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

Pinar Arslan, Michele Corazza, Elena Cabrio, Serena Villata

To cite this version:


HAL Id: hal-02020829
https://hal.archives-ouvertes.fr/hal-02020829
Submitted on 15 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 per post. This comment-level annotation revealed explicit 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.

CSC CONCEPTS
• Computing methodologies → Natural language processing;

KEYWORDS
Cyberbullying detection, emotion classification, sentiment analysis, social media

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ACM Reference Format:

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 sessions1 (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 $\kappa = 0.668$ for emotion, $\kappa=0.694$ for sentiment, and $\kappa=0.708$ 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.
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 of 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=-0.1696, 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.
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 de-