Automatic detection of cyberbullying on social media platforms

Ștefăniță Stan
University Politehnica of Bucharest
Splaiul Independenței 313,
București 060042
st.stan96@gmail.com

Traian Rebedea
University Politehnica of Bucharest
Splaiul Independenței 313,
București 060042
traian.rebedea@cs.pub.ro

ABSTRACT
The presence of cyberbullying on the Internet has grown alarmingly in recent years. Teenagers and children are the most affected by this phenomenon that is often the cause of higher suicide rates and social isolation. The detection and prevention of cyberbullying depends firstly on its correct understanding and secondly on the correct selection of a classification model trained on features that have a high discrimination factor between cyberbullying and non-cyberbullying. In this paper, we aim to create an automatic detection model for cyberbullying posts that is not biased towards a specific social media platform or a certain type of bullying. We describe the method we used for selecting the best features for two different classifiers trained on datasets collected from Twitter and Formspring. Next, we explain how we use the predictions made by these classifiers for labelling a new dataset collected by us from Twitter. The results of the automatic classification of the dataset have been compared to the manual classification of a sample of data from it, resulting in a rate of agreement larger than 50% between automatic detection and human annotation.

Author Keywords
Cyberbullying detection; Natural language processing; Social media; Online Aggressivity.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. .

General Terms
Human Factors; Languages; Measurement.

INTRODUCTION
Bullying is an increasingly frequent phenomenon, making its way into the Internet space, in the form of cyberbullying, along with the rising popularity of social media platforms. During a campaign meant to fight cyberbullying led by Bitdefender in 2017, *NU TASTA URA* (In English, *DON'T SEND HATE*)[1], an analysis of the cyberbullying phenomenon was made by questioning teenagers that were part of the Facebook group *Offtopic*, infamous for aggressive interactions between its members. The results have shown that about 80% of the questioned teenagers were victims of some level of cyberbullying (private or public), 66% of which did not tell anyone about it or asked for help. Moreover, only 36% of teenagers that witnessed instances of cyberbullying reported it or did something to help the victim. One of the characteristics of this phenomenon that makes it harder to notice than regular bullying is that it usually happens online, where parents or teachers may not see it. Also, its persistent and permanent nature (cyberbullying usually happening over a long period of time and instances of it remaining forever on the Internet space) have the capability of destroying the reputation of the aggressor, as well as of the victim. Therefore, research efforts were put into detecting and solving cyberbullying as quickly as possible in order to prevent the issues that may arise from it.

Most studies regarding this topic were done in the social and psychological fields to understand how this behavior appears, as well as how it affects both victims and bullies. Only recently it became a subject of interest for the information technology community, research being made on social media networks in order to find solutions for quick detection and prevention of cyberbullying. What scientists quickly discovered was that several issues arise when trying the traditional methods of text classification on this problem. The main issue is that there is not a consensus regarding a clear definition of cyberbullying, thus different researchers use different definitions, making future work harder, as there is not a solid basis for comparison. However, most of the definitions make references to the persistent and aggressive nature of cyberbullying. In this paper we use the definition given by Smith et al. [1] that defines cyberbullying 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". Another issue is generated by the bias of the already few publicly available datasets that are specialized on a specific type of aggression (e.g. sexism, racism, homophobia) or include data from only one social media platform. Finally, selecting the best combination of features that would work well on the training dataset as well as on future data is not an easy task due to changes in language or topic drift.

[1] The Bitdefender NU TASTA URA campaign site: nutastaura.bitdefender.ro, last accessed 25 July 2020.

RELATED WORK

The low number of publicly available datasets annotated for cyberbullying as well as a lack of variety among them, most of them targeting a single social media platform and containing out of context posts, make cyberbullying detection a complex problem to solve. That is the reason research mostly focuses on finding solutions for a specific social media platform or type of cyberbullying.

One of the more recent solutions regarding detection of cyberbullying on Twitter proposed by Chatzakou et al. [2] is based on the selection of features with a high discriminatory factor for more accurate classification. The dataset used for classification was collected over a period of 3 months and the goal was to label the users, rather than tweets. Therefore, a set of tweets posted by the same user was given to the annotator, that had to put the user into one of the categories: aggressive user, bullying user, spammer, or normal. This was done in order to consider the repeatability characteristic of cyberbullying, an accurate classification of bullying being hard to make without an ongoing communication between the bully and the victim.

Another study of Twitter posts was conducted by Al-garadi et al. [3]. A series of different binary classification models (Support Vector Machines, Random Forests, and Logistic Regression) and combination of features were tested to find the best classifier. Some of the features selected for training were number of followers, number of posts, number of tags, gender, and age of user. A ranking of those models and features shows that the most relevant features for classification were the presence of vulgarities and the gender of the user. We introduced such features in our solution as well.

Van Hee et al. [4] aimed to collect and use a dataset from Ask.fm, a social media platform based on question-answer type of posts, like Formspring from which one of our datasets is collected. A particularity of this study is that the manual annotators were guided to label the posts into a larger variety of categories that may indicate some form of cyberbullying, namely threats, insults, sexually explicit messages, texts encouraging a bully. Moreover, they introduce 2 new possible roles in the cyberbullying act: instigator (person that encourages a bully) and vigilante (person that steps up for the victim).

The importance of selecting an efficient combination of training features was illustrated in a study published by Dadvar et al. [5] where a dataset comprising comments collected from Youtube videos is used for training and analysis. Features regarding the user were found to bring high information value to the predictions, therefore they introduced features like the age and sex of the poster, or even the way in which they behave online, details like the time of the posting or the number of followers a user has, being also considered to be important.

In relation to the previous work, we propose to use traditional natural language processing methods as they are faster, explainable and require less computation. We aim to eliminate the shortcomings generated by the bias of using only a social media platform by combining models trained on 2 different datasets. Moreover, we intend to select the best combination of manually crafted features. To test our model architecture, we also collected Twitter posts over a period that are fed into our models to predict their class. At the end, we test a sample of those predictions against the results of manually annotations for the same data.

METHOD

Datasets

For training our models we combined 3 publicly available datasets. The first two were used for training the cyberbullying detection model while the third one, collected from Facebook, was used for training an aggressivity classifier. The aggressivity classifier was then used to compute the aggressivity factor feature for each post in the first two datasets. As it can be seen in the table below the number of positive examples, namely cyberbullying instances, in the training datasets is much lower than the number of negative examples, especially in the case of the Formspring datasets. In order to try and minimize the issues that may arise from this, we tried duplicating some positive examples and insert them into the datasets as well as change the class weight parameter of the model.

Dataset source	No. of classes	No. of entries	% of positive examples
Cyberbullying datasets			
Twitter	3	2 751	40.93 %
Formspring	2	12 773	6.08 %
Aggressivity dataset			
Facebook	3	12 000	57.90 %

Table 1. Datasets summary

Dataset source	Twitter	Formspring
Vulgarities	13.04%	18.21%
Vulgarities out of bullying posts	17.05%	65.72%
Aggressivity	65.53%	55.21%
Aggressivity out of bullying posts	72.29%	73.96%

Table 2. Vulgarity and aggressivity in cyberbullying datasets

After training the aggressivity classifier on the Facebook dataset, we proceeded to analyze the datasets which were used for the cyberbullying classification. Two factors were considered important, namely the presence of vulgarities

and whether a post has an aggressive tone. Therefore, we used the aggressivity classifier to predict the classes of these datasets and a list of popular vulgar words. We can see in Table 2 that vulgarity is not directly correlated with cyberbullying, only 17.05 % of the bullying examples from Twitter also containing vulgarities. However, in the last row of the table we can see that over 70 % of bullying examples from both datasets are also considered to be aggressive, aggressivity being often associated with bullying.

Training Features

Inspired from previous works, we manually crafted and selected training features that are relevant for cyberbullying detection. These features were further grouped into different categories in order to test which type of features brings the most information to the cyberbullying classification. Those categories, along with the features included in them, are further explained in this subsection.

Content Features

- **Presence and frequency of vulgar words.** While there is not a clear relationship between vulgar words and cyberbullying, some studies that analyzed different feature combinations discovered that vulgarity related features are a high discriminatory factor for cyberbullying detection [6]. Another study focused on the presence of vulgar words on Twitter [7] found that compared to other social media platforms, Twitter posts have more vulgar words. This is relevant for our study too, considering that the datasets we collected include posts from Twitter.
- **Presence of hyperlinks.** We chose this attribute after analyzing a sample of bullying posts where hyperlinks were frequent. Most of the times these were links to photos in the form of rude memes addressed to a person.
- **Frequency of capital letters.** In the absence of information given in verbal communication, like tone and emphasis, we use other indicators to suggest the way we intend our message to be read. The use of all capital words or sentences is meant to put emphasis on certain words or may suggest a raised tone of the post's author. In relation with cyberbullying, the use of capital letters may be associated with an aggressive tone or yelling.
- **Superlatives.** Like capital letters, superlatives may be used in textual communication to emphasize a certain message. For this feature, we considered the presence of words indicating superlatives (e.g. *very*, *the most*) as well as the presence of words from a list of superlatives we comprised.
- **Post length.** We introduced this feature in order to create a clearer separation between shorter posts that are vulgar or aggressive and longer posts that have a lower aggressivity value but have a big change of being considered cyberbullying through their content.

Subjectivity Features

- **Presence of second person pronouns.** The use of second person pronouns indicates that the content of the post is directly targeting a specific person. In combination with insults or an aggressive tone, this might indicate the presence of cyberbullying.
- **Presence of first-person pronouns.** By analyzing the results of some early versions of our models we observed that some falsely classified as bullying posts were written in first person. Most of those were self-denigration, meant to be light-hearted self-jokes.
- **Presence of mentions.** Similar to second person pronouns, mentions also indicate who the content of a post is targeted at.
- **Vulgarity-Pronoun/Vulgarity-Mention pairs.** This feature is used to indicate whether a vulgar word was addressed to a person. For a post, we verify if a vulgar word is at most 2 words apart from a mention or pronoun. The list of vulgar words was created by collecting banned words from different social media platforms and is available online[2].

Aggressivity Features

The overall aggressivity of a post can be a good indicator of the author's mood, indicating a possible feeling of frustration or fury. Moreover, several studies consider aggressiveness as a definitory factor of cyberbullying [2, 3], while there is still much debate regarding differences between aggressivity and cyberbullying and whether they represent the same thing. The value for this feature was computed by using the aggressivity classifier trained on the Facebook dataset.

General Content Features

Finally, we introduced features regarding the words present in a post. To compute the values for these features we considered both single appearances of words, as well as pairs of words. We considered the most frequent 10k such n-grams, resulting in 10k different numerical features.

Algorithms

For selecting the best classification algorithm for this problem, we conducted several tests with different combinations of training features and algorithms and recorded the metrics for them. Since for cyberbullying detection we would rather have a larger number of false positive examples, rather than false negatives, the most important metrics for us are the recall and precision. Out of Support Vector Machines (SVM), Random Forest, and Logistic Regression, SVM had the best score overall and therefore was chosen as our classifier. More, SVMs are used in almost all studies of cyberbullying detection either as the main algorithm or in comparison to others [4, 6].

[2] The entire code is available online in the repository: https://github.com/stefanx17/cyberbullying_detection

Figure 1. Cyberbullying classification steps

CLASSIFICATION PIPELINE DETAILS

Pre-processing and Feature Computing

Dataset Cleanup

This step was only applied for the dataset containing examples collected by us from Twitter. We considered that some examples may introduce unnecessary noise in the classification process and filtering them out would solve this issue. Therefore, we eliminated all retweets and duplicated entries from the training dataset.

Text Pre-processing

As shown in Figure 1, before computing the training features values, we cleaned-up the text by eliminating or replacing some elements in the text that do not bring any discriminatory information to the classification. We did this by using regular expressions applied to the text.

The modifications applied to the text are as follows:

- **Remove mentions and hyperlinks from the text.** Mentions represent references to other users and are marked in the text by the presence of the "@" symbol before the name of another user, while hyperlinks are marked by the presence of "http://" or "https://".
- **Restricting sequences of the same characters to a length of maximum 2.** Repeated consecutive appearances of the same character can either be typos or intended by the author to suggest the way the text is meant to be read. However, even in the latter case, there is no rule to how long that sequence can be, resulting in different length sequences that have the same meaning but are not considered to be the same when we compute features.
- **Elimination of unknown characters.** We chose to remove all non-ASCII characters.
- **Elimination of punctuation marks.** We remove all punctuation marks except the apostrophes or combinations of marks that may represent an emoticon.
- **Elimination of text sectioning rules.** This is applied only to the Formspring dataset where posts represent question-answer pairs, indicated by the presence of 'Q:' and 'A:'. We choose to eliminate those and consider the whole text of the post for training the model. We made this decision because splitting the text would make it hard to decide if bullying is present in the question or in the answer.

Hyperparameter	Tested values
Kernel type	{Linear; Poly; Rbf; Sigmoid}
C penalty	{0.001, 0.003, 0.01, ... ,100, 300, 1000}
Class weight	Default; Balanced

Table 3. SVM hyperparameters

Computing the Training Features

The only features that are computed before the text-processing are the ones regarding the presence of mentions and hyperlinks, as they are removed during the cleanup step. These are binary features, so their values are either 1, if these elements are present in the text, or 0 otherwise. The numerical features were computed by counting the number of appearances of certain words in the text of the post (the number of vulgar words, for example).

For the aggressivity feature, we used the classifier that predicts the aggressivity of a post on a scale from 0 to 1. We trained this classifier on the dataset specifically annotated for aggressivity and used it to classify all the posts in our training cyberbullying datasets.

Finally, for the features regarding the general content of a post we made use of modules designed for text feature extraction in the scikit-learn library [8]. We used the CountVectorizer module for determining the vocabulary of our dataset, composed of the most used unigrams and bigrams, and TfidfTransformer to get the final value for each feature.

Classifiers Architecture

To obtain the classifier used for detecting cyberbullying in a live collected dataset from Twitter, we needed to train and use several models, one for determining the aggressivity of a post and one for each cyberbullying dataset. As previously stated, we primarily selected SVM, for which we further did a grid search to select the best combination of hyperparameters. The hyperparameters we varied, as well as their values, can be seen in Table 3.

Aggressivity Classifier

For getting the value of the aggressivity feature we trained another model on a dataset annotated for aggressivity collected from Facebook. As features, we used TF-IDF using the most frequent 10k n-grams (n=1..3). As for the cyberbullying classifiers, the best results were obtained when using an SVM model.

Cyberbullying Classifier

In order to determine and select the best feature combinations we conducted several experiments by training the models with different features and comparing the metrics for each case. Before these experiments we conducted a grid search in the space of the hyperparameter values to determine the best values for the hyperparameters. The grid search was done by only using features described in the baseline configuration, namely TF-IDF features. All combinations have been tested, choosing variants that deliver the best results on each of the three machine learning algorithms used. After determining the values of the hyperparameters to be used when training our machine learning models, several tests were carried out, for each of the two classification models, training one at a time with different attributes. We will continue by presenting the attribute configurations chosen to be tested.

- **Baseline.** Only contains features consisting of unigrams and bigrams frequency (TF-IDF). We chose this as our baseline as these are the most common features for text classification.
- **Baseline + Content Features.** This feature configuration will show us how important is the content of a message (presence of certain words or phrases) for detecting cyberbullying.
- **Baseline + Content Features + Subjectivity Features.** In addition to the previous configuration we included subjectivity features, indicating who the content of a message was targeted at.
- **Baseline + Content Features + Subjectivity Features + Aggressivity Features.** This configuration contains all the features we computed for the datasets.
- **Baseline + Aggressivity Features.** We considered this configuration in order to find out how correlated is aggressivity with cyberbullying and to see if the aggressivity of a post is a strong enough feature to determine if it can be labelled as cyberbullying or not.
- **Content Features + Subjectivity Features + Aggressivity Features.** We want to see how well our manually crafted features do when we do not include TF-IDF features for training our models.

Collection and Classification of Twitter Data

For the last part of this research we intend to test our classifiers on a live dataset collected by us from Twitter. This data will be cleaned and pre-processed and then fed into our classifier to predict each post's class. Lastly, we will compare the automatic detection against a sample of manually annotated data to see how well our classifier performs on real live tweets.

Collection of Twitter Datasets

For testing our classifier, we decided to collect new data from Twitter that we will annotate both automatically with our trained classifier and manually with 3 different human annotators. Therefore, we collected two datasets, one with random posts and one by searching for keywords that may indicate the presence of cyberbullying. The keywords were taken from a list that contains the most used words in cyberbullying posts from the datasets we used for training. To increase the possibility of a higher variety of topics, we collected posts every day for a period of 6 weeks in May-June 2019. For data collection, we use the API provided by Twitter and tweepy (https://www.tweepy.org/), a Python library that provides methods for post searching.

Because the datasets we use for training our models mostly contain English posts, we chose to only collect posts written in English. Therefore, the filters we introduce in our posts search call were the language of the post and a list of keywords for one of the datasets.

Automatic Classifier Architecture

For the classification of the newly collected data we decided to use two previously trained classifiers, one on the Formspring dataset and the other on the Twitter dataset (see Table 1). From both those classifiers we obtained a probability of the presence of cyberbullying in a tweet that we combined to obtain a final prediction. We did this by computing the median of those predictions, except for when at least one of the classifiers predicted the cyberbullying class with a very high confidence (more than 85%), in which case the post was automatically considered to be cyberbullying even if the median indicated otherwise. We consider this exception to be necessary as one of the classifiers might not be sensible to a certain type of cyberbullying, resulting in a lower cyberbullying score even if the post has some cyberbullying indicators.

Manual Annotation of Posts

From each dataset we selected a sample of 100 posts that had the highest cyberbullying score to be manually annotated. Therefore, in the first dataset collected based on the presence of words that may indicate cyberbullying all the selected posts were labelled as cyberbullying by the automatic classifier, while only 45% of posts in the random dataset were manually labelled as cyberbullying.

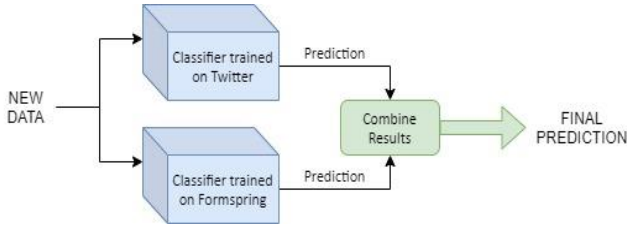

Figure 2. Cyberbullying classifier architecture

Bullying
The presence of cyberbullying is obvious and one or more of the indicators are present: • Humiliation of another person • Vulgar words • Indicators that suggest the repeatability of cyberbullying • Aggressive tone used by the author of the post • Other indicators considered to be relevant by the manual annotators
Maybe Bullying
The presence of cyberbullying is not obvious, but some indicators of cyberbullying are present in the text of the post. Some relevant indicators are: • One or more indicators presented previously in the definition of cyberbullying • General references that intend to humiliate/harm a category of people (by race, sexuality), but are not directed at a specific person (e.g. "All gays must be killed") • The possibility to identify a context in which that post might be considered cyberbullying (e.g. „Just end it already" might be an allusion to suicide in some cases)
Not Bullying
It is obvious that cyberbullying is not present, as none of the previous indicators are satisfied.

Table 4. Cyberbullying classification guidelines

After selecting the posts, we determined 3 classes in which the manual annotators can sort the posts they were given. For the manual annotation, 3 people were asked to determine the class of each post based on the text of the post as well as on the guidelines described in Table 4.

Finally, after the manual annotation, we analyzed the results and computed metrics of the given answers. One of the metrics we computed was the inter-rater agreement rate, used to determine the similarity of the answers from different annotators. The second one refers to the frequency of cyberbullying according to our classifiers' opinions. To determine this percentage, we split our samples in groups of 20 posts and analyzed them individually.

To compute the acceptance rate between the classifiers, each posts received a score:

- 3: If all the annotators chose the same label for the post
- 1: If at least two of the classifiers chose the same label
- 0: If all the classifiers chose a different label for the post

Finally, we combined these scores and obtained an overall acceptance rate by using the following formula:

$$Acceptance\ rate = \frac{\sum_{p \in Posts} score_p}{||Posts|| \times MAX\_score}$$

RESULTS

Cyberbullying Classification Results

The configurations selected for testing our detection algorithms can be seen below and were further explained in the previous section. For each new combination, we maintained the previous features and added a new category until we reached a configuration where we use all the features selected by us (D). Then, we also test how aggressivity features alone influence the results and we also test a feature combination where we don't use TF-IDF.

A. Baseline (TF-IDF)
B. Baseline + Content Features (CF)
C. Baseline + CF + Subjectivity Features (SF)
D. Baseline + CF + SF + Aggressivity Features
E. Baseline + Aggressivity Features
F. CF + SF + Aggressivity Features

The performance metrics are precision, recall, f1-score, and accuracy. In the case of cyberbullying, the most relevant metric is the recall, as we want the percentage of undetected cyberbullying posts to be as low as possible. The f1-score is also relevant as we want to maintain a balance between precision and recall, too many posts wrongly labelled as cyberbullying not being good either.

The results of using our classifier architecture on the two different datasets can be observed in Table 5. On a first look, we can clearly see that better results were obtained for the Twitter dataset, the Formspring dataset having a much lower precision, thus ignoring many of the cyberbullying posts. Some of the reasons for this difference could be concerning the different platforms these datasets were collected from, as well as different periods of times, and different guidelines for manual annotation. Analyzing the different feature combinations, we can see a small improvement in adding subjectivity and content features, especially in the case of the Formspring dataset.

Cfig / Metric	A	B	C	D	E	F
Twitter dataset						
Precision	63.5	63.2	**64.9**	64.4	63.4	49.5
Recall	**82.4**	79.6	80.5	80.5	78.7	50.9
F1	71.7	70.4	**71.9**	71.6	70.2	50.2
Accuracy	74.6	73.9	**75.3**	75.0	73.9	60.5
Formspring dataset						
Precision	34.0	35.4	36.1	**37.1**	34.3	36.0
Recall	63.0	60.2	64.3	63.0	60.2	**65.7**
F1	44.2	44.6	46.3	**46.7**	43.7	46.6
Accuracy	90.9	91.4	91.4	**91.7**	91.1	91.3

Table 5. Classification test results

	Bullying Sample May-June 2019	Random Sample May-June 2019
Sample size	6783	3144
Bullying %	6.04 %	1.46 %
Aggressivity %	48.68 %	55.69 %

Table 6. Results of automatic classification

	Bullying Sample May-June 2019	Random Sample May-June 2019
Bullying %	53.00 %	36.00 %
Inter-rater Reliability	67.33 %	68.67 %

Table 7. Results of manual classification

Twitter Sample Classification Results

Automatic Classification Results

As it is illustrated in Table 6, the dataset collected by searching for words that appear often in cyberbullying posts has a higher percentage of posts classified as cyberbullying, almost 9 times more posts than in the case of the randomly collected samples from Twitter. In the case of aggressivity, both datasets have a high presence of it in their entries, about half of the posts being labeled as aggressive. Even though this percentage is troubling, the fact that is way different than the percentage of cyberbullying further shows that these two phenomena should not be confused with each other, each having their particularities.

Human Classifiers Results

To obtain the following statistics, we selected a sample of 100 posts from each dataset by the probability of cyberbullying assigned to each of them by the automatic classifier. Before presenting them to the human classifiers we randomized their order and hid their cyberbullying score. Then, the human classifiers were asked to assign each post to one of the three possible classes: Bullying, Maybe Bullying, Not Bullying by following the guidelines presented in Table 4. As we consider important in cyberbullying detection to consider any indication that this phenomenon appears in a post, for the statistics below we label the post as cyberbullying if at least 2 out of the 3 annotators classified it as Bullying or Maybe Bullying.

As with the automatic detection, the random sample has a lower presence of cyberbullying, this is partly due to the fact that more than half of the posts were not classified as cyberbullying by the automatic detector either.

Also, we can observe in Table 7 that the inter-rater reliability score is quite high, being higher than 65% for both datasets. This indicates that people have an acceptable rate of agreement when it comes to identifying cyberbullying. The reason why we don't have a higher rate of agreement can be the fact that this is a hard phenomenon to identify in small posts taken out of context, even when it comes to classifiers presumed to have a clear understanding of cyberbullying and its indicators.

Finally, considering the relatively low percentage of bullying found in the randomly collected sample (1.46 %), we think that a solution based on automatic detection, paired with human classifiers in the form of moderators on different social media platforms could help reduce the negative impact of cyberbullying on the Internet. A simple and straightforward solution would be having a classifier app triggered by common indicators of cyberbullying (similar to how we collected our bullying sample), then a moderator should be notified of any possible presence of cyberbullying behavior in order to investigate and take a final decision regarding the suspected post.

CONCLUSIONS

Cyberbullying has become an increasingly larger problem in recent years, affecting the mental health and safety of people, especially when it comes to children and teenagers. Therefore, solutions to best handle this situation and try to solve are of high interest for many organizations. Thus it has become a topic of interest for machine learning research meant to detect cyberbullying as accurately as possible.

After conducting several experiments with different training features combinations, we found out that the features introduced by us improve the classifications, in all cases the metrics being better than the baseline configuration. However, these must be used in combination with general content features (TF-IDF) in order to provide good results.

Our next goal was to engineer a model capable of accurately classifying new data and not be biased towards a certain social media platform. We tried to eliminate this bias by combining classifiers trained on 3 different social media platforms (Twitter, Formspring, and Facebook) in order to get a final prediction. To test out the proposed method, we collected two new datasets from Twitter (one random and one based on cyberbullying indicators) that we automatically classified with our model. We compared the results of automatic classification against a sample of manually annotated posts and we discovered that more than 50% of the posts detected as cyberbullying were also labeled as cyberbullying by human annotators. We consider this score to be very good, as automatic detection can be doubled by a human moderator that can make a final decision and decide whether to take action or not.

REFERENCES

1. Smith, P. K., del Barrio, C., & Tokunaga, R. S. (2012). Definitions of Bullying and Cyberbullying: How Useful Are the Terms?. In Principles of Cyberbullying Research (pp. 54-68).
2. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. In Proc. ACM (2017) on web science conference (pp. 13-22).
3. Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior, 63, 433-443.
4. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PloS one, 13(10).
5. Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In European Conference on Information Retrieval (pp. 693-696).
6. Dadvar, M., Jong, F. D., Ordelman, R., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In Proc. Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012).
7. Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014). Cursing in english on twitter. In Proc. 17th ACM conference on Computer supported cooperative work & social computing (pp. 415-425).
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 2825-2830.