Overwhelmed by Negative Emotions? Maybe You Are Being Cyber-bullied!

Pinar Arslan, Michele Corazza, Elena Cabrio, Serena Villata

Overwhelmed by Negative Emotions? Maybe You Are Being Cyber-bullied!

Pinar Arslan, Michele Corazza, Elena Cabrio, Serena Villata
Université Côte d'Azur, CNRS, Inria, I3S, France
{pinar.arslan, michele.corazza}@inria.fr, {elena.cabrio, serena.villata}@unice.fr

ABSTRACT

With the increasing number of interactions, social media users have been vulnerable to intentional aggressive acts and cyberbullying instances. In this paper, first, we carry out a message-level cyberbullying annotation on an Instagram dataset. Second, we use the correlations on the Instagram dataset annotated with emotion, sentiment and bullying labels. Third, we build a message-level emotion classifier automatically predicting emotion labels for each comment in the Vine bullying dataset. Fourth, we build a session-based bullying classifier with the use of n-grams, emotion, sentiment and concept-level features. For both emotion and bullying classifiers, we use Linear Support Vector Classification. Our results show that "anger" and "negative" labels have a positive correlation with the presence of bullying. Concept-level features, emotion and sentiment features in different levels contribute to the bullying classifier, especially to the bullying class. Our best performing bullying classifier with n-grams and concept-level features (e.g., polarity, averaged polarity intensity, moodtags and semantics features) reaches to an F1-score of 0.65 for bullying class and a macro average F1-score of 0.7520.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;

KEYWORDS

Cyberbullying detection, emotion classification, sentiment analysis, social media

ACM Reference Format:
Pinar Arslan, Michele Corazza, Elena Cabrio, Serena Villata, Université Côte d'Azur, CNRS, Inria, I3S, France, {pinar.arslan, michele.corazza}@inria.fr, {elena.cabrio, serena.villata}@unice.fr, . 2019. Overwhelmed by Negative Emotions? Maybe You Are Being Cyber-bullied!. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, Article 4, 3 pages. https://doi.org/10.1145/3297280.3297573

1 INTRODUCTION

Internet can constitute a risk to the society, albeit being a very useful tool. One instance of this risk is *cyberbullying*. Cyberbullying is defined as "an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself" [7]. Many cyberbullying victims struggle with emotional burden, such as emotional stress, in that these victims may often face being threatened or offended in social media platforms. The automatic detection of cyberbullying, therefore, has benefits to the society, including avoiding vulnerable individuals' (e.g., teenagers) encounter with cyberbullying, and hence, minimizing any potential mental health conditions directly or indirectly caused by cyberbullying.

The main contribution of this paper is threefold. First, we provide a 1000-comment Instagram dataset annotated for cyberbullying, emotion and sentiment per Instagram post. With this dataset, we identify which emotion and sentiment features correlate with bullying instances. Second, we build a message-level emotion classifier which is then used to automatically predict emotion labels for each comment of the bullying dataset. Third, we build a session-level bullying classifier with the use of the following features: n-grams, emotion and sentiment features, message-level emotion and concept-level features. The impact of these features on the bullying classifier is unveiled. For both classifiers, Linear Support Vector Classification is implemented. Although emotion and sentiment features were employed in the cyberbullying detection tasks in the literature [5, 8], our study involves a larger set of emotion and sentiment features in different levels from various resources (e.g., Emolex, SenticNet).

2 TASKS

2.1 Correlation Analysis

In order to investigate the relationship of emotion and sentiment-related information with bullying instances, Pearson's correlation coefficient and 2-tailed p-value were measured on an Instagram dataset. Specifically, a portion of Instagram dataset [2] was annotated per post. This comment-level annotation revealed explicit correlation of bullying with emotion and sentiment features. We randomly selected 10 media sessions[1] (i.e., 5 bullying and 5 no bullying sessions) from the Instagram dataset. Two annotators from linguistics annotated 1000 Instagram comments, which were obtained from 10 sessions, with emotion, sentiment and bullying labels. The annotation was addressed using the following emotion, sentiment and bullying labels: *anger*, *fear*, *joy*, *sadness*, *other*, *no emotion*, *positive*, *negative*, *neutral*, *bullying* and *no bullying*. We computed the inter-annotator agreement on a subset of the annotated dataset using Cohen's Kappa. We obtained κ = 0.668 for emotion, κ=0.694 for sentiment, and κ=0.708 for the bullying annotations, meaning substantial agreement for all the tasks. Here is

[1] A media session is the thread of comments following a picture.

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04.
https://doi.org/10.1145/3297280.3297573

Table 1: Frequencies and percentages of annotated labels

Labels	Bullying	No Bullying	Frequency	Percentages
anger	118	5	123	12.3%
fear	2	0	2	0.2%
sadness	9	4	13	1.3%
joy	130	155	285	28.5%
other	73	28	101	10.1%
no emotion	284	192	476	47.6%
positive	143	164	307	30.7%
negative	158	12	170	17%
neutral	315	208	523	52.3%
bullying	61	0	61	6.1%
noneBll	555	384	939	93.9%

an example of annotated comment[2] with the labels anger, negative and bullying: "Shove off baby ugly @username". Table 1 shows the frequencies and percentages of the annotated labels for the sessions with overall bullying and no bullying labels.

2.2 Baseline Systems and Pre-processing

We use two baselines: majority-class and n-gram based baselines applied to 10-fold cross-validated emotion and cyberbullying datasets. Our emotion dataset with 2808 messages and 5 emotion labels (i.e., 577 "anger", 567 "fear", 690 "joy", 397 "sadness", 577 "other") was the combination of the WASSA-2017 [3] dataset (training and development sets)[3] and the annotated Instagram dataset where the "no emotion" tags were renamed as "other" tags. Our cyberbullying dataset with 970 sessions (i.e., 304 "bullying", 666 "no bullying") was the Vine dataset [6]. We used the word unigram-based emotion classification baseline on which only word tokenization was applied. Our cyberbullying classification baseline system was based on word (1,2) and character (3, 4, 5) n-grams. We implemented LinearSVC for the n-gram based baseline systems with TF-IDF weighting schemes. We addressed the following pre-processing steps for the cyberbullying baseline system, emotion and cyberbullying classification systems: cleaning the format of texts, word tokenization, tagging of URLs and usernames, removal of hashtag and stopwords, addition of a whitespace before and after punctuations, use of placeholders for adversative conjunctions, negative items and numbers, and stemming.

2.3 Emotion and Cyberbullying Classification

Our message-level emotion[4] and session-based bullying classifiers were based on LinearSVC. After a hyperparameter search, we selected penalty parameter `C` as 1.0 and `class_weight` as balanced. A 10-fold cross-validation was applied to the emotion and cyberbullying datasets. Upon building an emotion classifier with optimal performance, the whole emotion dataset was used as a training set and the emotion classifier was tested on the whole bullying dataset. In this way, we obtained automatically predicted emotion labels for each Vine comment. The following features were experimentally tested for the two classifiers: word n-grams (i.e., unigrams, bigrams), character n-grams (i.e., trigrams, fourgrams, fivegrams), emotion and sentiment features (i.e., word-level emotion and sentiment tags extracted from EmoLex [4]), SenticNet features (i.e., polarity, averaged polarity intensity per message or session, moodtags, semantics features obtained via SenticNet[1] knowledge base and input concepts), and message-level emotion features (i.e., automatically predicted emotion features used for the bullying classifier). We tested the contribution of each individual feature and concatenated the features on the classifiers. Except for the averaged polarity intensity, TF-IDF weighting schemes were applied to all the features. Our best performing emotion classifier is comprised of the first four features. Our best performing cyberbullying classifier was based on word and character n-grams used with concept level SenticNet features.

3 RESULTS AND DISCUSSION

The correlation results reveal a *strong* positive association for the pairs "anger-'bullying" (r=0.6805, $p < 0.05$) and "negative-bullying" (r=0.5631, $p < 0.05$), a *small* negative association for the pairs "joy-bullying" (r=-0.1609, $p < 0.05$), "no emotion-bullying" (r=-0.2429, $p < 0.05$), "positive-bullying" (r=-0.1696, $p < 0.05$), "neutral-bullying" (r=-0.2668, $p < 0.05$) and "other-bullying" (r=-0.0854, $p < 0.05$). The emotion labels "sadness" and "fear" have no significant correlation with the bullying instances. This lack of association might have stemmed from the fact these labels were only very few in number. We can conclude that bullying bearing messages can be detected more easily based on the emotion and sentiment labels of the messages.

We experimented various features in isolation and group on the emotion and cyberbullying classifiers. McNemar's test was applied to compare significant differences between the two systems based on the contingency table. Table 2 presents the results of emotion classification systems with macro average F1-scores[5]. The biggest contribution of individual features was obtained with word unigrams (i.e., 0.79), character fourgrams (i.e., 0.78) followed by character fivegrams (i.e., 0.77). The system with word unigrams was significantly different from the one with character fourgrams and fivegrams. The best performing emotion classifier (i.e., 0.82) with all features showed a significant difference compared to the baseline systems and the system with word unigrams. This suggests each single feature contributes in different aspects rendering a more sensitive emotion classification possible.

Table 3 displays the macro average F1-scores of the cyberbullying classification systems. The individual feature with the highest F1-score was character fourgrams (i.e., 0.7376). The system with word n-grams and character trigrams (i.e., 0.7497) was the only one system with a significant difference from the n-gram based baseline. The highest F1-score (i.e., 0.7520) was obtained with the system containing all n-gram features and SenticNet features, which shows the importance of polarity, averaged polarity intensity, moodtags and

[2]Warning: The examples on the paper include very explicit language. These contents do not reflect the views of the authors. It is, however, necessary to use such data despite its offensive nature as it is the only way to find methods to automatically master this kind of contents on the Web.

[3]We used the tweets with an emotion intensity score of 0.50 or higher.

[4]A message is either a tweet or an Instagram comment for the emotion dataset, and it is a Vine comment for the bullying dataset.

[5]F1-scores in bold show the best performing models.

Table 2: F1-scores of emotion classification systems

Majority Baseline	Unigram Baseline	word-1gram	word-2gram	char-3gram	char-4gram	char-5gram	EmoLex	SenticNet	Anger	Fear	Joy	Other	Sadness	Macro Average
✓									0.00	0.00	0.39	0.00	0.00	0.08
	✓								0.20	0.50	0.46	0.65	0.32	0.45
		✓							0.79	0.85	0.81	0.72	0.78	0.79
			✓						0.45	0.52	0.65	0.57	0.44	0.53
				✓					0.73	0.78	0.79	0.75	0.72	0.75
					✓				0.76	0.83	0.81	0.76	0.76	0.78
						✓			0.76	0.83	0.80	0.75	0.74	0.77
							✓		0.51	0.64	0.69	0.68	0.57	0.62
								✓	0.73	0.84	0.76	0.68	0.78	0.76
		✓	✓						0.77	0.84	0.81	0.74	0.78	0.79
		✓	✓	✓					0.76	0.83	0.82	0.77	0.76	0.79
		✓	✓	✓	✓				0.77	0.85	0.83	0.77	0.79	0.80
		✓	✓	✓	✓	✓			0.78	0.85	0.83	0.76	0.78	0.80
		✓	✓	✓	✓	✓	✓		0.77	0.85	0.83	0.76	0.78	0.80
		✓	✓	✓	✓	✓	✓	✓	**0.79**	**0.88**	**0.85**	**0.76**	**0.81**	**0.82**

Table 3: F1-scores of cyberbullying classification systems

Majority Baseline	word-1gram	word-2gram	char-3gram	char-4gram	char-5gram	EmoLex	pred.emotion	SenticNet	Bullying	No Bullying	Avg
✓									0.00	0.81	0.4071
	✓								0.62	0.83	0.7277
		✓							0.36	0.83	0.5946
			✓						0.62	0.84	0.7333
				✓					0.62	0.85	0.7376
					✓				0.61	0.85	0.7290
						✓			0.56	0.66	0.6113
							✓		0.61	0.77	0.6875
								✓	0.62	0.81	0.7154
	✓	✓							0.59	0.84	0.7142
	✓	✓	✓						0.64	0.86	0.7497
	✓	✓	✓	✓					0.63	0.86	0.7419
	✓	✓	✓	✓	✓				0.62	0.86	0.7374
	✓	✓	✓	✓	✓	✓			0.63	0.85	0.7407
	✓	✓	✓	✓	✓		✓		0.64	0.85	0.7466
	✓	✓	✓	✓	✓			✓	**0.65**	**0.86**	**0.7520**
	✓	✓	✓	✓	✓	✓	✓		0.64	0.85	0.7442
	✓	✓	✓	✓	✓	✓	✓	✓	0.64	0.84	0.7412

semantics on the bullying classifier. SenticNet (i.e., 0.65) and predicted emotion features (i.e., 0.64) had larger impact than EmoLex features (i.e., 0.63) on the bullying classifier. The three features significantly increased the performance of bullying class while they decreased the performance of the no bullying class. This proves that the bullying class necessitates emotion, sentiment and semantics features to perform well while the no bullying class gets better with the use of more abstract features such as character n-grams. The system with the highest macro average F1-score of 0.7520 had 25% error rate which might stem from the words with multiple EmoLex labels (e.g., bitch: sadness, anger, fear, disgust and negative), incorrectly predicted emotion labels, idiomatic expressions (e.g., shove up your ass) and internet slangs (e.g., Gtfo). Our findings show that SenticNet features including emotion, sentiment and semantics-related features, are useful complementary features, and hence, advisable while analyzing emotion and bullying in text classification problems.

4 CONCLUSION AND FUTURE WORK

In this study, we carried out the first (to the best of our knowledge) message-level cyberbullying annotation on an Instagram dataset. Each Instagram post was annotated with bullying, emotion and sentiment labels. Second, we calculated correlations on the annotations, which unveiled positive correlations for the pairs "anger-bullying" and "negative-bullying" and negative correlations for several pairs (e.g.,"no emotion-bullying"). Third, we adopted an approach to detect cyberbullying events, which was firstly to build a message-level emotion classifier and then to let the emotion classifier predict emotion labels for each Vine comment. We empirically showed that not only the correlation but also SenticNet, emotion and sentiment features in different levels impact on the cyberbullying detection task addressed on the Vine dataset, improving the results of the cyberbullying classification task, especially for the bullying class. There are several future research lines to be considered such as to include more emotional aspects from social media texts, extend our annotated dataset, add new features (e.g., word embeddings, slangs), use social network features, and tackle the issue of multilinguality.

ACKNOWLEDGMENTS

This work was funded by the CREEP project (http://creep-project.eu/), a Digital Wellbeing Activity supported by EIT Digital in 2018.

REFERENCES

[1] Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of AAAI, 2018.*

[2] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the Instagram social network. In *International Conference on Social Informatics, 2015.* Springer, 49–66.

[3] Saif M. Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In *Proceedings of WASSA, 2017.* Copenhagen, Denmark.

[4] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[5] Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang. 2014. Semi-supervised learning for cyberbullying detection in social networks. In *Australasian Database Conference, 2014.* Springer, 160–171.

[6] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In *Proceedings of ACM, 2015.* 617–622.

[7] Robert Slonje and Peter K Smith. 2008. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology* 49, 2 (2008), 147–154.

[8] Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. 2012. Fast learning for sentiment analysis on bullying. In *Proceedings of ACM, 2012.* 10.