyberbullying detection becomes important.

Currently, typical cyberbullying detection approaches employ text analysis subtasks such as pre-processing, feature extraction, feature selection, and classification to identify online harassing events. Despite such a well-defined pipeline, there exist very few works in the literature aiming at detecting cyberbullying in textual data from social media written in other languages different from the English language [10-13]. Furthermore, there are a limited number of works trying to solve the automatic cyberbullying detection problem in Spanish languages [14-17].

In this work, we propose to compare four machine learning algorithms for detecting cyberbullying on Twitter textual data written in the Peruvian Spanish language. To reach our goal, we have built an annotated text messages dataset from Twitter written in Peruvian Spanish. The dataset was validated with the help of human participants through an online service specially created to verify and annotate the offensive content according to no harassment, direct harassment, hate speech, and sexual harassment [18,19]. Then, we have used Natural Language Processing (NLP) techniques for pre-processing and subsequent feature extraction. Finally, we have trained and assessed the performances of a Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Multinomial Logistic Regression (MLR) classifiers.

The rest of the paper is structured as follows. The next section presents the related works aiming to automatically detect cyberbullying in social media. Section III describes our methodology to perform the automatic detection of cyberbullying events on Twitter. Section IV provides details of the experimental procedure adopted to perform classifiers' performance comparison as well as presents and discusses the experimental results. Finally, Section V summarizes our research findings along with suggestions for potential future work.

II. Related Works

In recent years, automatic cyberbullying detection in social media has attracted the attention of the scientific community. The early works of Dinakar et al. [19] and Yin et al. [20] demonstrate the researchers' interest in detecting cyberbullying events in social media textual data using supervised machine learning tools. In 2016, Di Capua et al. [21] explored ways of combining semantic, syntactic, sentiment, and social features

within the machine learning pipeline to detect cyberbullying on large data streams from YouTube, Twitter, and Formspring. Chatzakou et al. [22] studied text features, user features, and network-based features to find the set of features that best distinguish bullies and aggressors, thus, detecting bullying and aggressive behavior on Twitter. Later, Davison et al., [23] aimed to identify different types of cyberbullying on Twitter data via a multiclass classifier. Park and Fung [10] employed traditional supervised classifiers and neural network-based models to identify sexist and racist posts on Twitter. Chen et al. [11] aimed to find the best suited supervised classifier at detecting harassment in manually labeled social media comments from Twitter and Facebook. Most recently, Lee et al. [12] investigated the efficacy of traditional machine learning and neural networks-based models at detecting abusive language on a Twitter dataset. Hani et al. [13] extended the work of Reynolds et al. [24] at detecting cyberbullying events in text messages from Formspring.me by introducing a set of new classifiers. Such a group of works exposes the scientific efforts made to detect cyberbullying from textual data written in the English language.

However, the cyberbullying issue is common across countries and languages. In this sense, Ptaszynski et al. [25] developed a systematic approach upon machine learning techniques to automatically detect cyberbullying entries in the Japanese language. Van Hee et al. [26] trained an SVM classifier in a Dutch text messages dataset collected from Ask.fm social network to identify seven cyberbullying-related categories, thus, detecting cyberbullying events. Similarly, Del Vigna et al. [27] assessed the SVM and a neural network-based classifier on the task of hate speech recognition upon a manually annotated Italian corpus of Facebook. Özel et al. [28] considered a feature selection stage within the machine learning pipeline, to detect cyberbullying in Turkish text messages using labeled data from Instagram and Twitter. Haidar et al. [29] presented a machine learning-based approach to detect cyberbullying in the Arabic language from Twitter textual data collected across the Middle East Region countries. Furthermore, Mouheb et al. [30] presented a real-time cyberbullying detection system in Twitter streams that classify bullying messages according to the offensive strength. On the other hand, Bai et al. [31] focused on detecting offensive speech in German social media through a binary classification scheme that considers traditional supervised classifiers and neural networks models. Most recently, Nurrahmi and Nurjanah [32] employed text processing and machine learning techniques to detect bullies from the automatic analysis of Twitter posts written in the Indonesian language. Also, Febriana and Budiarto [33] constructed a dataset of Twitter posts collected during the presidential election period in Indonesia to promote the detection of hateful speech and tested its usefulness by submitting it to a basic sentiment analysis model. Win [34] used the SVM algorithm on a set of textual data collected from Facebook in the Myanmar language to discriminate bullying messages.

In a different direction, some authors have addressed the cyberbullying detection task through the implementation of multilanguage cyberbullying detection platforms. For instance, Unsvåg and Gambäck [35] conducted experiments on Twitter text messages written in English, Portuguese, and German languages to measure the effects of including Twitter user's features on the hate speech classification task. The authors observed that tweets with similar content written in different languages hinder the classifiers' performances. Pawar and Raje [36] modeled linguistic patterns upon a hand-labeled bilingual (Hindi and Marathi languages) dataset using Machine Learning and Natural Language Processing techniques to detect cyberbullying in Twitter and Internet forums. Moreover, Steimel et al. [37] experimented with a general cyberbullying detection model across multiple languages (English and German) with data collected from Twitter. Their findings showed that multilingual classifier optimization is not possible even in environments that use comparable datasets.

Despite the efforts to tackle cyberbullying detection in social media, the works aiming at detecting offensive behavior in the Spanish language are yet scarce. For instance, Gómez-Adorno et al. [14] addressed the detection task as a binary classification problem, employing supervised Machine Learning models to detect aggressive tweets, a cyberbullying-related topic, in a Mexican-Spanish language dataset proposed in the 2018 edition of MEX-A3T contest. Similarly, Molina-Gonzáles et al. [15] proposed an ensemble of supervised classifiers to identify offensive messages on the 2019 edition of MEX-A3T. Gutiérrez-Esparza et al. [16] developed a classification model to detect cyberbullying events (i.e., racism, violence based on sexual orientation, and violence against women) on a Mexican-Spanish textual dataset collected from Facebook. The authors highlight the participation of school professors and psychologists, with experience in evaluation and intervention in cases of bullying, during the annotation process. Finally, in a more recent study, López-Martínez et al. [17] proposed an online-tool capable of detecting cyberbullying from tweets written in Spanish. The authors combined Open Source Intelligence tools with Natural Language Processing techniques to compile information from the victim's Twitter account and analyzed tweets from every follower.

III. Methodology

Currently, there exist several works focused on detecting cyberbullying in social media. However, the vast majority focuses on text analysis in the English language due to the availability of resources for text analysis, including textual datasets. Such a lack of works aiming for cyberbullying detection in other languages is primarily due to language variants and its grammar complexity. Language variants are specific to a region and vary according to demographic and social factors, such as the appearance of words according to the dialect, idioms, and colloquialisms [16,38]. Language grammar complexity, on the other hand, is attributed to morphology and syntax rules, such as gender and number derivations, verb conjugations, enclitic forms, superlatives, and diminutives suffixes, among others [39]. Therefore, it is paramount to consider both aspects when acquiring textual data intended to model cyberbullying in social media.

In this work, we propose the automatic detection of cyberbullying through the identification of its four categories in an analysis of Spanish tweets collected from Twitter users

resident in Peru. Our method combines Natural Language Processing (NLP) and Machine Learning (ML) techniques to establish a correspondence between the users' tweets and the types of cyberbullying, namely, no harassment, direct harassment, hate speech, and sexual harassment [18,19]. A class label is assigned to a tweet according to the conventional four-stage classification scheme, as shown in Fig.1 the Dataset Collection stage gathers a set of tweets from Peruvian Twitter users; the Pre-Processing stage improves the data quality by removing inconsistencies from the tweets; the Feature Extraction stage obtains a compact representation (x) of a tweet; finally, the Model Selection Stage choose the best-suited classifier to solve the automatic cyberbullying detection problem via a classifiers' performance comparison.

A. Dataset Collection

In this work, we have constructed and made publicly available[1] a dataset consisting of a collection of 10,096 tweets in Spanish from comments and interactions between Peruvian Twitter users with the help of the Streaming API[2] tool. We collected the dataset during August 2019 and January 2020 from users with an age range between 14 and 60 years old. To ensure class discriminability among tweets, we included common words, jargons relative to Peruvian people, and offensive words during the tweet retrieval process. Furthermore, we have added a geographical delimitation filter after the tweets retrieval process to ensure that the collected tweets belong to Peruvian users only. The filter is part of the Streaming API tool, which is composed of delimiting quadrants with the latitude and longitude coordinates of the different regions of Peru.

The collected tweets were labeled with the help of human participants, who were mostly undergraduate students from the last year of Psychology, Communications, and Law schools from different universities in Peru. The participants evaluated a set of twenty randomly selected tweets via a website specially created to guarantee anonymous sessions not to reveal the participant' identities. In one session, a participant assigns a class label to each tweet from the set of twenty tweets according to the four cyberbullying categories. Moreover, we made cyberbullying categories definitions available throughout the labeling process, and we also ensured that a tweet gets evaluated by at least three different participants to avoid labeling conflicts [40].

Finally, after applying the region based filtering and tweet labeling processes, we obtained a dataset comprised of 10,096 tweets, which class distribution corresponds to 5122, 2127, 1000, and 1847 observations for the no harassment, direct harassment, hate speech, and sexual harassment, respectively.

B. Pre-Processing

In this stage, we performed a set of transformations over the original tweets in the dataset to enhance data quality and facilitate its processing for further analysis. In this sense, we first removed symbols, hash tags, mentions, digits, emoticons, and web links from the dataset. Then, we eliminated repetitive characters, using regular expressions, to correct spelling errors except for the consecutive characters r, l, c, and e, because they represent single sound letters, e.g., "aburrido", "llamada", "acción", "reenviar". Then, we converted all the tweets to lowercase to standardize the data. After that, we applied a word tokenization technique overall the tweets to translate the Peruvian jargon to words with the closest meaning in the Spanish dictionary, e.g., "yapa" to "extra" or "monse" to "aburrido". Finally, we eliminated the stopwords, such as y, a, pero, que, tu, among others, because they often are irrelevant to the tweets analysis in further steps.

C. Feature Extraction

The feature extraction stage aims at establishing relationships between words in a tweet that might help discriminate the intent of abuse. Therefore, here, we used a set of techniques oriented to the semantic and syntactic analysis among words, whose objectives are to relate groups of words to establish the intention and context in which they were used. To perform the semantic analysis, we used stemming and lemmatization techniques implemented with a neutral Spanish dictionary in the Snowball Stemmer [3] and Spacy [4] tools, respectively. On the other hand, we based the syntactic analysis on the n-gram technique, specifically in its bi-gram and tri-gram variants, using the nltk[5] library. It is worth mentioning that we applied these techniques before the stopwords removal in the pre-processing stage to maintain the context of the message, e.g., "no eres tonto" is different from "eres tonto". Subsequently, we used the TF-IDF statistical measure to obtain numerical representations of the tweets and the frequency of their words, allowing us to know the degree of importance of a feature. Specifically, we complemented the stemming and lemmatization semantic representation techniques with the TF-IDF technique, and the bi-grams and tri-grams syntactic feature extraction techniques with the TF-IDF method.

D. Model Selection

The model selection stage's purpose is to select the best-suited classifier in detecting the four types of cyberbullying from tweets posted in the Peruvian Spanish language. Hence, we conducted a performance comparison among the most common supervised algorithms for text classification problems. Specifically, we trained a Naive Bayes (NB), Multinomial Logistic Regression (MLR), and Random Forest (RF) classifiers, which are suitable when working with a large number of features [41-43]. We also trained a Support Vector Machine (SVM) classifier, which has proven to behave well in text classification tasks with small class samples [44]. Such models were implemented using the Scikit-Learn[6] library for Python and were set to work upon their by default parameters.

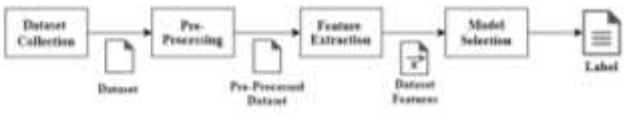

Fig. 1. Overview of our Methodology.

[1] Available in: https://github.com/ximenamar/sp_tweets_cyberbullying
[2] Available in: https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter
[3] Available in: https://snowballstem.org/download.html
[4] Available in: https://spacy.io/api/lemmatizer
[5] Available in: http://www.nltk.org/api/nltk.html?highlight=ngram
[6] Available in: https://scikit-learn.org/stable/

IV. Result Analysis

In order to assess the performances of the four classification algorithms on the cyberbullying detection task over Twitter textual data written in Peruvian Spanish language, we performed a dataset partitioning into a training and testing sets according to a 70% and 30% proportions, respectively. Moreover, we include data under-sampling scheme in our experiments to examine whether the data balancing improves the classifiers' performances. Specifically, we randomly selected data from the majority classes to compensate for such imbalance. In this way, we evaluate the classifiers' performances based on 10-fold cross-validation procedure over two datasets: an imbalanced dataset, which maintains the original class distribution, and a balanced dataset, which contains approximately four thousand observations equally distributed among the classes. Finally, we assessed the classifiers based on the average of the accuracy, precision, recall, and F1-Score performance metrics. Next, we report and discuss the results obtained from such experimental procedure.

A. *Classifiers' Assessment on the Imbalanced Dataset*

TABLE Isummarizes the classifiers' performance scores on detecting the cyberbullying in an imbalanced dataset. The performance metrics correspond to the average and the standard deviation (in parentheses below the average score) of the accuracy, precision, recall, and F1-score, respectively, for the semantic (Stemming and Lemmatization) and syntactic (Bi-grams and Tri-grams) data representations schemes combined with the TF-IDF.

In general, the results indicate that the classifiers using the semantic schemes to represent the textual data performed significantly better compared to their syntactic-based counterparts. We attribute this behavior to Spanish language properties, such as the use of proper nouns next to potentially relevant words. While semantic schemes for textual data representation consider the relevance of a word via its occurrence throughout the dataset, the syntactic schemes ponder the appearance of compositions of words, reducing their representatively in the dataset.

Further analysis of the classifiers' performances based on semantic schemes reveals that the stemming-based classifiers performed slightly better than lemmatization-based classifiers; these differences in the results are due to the feature extraction techniques principles. Whereas stemming removes affixes and suffixes to obtain word roots, lemmatization transforms words into their dictionary form, which turns the classification of textual data a challenging task, especially in languages with complex morphology [45].

TABLE I. Classifiers' Performance Metrics on a Imbalanced Dataset

Performance Metrics	**Models**	**Feature Extraction Schemes**			
		Stemming & TF-IDF	***Lemmatization & TF-IDF***	***Bi-grams & TF-IDF***	***Tri-grams & TF-IDF***
Accuracy	NB	0.674 (+/-0.003)	0.667 (+/-0.006)	0.630 (+/-0.006)	0.622 (+/-0.006)
	RF	**0.797** (+/-0.002)	0.792 (+/-0.007)	0.751 (+/-0.005)	**0.722** (+/-0.001)
	SVM	0.792 (+/-0.003)	**0.793** (+/-0.007)	**0.753** (+/-0.007)	0.710 (+/-0.004)
	MLR	0.764 (+/-0.007)	0.750 (+/-0.008)	0.678 (+/-0.007)	0.628 (+/-0.005)
Precision	NB	**0.834** (+/-0.139)	**0.838** (+/-0.148)	0.626 (+/-0.386)	0.629 (+/-0.392)
	RF	0.814 (+/-0.111)	0.819 (+/-0.097)	0.820 (+/-0.119)	0.850 (+/-0.130)
	SVM	0.801 (+/-0.087)	0.817 (+/-0.100)	**0.838** (+/-0.109)	**0.871** (+/-0.124)
	MLR	0.822 (+/-0.111)	0.818 (+/-0.117)	0.835 (+/-0.153)	0.855 (+/-0.156)
Recall	NB	0.370 (+/-0.361)	0.348 (+/-0.369)	0.281 (+/-0.415)	0.272 (+/-0.419)
	RF	0.712 (+/-0.155)	0.694 (+/-0.181)	0.571 (+/-0.285)	**0.495** (+/-0.322)
	SVM	**0.715** (+/-0.169)	**0.710** (+/-0.173)	**0.581** (+/-0.281)	0.491 (+/-0.330)
	MLR	0.603 (+/-0.222)	0.598 (+/-0.255)	0.382 (+/-0.362)	0.277 (+/-0.415)
F1-Score	NB	0.402 (+/-0.234)	0.368 (+/-0.245)	0.252 (+/-0.304)	0.232 (+/-0.3021)
	RF	**0.752** (+/-0.113)	0.739 (+/-0.119)	0.627 (+/-0.191)	**0.552** (+/-0.225)
	SVM	0.748 (+/-0.119)	**0.750** (+/-0.120)	**0.642** (+/-0.186)	0.545 (+/-0.236)
	MLR	0.669 (+/-0.134)	0.664 (+/-0.156)	0.410 (+/-0.263)	0.245 (+/-0.303)

Despite the unbalanced characteristic of the dataset, a classifier-based analysis exhibits the SVM and RF models as the best two performing classifiers, with small differences in scores for both classifiers. Regarding the average scores to all the metrics, the SVM obtained the best scores in most of the evaluated cases, whereas the RF obtained the lowest standard deviation values. We attribute these behaviors to the classifiers' training characteristics. On the one hand, the SVM classifier defines a decision surface based on the most representative samples within the training set, which battles the imbalance. On the other hand, the RF classifier bootstrap characteristic randomly selects a subset of training samples to build a tree within the forest; however, the training subsets are majority different among trees in the forest, thus overcoming the imbalance on the dataset.

B. *Classifiers' Assessment on the Balanced Dataset*

Similar to TABLE I, 0presents the classifiers' performance scores obtained from their execution on a balanced dataset. The results reinforce the classifiers' performance behavior elicited from the feature-based analysis on an imbalanced dataset.

In a classifier-based analysis, however, the results show that in general, the SVM classifier performed better than the rest of classifiers, closely followed by the RF classifier. In a classifier-based analysis, however, the results show that in general, the SVM classifier performed better than the rest of classifiers, closely followed by the RF classifier. We believe that this is due to the linear kernel used during the SVM training, which makes the SVM performs better in tasks with high-dimensional feature spaces [46], such text classification for cyber bullying detection.

TABLE II. CLASSIFIERS' PERFORMANCE METRICS ON A BALANCED DATASET

Performance Metrics	**Models**	**Feature Extraction Schemes**			
		Stemming & TFIDF	***Lemmatization & TFIDF***	***Bi-grams & TFIDF***	***Tri-grams & TFIDF***
Accuracy	NB	0.761 (+/-0.008)	0.763 (+/-0.008)	0.651 (+/-0.006)	0.467 (+/-0.020)
	RF	0.797 (+/-0.009)	0.786 (+/-0.008)	0.621 (+/-0.013)	0.539 (+/-0.009)
	SVM	**0.805** (+/-0.007)	**0.805** (+/-0.009)	**0.689** (+/-0.013)	**0.612** (+/-0.015)
	MLR	0.795 (+/-0.010)	0.791 (+/-0.008)	0.672 (+/-0.009)	0.609 (+/-0.015)
Precision	NB	0.760 (+/-0.110)	0.750 (+/-0.020)	0.648 (+/-0.038)	0.625 (+/-0.179)
	RF	**0.815** (+/-0.088)	**0.808** (+/-0.109)	0.680 (+/-0.179)	0.673 (+/-0.213)
	SVM	0.811 (+/-0.072)	0.802 (+/-0.081)	0.675 (+/-0.116)	0.665 (+/-0.187)
	MLR	0.809 (+/-0.091)	0.795 (+/-0.083)	**0.691** (+/-0.137)	**0.687** (+/-0.193)
Recall	NB	0.766 (+/-0.117)	0.756 (+/-0.098)	0.649 (+/-0.168)	0.476 (+/-0.316)
	RF	0.799 (+/-0.103)	0.781 (+/-0.125)	0.627 (+/-0.113)	0.528 (+/-0.261)
	SVM	**0.806** (+/-0.088)	**0.796** (+/-0.074)	**0.661** (+/-0.122)	**0.570** (+/-0.187)
	MLR	0.794 (+/-0.091)	0.783 (+/-0.107)	0.659 (+/-0.151)	0.564 (+/-0.216)
F1-Score	NB	0.758 (+/-0.058)	0.748 (+/-0.065)	0.634 (+/-0.062)	0.428 (+/-0.136)
	RF	0.798 (+/-0.053)	0.780 (+/-0.061)	0.636 (+/-0.114)	0.522 (+/-0.163)
	SVM	**0.805** (+/-0.064)	**0.796** (+/-0.070)	**0.662** (+/-0.104)	**0.575** (+/-0.119)
	MLR	0.795 (+/-0.056)	0.782 (+/-0.065)	0.659 (+/-0.113)	0.568 (+/-0.117)

VI. Conclusions and Future Work

In this work, we have proposed a machine learning classifiers' comparison to detect cyberbullying on Twitter posts written in the Peruvian Spanish language. The classifiers were trained upon a set of text messages collected from Twitter users resident in Peru. Moreover, the dataset content was validated by matter-related participants, i.e., psychologists, sociologists, among others, through a web application. We conducted experiments over imbalanced and balanced versions of the dataset using feature extraction schemes, which involve the combination of semantic and syntactic techniques from the Natural Language Processing field.

The experimental analysis demonstrated that semantic-based schemes for text representation are better than syntactic-based schemes. Moreover, classifiers working upon stemming features showed superior from those using lemmatization features. Furthermore, the Support Vector Machine classifier has shown a consistent performance among the feature extraction schemes despite the different performances showed by the classifiers in both datasets, obtaining superior results in the balanced dataset.

In our experiments, we relied on a pre-processing scheme based on traditional text processing techniques, such as the removal of repetitive characters, emoticons, stop words, and so on, to easy the classifiers' training. However, it would be interesting to assess the classifiers' performances over tweets that include emoticon characters as they are often used to reinforce emotions in text messages.

Finally, in this work, we have translated common jargons in Peruvian Spanish language to their dictionary equivalent, so to be part of the training process. However, it would be interesting to include jargon into a pre-defined Spanish language lexicon and assess the classifiers' performances

References

[1] J. L. Pedreira Massa and H. S. Basile, "El acoso moral entre pares (bullying)," Construção psicopedagógica, 2011, vol. 19, no. 19, pp. 8–33.

[2] M. Newall, "Global views of cyberbullying," 2018. [Online]. Available: https://www.ipsos.com/en/global-views-cyberbullying.

[3] V. Dalla Pozza, A. Di Pietro, S. Morel, and E. Psaila, "Cyberbullying among young people," 2016. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571367/IPOL_STU(2016)571367_EN.pdf.

[4] M. Anderson, "A majority of teens have experienced some form of cyberbullying," 2018. [Online]. Available: https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/.

[5] DQ Insitute, "The 2018 DQ impact report," 2018. [Online]. Available: https://www.dqinstitute.org/2018DQ_Impact_Report/#Cyber-Pandemic.

[6] J. Vitak, K. Chadha, L. Steiner, and Z. Ashktorab, "Identifying women's experiences with and strategies for mitigating negative effects of online harassment," in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17), 2017, pp. 1231–1245.

[7] M. Fan, L. Yu, and L. Bowler, "Feelbook: A social media app for teens designed to foster positive online behavior and prevent cyberbullying," in Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16), 2016, pp. 1187–1192.

[8] B. Belsey, "Cyberbullying: An emerging threat to the "always on" generation," 2006. [Online]. Available: http://www.billbelsey.com/?p=1827.

[9] J. Suler, "The online disinhibition effect," Cyberpsychology & behavior, 2004, vol. 7, no. 3, pp. 321–326.

[10] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on twitter," in Proceedings of the First Workshop on Abusive Language Online, 2017, pp. 41–45.

[11] J. Chen, S. Yan, and K.-C. Wong, "Aggressivity detection on social network comments," in Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI '17), 2017, pp. 103–107.

[12] Y. Lee, S. Yoon, and K. Jung, "Comparative studies of detecting abusive language on twitter," in Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 101–106.

[13] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning," International Journal of Advanced Computer Science and Applications, 2019, pp. 703–707.

[14] H. Gómez-Adorno, G. B. Enguix, G. E. Sierra, O. Sánchez, and D. Quezada, "A machine learning approach for detecting aggressive tweets in spanish," in Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), 2018, pp. 102–107.

[15] M. D. Molina-Gonzáles, F. M. Plaza-del-Arco, M. T. Martín-Valdivia, and L. A. Ureña-López, "Ensemble learning to detect aggressiveness in mexican spanish tweets," in Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), 2019, pp. 495–501.

[16] G. O. Gutiérrez-Esparza, M. Vallejo-Allende, and J. Hernández-Torruco, "Classification of cyber-aggression cases applying machine learning," Applied Science, 2019, vol. 9, no. 9, Article ID 1828.

[17] A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, "CyberDect. A novel approach for cyberbullying detection on twitter," International Conference on Technologies and Innovation, Spring, Cham, 2019, pp. 109–121.

[18] J. Golbeck, Z. Ashktorab R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, Q. Gergory, R. K. Gnanasekaran, et al., "A large labeled corpus for online harassment research," in Proceedings of the 2017 ACM on Web Science Conference (WebSci), 2017, pp. 229–233.

[19] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," Proc. Fifth International AAAI Conference on Weblogs and Social Media (SWM' 11), 2011, pp. 11–17.

[20] D. Yin, Z. Xue, L. Hong, and B. Davison, "Detection of harassment on Web 2.0," in Proceedings of the Content Analysis in the WEB 2.0, 2009, vol. 2, pp. 1–7.

[21] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 432–437.

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

[23] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017), 2017, pp. 1–4.

[24] K. Reynolds, A. Kontostathis and L. Edwards, "Using machine learning to detect cyberbullying," 2011 10th International Conference on Machine Learning and Applications and Workshops, 2011, pp. 241–244.

[25] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, and K. Araki, "Machine learning and affect analysis against cyber-bullying," in Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents, 2010, pp. 7–16.

[26] C. Van Hee, B. Verhoeven, E. Lefever, G. D. Pauw, W. Daelemans, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," International Conference Recent Advances in Natural Language Processing (RANLP), 2015, pp. 672–680.

[27] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me hate me not: Hate speech detection on facebook," in Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), 2017, pp. 86–95.

[28] S. A. Özel, E. Saraç, S. Akdemir and H. Aksu, "Detection of cyberbullying on social media messages in Turkish," 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 366–370.

[29] B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content," 2017 1st Cyber Security in Networking Conference (CSNet), 2017, pp. 1–8.

[30] D. Mouheb, M. H. Abushamleh, M. H. Abushamleh, Z. A. Aghbari and I. Kamel, "Real-time detection of cyberbullying in Arabic twitter streams," 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 2019, pp. 1–5.

[31] X. Bai, F. Merenda, C. Zaghi, T. Caselli, and M. Nissim, "Rug at germeval: Detecting offensive speech in German social media," in Proceedings of the GermEval 2018 Workshop, 2018, pp. 63–70.

[32] H. Nurrahmi and D. Nurjanah, "Indonesian twitter cyberbullying detection using text classification and user credibility," 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 543–548.

[33] T. Febriana and A. Budiarto, "Twitter dataset for hate speech and cyberbullying detection in indonesian language," 2019 International Conference on Information Management and Technology (ICIMTech), 2019, pp. 379–382.

[34] Y. Win, "Classification using support vector machine to detect cyberbullying in social media for Myanmar language," 2019 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), 2019, pp. 122–125.

[35] F. Unsvåg and E. Gambäck, "The effects of user features on twitter hate speech detection," in Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 75–85.

[36] R. Pawar and R. R. Raje, "Multilingual cyberbullying detection system," 2019 IEEE International Conference on Electro Information Technology (EIT), 2019, pp. 040–044.

[37] K. Steimel, D. Dakota, Y. Chen, and S. Kübler, "Investigating multilingual abusive language detection: A cautionary tale," in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019, pp. 1151–1160.

[38] S. Dhuliawala, D. Kanojia, and P. Bhattacharyya, "SlangNet: A wordNet like resource for english slang," Language Resources and Evaluation Conference (LREC 2016), 2016, pp. 4329–4332.

[39] S. Rodríguez and J. Carretero, "A formal approach to spanish morphology: the coes tools," Procesamiento del Lenguaje Natural, 1996, vol. 19, pp. 118–127.

[40] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 632–642.

[41] L. Breiman, "Random forests," Machine Learning, 2001, vol. 45, pp. 5–32.

[42] J. Abellán and J. G. Castellano, "Improving the naive bayes classifier via a quick variable selection method using maximum of entropy," Entropy, 2017, vol. 19, no. 6, p. 247.

[43] D. W. Hosmer and S. Lemeshow, "Applied logistic regression," New York:Wiley, 2000.

[44] A. C. Gay Thomé, "SVM classifiers – concepts and applications to character recognition," in Advances in Character Recognition, Xiaoqing Ding, IntechOpen, 2012.

[45] V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: A comparison of retrieval performances," in Proceedings of SCEI Seoul Conferences, 2014, pp. 174–179.

[46] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," Pattern Recognition, 2016, vol. 58, pp. 121–134.