2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Unsupervised Cyber Bullying Detection in Social Networks

Michele Di Capua
Department of Computer Science
University of Milan
Milan, Italy
michele.dicapua@unimi.it

Emanuel Di Nardo, Alfredo Petrosino
Department of Science and Technology
University of Naples "Parthenope"
Naples, Italy
{emanuel.dinardo, petrosino}@uniparthenope.it

***Abstract*—Modern young people ("digital natives") have grown in an era dominated by new technologies where communications are pushed to quite a real-time level, and pose no limits in establishing relationships with other people or communities. However, the speed of evolution does not allow young people to split consciously acceptable behaviors from potentially harmful ones and a new phenomenon known as cyber bullying is emerging with increasing evidence, attracting the attention of educators, and media. Cyber bullying is defined as "willful and repeated harm inflicted through the use of electronic devices" [1]. In this paper we propose a possible solution for automatic detection of bully traces over a social network, using techniques derived from NLP (Natural Language Processing) and machine learning. Specifically, we shall design a model inspired by Growing Hierarchical SOMs, able to cluster efficiently documents containing bully traces, built upon semantic and syntactic features of textual sentences. We fine-tuned our model to work with the social network Twitter, but we also tested the model against other social networks such as YouTube and Formspring. Finally, we report our results, showing that the proposed unsupervised approach could be effectively used with good performances in some scenarios.**

***Index Terms*—clustering; cyber bullying; self-organizing map; sentiment analysis; social networks; unsupervised learning.**

I. Introduction

With the spread of mobile technologies, cyber bullying has become an increasing problem, especially among teenagers. Awareness has also increased, due to some episodes of suicide. According to recent studies almost 43% of teenagers in the U.S. revealed to be victims of cyber bullying. It is, therefore, evident that the availability of tools that can automatically identify possible behaviors classified as cyber bullying, can be really useful to prevent situations of "risk" to the victim. Even if the problem is now heavily considered from a social point of view, computational studies in this field are largely yet unexplored and only few researches on cyber bullying are available. We propose a possible solution for automatic detection of the bully traces, i.e. social media posts containing harmful text or sentence that could possibly lead to a cyber bullying episode. We shall show that using both techniques derived from NLP, in the pre-processing data stage, and the subsequent adoption of unsupervised machine learning algorithms, for the detection phase, can lead to reliable results. We propose here a new model of cyber bullying detection also on the basis of Sentiment Analysis approach, considering, as an assumption, that a cyber bullying post is an extremely negative message. The unsupervised approach we pursue is aimed to avoid manual labelling of the datasets that are huge and imposing any a priori assumption about possible classes. In the following sections, after a background introduction and a view on the state of the art in this area, we present our cyber bullying detection model and we present some results obtained by applying the model to some dataset.

II. Background

Research in sociology and in psychiatry can provide important algorithmic insights to define models in order to detect bully traces. From a psychological point of view there are some special features that differ cyber bullying from the traditional bully behavior. First, the absence of relationship between victim and bully. In fact, for those who suffer harassment is even more difficult to defend themselves against this phenomenon because very often the victims cannot even identify who the bully is. Often the bully hides himself behind false names and not having direct contact with the victim lowers his inhibition. The aggressor does not always receive communications by the victim that may mitigate or modify his aggressive behavior. Last, the lack of space and time limits represents another fundamental aspect: cyber bullying can damage the privacy of the victim, at any time of day or night. Given the characteristics of virtual communication it is also necessary to reconsider the criterion of repetition inside cyber bullying phenomenon; in fact, only one simple information (message, video, or a photo) disclosed to many people through the Internet or smart phones, can cause damage to the victim regardless of its repetition, being able to be viewed and re-transmitted by many people at different times, remaining available to the community for a long time.

III. Related Works

While cyber bullying is a well-studied problem from a social point of view, only recently it has attracted the attention of computer scientists, especially towards automatic detection tasks. For this reason, only relatively few articles on the subject and very few datasets are available.

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

Yin, et al. [2] adopted a supervised learning technique for detecting harassment, using a bag of words model based on content, sentiment and contextual features of documents to train an SVM classifier. The authors used a combined model based on sentiment and contextual features, reaching, with a support vector machine learner, a recall level of 61.9%.

Dadvar et al., [7] analyzed the gender approach within the cyber bullying detection problem, applied to the social network MySpace, a platform that offers an interactive, user-submitted community of friends with personal profiles, blogs, groups, etc. Authors investigated the content of the posts written by the users but regardless of user's profile information. They used an SVM model to train a specific gender text classifier. The dataset consists of about 381.000 posts. The results obtained by the gender based approach improved the baseline by 39% in precision, 6% in recall, and 15% in F-measure.

At MIT, Dinakar et al. [4] applied different binary and multiclass classifiers on a manually labeled corpus of You Tube comments. This approach reached 66.7% of accuracy. Also, in this case authors used an SVM learner.

Kontostathis et al. [6] adopted a language based approach for cyber bullying detection. Authors collected data from Formspring.me, a "question and answer" social network, manually labelling data using Amazon's Mechanical Turk. Authors used rule based learning method and a bag-of-words approach based on a C4.5 decision tree learner and an instance-based learner. They identify true positives cyber bullying posts with an overall accuracy of 47.7%.

Xu, et al. [5] proposed different natural language processing techniques to identify bully traces and also defined the structure of a bully episode and possible related roles. Authors adopted Sentiment Analysis to identify roles and Latent Dirichlet Analysis to identify topics. Cyber bullying detection is formulated as a binary (positive/negative) classification problem and a linear SVM is trained with manually labelled dataset. The results reported 89% of cross validation accuracy, showing that even basic features and common classifier, can be useful to detect cyber bullying signals in text.

We can observe that most of these studies are based on supervised approaches, and usually adopt pre-trained classifiers to solve the problem, typically based on SVM. Data are manually labelled using online services or custom applications, and are usually limited only to a small percentage. NLP techniques are obviously wide adopted in all these works, due to the strict correlation between text analysis and cyber bullying detection. Mostly NLP tasks are performed at the preprocessing stage.

IV. Proposed Model

We want to develop a model of cyber-bulling aggression, based on a hybrid set of features, starting with classical textual features but also based on the so-called "social features". Our model will avoid a bag-of-words (BoW) approach because this approach does not consider the position of words in a sentence but also because in the BoW model the feature space can be significantly large. In order to accomplish our task we manually build some features considering the cyber bullying problem from different points of view. First, aggressive sentence (bully traces) can be pre-filtered using syntactic and semantic analysis, using NLP algorithms, in order to find, for example, bad words occurrences in a document. We also consider emotional traces, inside a document that could lead to a more precise detection, introducing in the model, also sentiment analysis features. Our assumption is that cyber bullying detection can be effectively formulated as a particular sentiment analysis problem. Then, we also introduce in our model features strictly related to the social network platform addressed.

We divide features in groups, to distinguish features based on pure text analysis from features related to statistical or social analysis approach. We propose to build the model onto 4 distinct features group, divided into:

- Syntactic features (F_{syn});
- Semantic features (F_{sem});
- Sentiment features (F_{sen});
- Social features (F_{soc});

So, our global set of features F, related to a document can be expressed as:

$$F = \{F_{syn}, F_{sem}, F_{sen}, F_{soc}\} \tag{1}$$

For each group we selected some features considered both from experience and from recent literature as partial good indicators of a cyberbullying sentence. For each document, we associate an input vector (2) as the weighted concatenation of the selected features:

$$x(t) = \alpha F_{syn} \oplus \beta F_{sem} \oplus \gamma F_{sen} \oplus \delta F_{soc} \tag{2}$$

where t is a sentence, F is a set of features, and α, β, γ and δ are weights related to each single group, that can be used to tune the model in particular context, i.e. these weights, correctly updated, could be used to apply the model to different social network platforms. In out tests, for the Twitter platform, these weights are all set to values > 1.

A. Syntactic features (F_{syn})

These features are generally obtained by statistical analysis of documents (tweets or sentences):

1) Bad words: From literature is quite evident and intuitive that some "bad" words make a text a suitable candidate to be labeled as a possible cyber bullying sentence. As just done in other works, we have identified a list of insults and swear words (550 terms), collecting these terms from different online available sources.

2) Bad words density: In our model we check also the density of "bad" words as a single feature. This features is equivalent to the number of bad words that appear in a sentence, for each severity level, divided by the words in the same sentence.

3) Badness of a sentence: We also add a feature to our model in order to measure the overall "badness" of a text. This feature is computed by taking a weighted average of the "bad" words (weighted by a severity assigned).

4) Density of upper case letters: This feature is based on Dadvar et al. [7] results. The presence of capital letters in a text message is selected as a feature, considering it as possible 'shouting' at someone behavior, as commonly treated in social networks netiquette. This feature is given by the ratio between the number of upper case letter and the length (number of chars) of the whole sentence.

5) Exclamations and questions marks: Just like capital letters, also exclamation points and question marks can be considered as emotional comments. We just stated that cyber bullying is related to an extreme case of sentiment analysis and so it can be connected to the strong (usually bad) emotions. With this premise, we consider helpful to introduce the number of exclamation points and question marks as a feature in our model.

B. Semantic features (F_{sem})

The semantic features adopted here are based on semantic meanings, e.g. "a person", that could represent some entities extracted from documents. The assumption behind introducing semantic features in our model is that some entities tend to be more correlated with cyber bullying and, more in general, with positive or negative sentiment. These correlations can help discovering similar structures and can increase the overall accuracy of bully traces detection. For the Twitter case, we use the part-of-speech tagger developed by Carnegie Mellon, to detect bigrams and trigrams structures in sentences.

1) Bigrams: Use of offensive words is a common way of harassing someone over the web. Trivially, the adoption of foul language may be considered a sign of a potential cyber bullying episode. It's also common that when people are harassing others, they commonly tend to use personal pronouns. Therefore a good indicator of harassment can be considered the usage of personal pronouns appearing near bad words. From literature has been also observed that second person pronouns, such as "*you*" and "*yourself*", are more relevant respect to other possible pronouns. Using Part Of Speech analysis, it's possible to detect, as a feature, the presence of commonly occurring bigram pairs in a bullying sentence such as "*you are*", "*yourself*", and so on.

2) Trigrams: A still open problem in text analysis, is the negation handling. A negation (such as "no" and "not") is near to a word which precedes it or follows it, i.e. "I do not like you". The adoption of N-Gram windows inside text can help at least to mitigate some controversial sentences that contain negations, so in our model we try to detect such structures using trigrams, in order to improve the accuracy of clustering.

C. Sentiment features (F_{sen})

Sentiment analysis and cyber bullying detection were topics strictly correlated. In a cyber bullying post there is a wide range of emotions that can be used both to identify victims and bullies. Twitter posts usually contain noisy text, i.e. text that does not follow the standard rules of orthography, syntax and semantics. For polarity evaluation of a tweet we filter out words shorter than k characters, where k is an integer experimentally fixed to $k = 3$.

1) Sentiment polarity of a sentence: The polarity score of a single tweet is computed as the average of the sum of the polarity of its words [10]. Given a tweet message *m* as a collection of *n* words $w_1, w_2, ..., w_n$, its polarity score is defined as the mean of polarity scores of all the terms. The polarity function is calculated by using the SentiWordNet[1] lexicon.

2) Emoticons: Emotional signals are any information that could be correlated with sentiment polarity of a sentence. Recently in social media, users adopt visual cues that are strongly associated with their emotional states. These cues, known as emoticons (or facial expressions), are widely used to show the emotion that a user's post represents. In order to integrate these features in our model, we have built a weighted list of common emoticons (including the recent emoji list), and for each tweet we calculate the average polarity of these emoticons, if any, in the sentence. Our list comprises about 300 emoticons and *emojis* with a sentiment level associated among the values: extremely negative, negative, neutral, positive, extremely-positive. Emoticon based polarity of a tweet is computed as the mean of polarity scores of all the emoticons found in the message.

D. Social features (F_{soc})

These features are related to social behavior and their peculiarities are strictly related to the social platform analyzed. Using only features extracted from a post itself to detect harassment could be not sufficient. Sometimes it's necessary to look at the context of a post to have a better understanding of the meaning. Often, users who are familiar with each other tend to communicate in a very informal way also adopting bad words or some terms that can appear to be harassing. So it's really important to take care about the general social behavior of a user (victim or bully), to better detect eventual bully traces in posts, avoiding false positive.

1) Direct User Tagging: A cyber bullying sentence typically is a direct harassment to a user, especially on a social network platform. On Twitter this can be easily detected by looking at the presence of the special @USER tag. In our model this feature is added to check if a tweet contains a direct addressing to someone.

2) Author profiling: This feature measures the politeness of the author of posts. As previously stated, some users, mainly teenagers, adopt as a standard way of communication the use of bad words and apparently offending slang utterances. Our model tries to reflect this behavior to avoid misleading posts.

3) Messages exchanged with a user: This feature tries to gain information about an eventually pre-existent discussion to which the current post analyzed belongs. In these cases having

[1] http://sentiwordnet.isti.cnr.it

an insight into the "history" of the active talk can be useful to determine if the post is a trace of cyber bullying.

V. An Unsupervised Approach

A supervised approach, with manual labelling of dataset, in case of social networks data analysis, potentially leads to a time consuming effort that could be unfeasible in certain scenarios. If we want to detect cyber bullying traces in a huge stream of data, as the ones produced by Twitter, an unsupervised approach can be more effective. The self-organizing map (SOM), also known as Kohonen map [8], is one of the most representative artificial neural network models compliant with the unsupervised learning paradigm.

As a result of the training process, similar input data are mapped in neighboring regions of the map. In the case of sentences classification, similar texts (with similar features) are grouped together. The general idea is then to display similar tweets in similar region of the map.

VI. Growing Hierarchical Self Organizing Map

One of the short comings of the SOM is its fixed architecture that must be initially defined. Dynamically growing variants of the SOM tend instead to produce big maps that are hard to manage. This has led to the development of the GHSOM [3], a dynamic architecture able to grow in a hierarchical way according to the data distribution, allowing a hierarchical decomposition and adapting itself to the requirements of the input space. GHSOM networks are well suited in the case of large collection of documents that needs to be classified.

The basic principle of the GHSOM is to use a hierarchical structure of multiple layers, where each layer consists of a number of independent SOMs. A single SOM is used at the root layer. For every unit in this map a SOM might be added to the next layer of the hierarchy. An example of a GHSOM architecture with 3 layers is showed in Figure 1. The training process of a GHSOM starts with a small map of 2x2 units at the first layer which is organized according to the standard SOM training algorithm [8]. Each single map is then expanded in order to represent the corresponding subset of data at the specific level of granularity.

VII. Methodology and Results

Automatic solutions related to cyber bullying detection are not properly studied in the past. This is one of the main reason for which there exists insufficient training datasets available. Some datasets are available instead on general sentiment analysis and all of them are used in supervised approaches. Although bullying messages are posted every day compared to hundreds of thousands of messages posted every second, they are very sparse. Collecting enough training data is an actual big challenge since random sampling will lead only to few bully messages. We selected two distinct datasets, recently published, related to the social network FormSpring.me and YouTube.

Our experiments are based on a GHSOM network, with a final grid of 50 x 50 neurons, and about 20 hand-crafted features as input layer, as described above. All the input vectors have been normalized between 0 and 1. The network has been trained with a learning rate of 0.7, and with 10.000 epochs. We have implemented the GHSOM network algorithm using the SOMToolbox[2] framework. Some pre-processing has been applied to the datasets, as stop words removal, punctuation removal and stemming, consistently with the original experimental protocols. To measure the goodness of generated clusters we use standard classification measures like *Precision*, *Accuracy*, *Recall* and *F-Measure*, an harmonized mean of precision and recall, also called F1 score. In all our tests we adopted K-fold cross validation. We trained and tested our GHSOM network respect to a K-folded dataset, applying a K-fold partitioning of data, with $K = 10$. Then the average error across all K trials has been computed. We have also applied a "random under sampling technique" that decreases the majority class so as to match the minority class, in order to balance the classes distribution.

A. Results on Formspring.me dataset

In our first experiment we considered the dataset from Kontostathis et al. [6], collected from the social network Formspring.me, from September 2011 to July 2012. Formspring.me is a question-and-answer platform where users invite others to ask and answer questions. Anonymity guaranteed by this platform makes it a fertile ground for episodes of cyber bullying. The data were manually labeled using Amazon Mechanical Turk, where 3 experts manually annotated about 13.000 questions and related answers, with an average of 6% of bullying posts. Totally, the dataset includes 20.921 questions and answers. We considered records having at least 3 positive annotations, i.e. records that can be labelled with a good certainty as bully traces. The dataset contains at the end 1.239 positive classes, and 19.682 negative classes. The dataset is clearly unbalanced for the training stage, so, as stated before, we choose to subsample data adopting a random subsampling technique, and providing a final balanced dataset with about 1.200 posts (with 650 negative classes and 550 positive classes). So, at each round of our K-fold cross validation ($K = 10$) we have about 1.100 sentences used as train dataset and about 100 sentence used as test dataset.

In fig. 2, we show examples of clusters identified by a K-means algorithm applied to generated network lattice, during a test run. We choose to evaluate a single cluster (the biggest one) that can be automatically discovered. We evaluate here the results of the GHSOM, superimposing on the winners neurons the supervised classes provided by the dataset. It's possible to see in fig. 3 how the dataset is topologically mapped into clusters, and how some of these clusters seem to be quite homogeneous in terms of associated classes. Clustering algorithms have been thus applied to see to what extent its possible to automatically detect these clusters in order to classify the input dataset.

[2]www.ifs.tuwien.ac.at/dm/somtoolbox

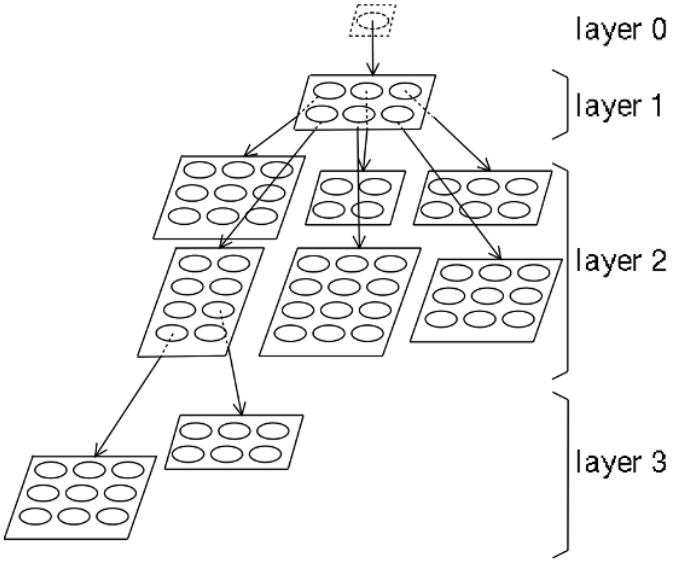

Fig. 1. Architecture of a trained GHSOM. [3]

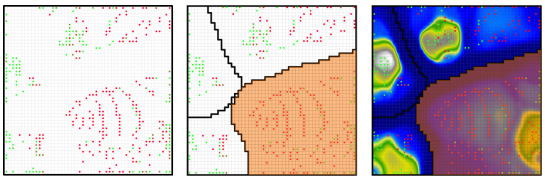

Fig. 2. Class distribution (left) over the GHSOM with K-means clustering (center) and smoothed density map (right). Red dots mean cyber bully traces while green dots mean non bully traces (FormSpring.me dataset).

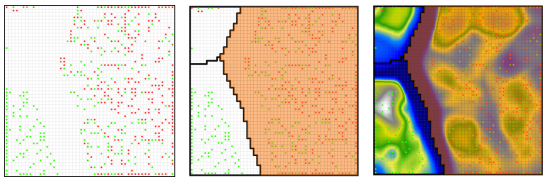

Fig. 3. Class distribution (left) over the GHSOM winners neurons with related K-means clustering (center) and smoothed density map (right). Red dots mean cyber bully traces while green dots mean non bully traces (Youtube dataset).

TABLE I
RESULTS OBTAINED ON FORMSPRING.ME DATASET

Precision	Accuracy	Recall	F1	Method
0.72	0.73	0.69	0.71	GHSOM
0.60	-	0.40	-	C4.5
-	-	0.67	-	SVM

TABLE II
AVERAGE RESULTS OBTAINED ON YOUTUBE DATASET.

Precision	Accuracy	Recall	F1	Method
0.60	0.69	0.94	0.74	GHSOM

Finally these results can be compared to the result provided by Kontostathis et al. [6] according to which the average precision across all documents is 47.7%. We can see that our unsupervised approach, together with the proposed set of features, performs reasonably well respect to the values of the supervised C4.5 decision tree [6], applied on a dataset size of 1.000 records. For completeness, we report that a true positive rate of 78.5% has been reached in [11] when the positive posts were overrepresented in the training dataset.

B. Results on Youtube dataset

We tested our model also on the dataset referred by the paper from Dadvar et al. [9]. The dataset represents 3.462 comments from YouTube crawled in 2012-2013. The history of activity of users are collected over 4 months (April-June 2012), together with profile information. A total of about 54.000 manually annotated comments over YouTube has been analyzed. Finally, the dataset has about 3.045 posts, where 419 posts contain cyber bullying traces. The dataset is unbalanced and so we adopted the same previous technique based on random under sampling. Also for this dataset we applied K-fold cross validation with K=10. For testing data we have 450 negative classes and 369 positive classes. Authors adopted a set of 11 features in three different categories and three machine learning methods, which use labelled training data: a Naive Bayes classifier, a classifier based on decision trees (C4.5) and Support Vector Machines (SVM) with a linear kernel, obtaining at the best 0,72 AUC score (ROC curve) with hybrid Naive Bayes classifier. Our results, based on unsupervised approach are shown visually in fig. 3. The precision, accuracy, recall and F1 score are shown in table II.

Some considerations must be done on the different results obtained for the YouTube dataset. First, user comments on YouTube are longer than the one usually posted in Form-

TABLE III
AVERAGE RESULTS OBTAINED ON TWITTER DATASET.

Precision	Accuracy	Recall	F1	Method
0.81	0.72	0.26	0.4	GHSOM
-	0.67	-	-	Naive Bayes

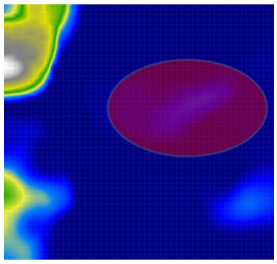

Fig. 4. Density of the BMU for the a real twitter data set phase using a smoothed data histogram.

Spring. Therefore textual analysis and some related syntactical features perform differently on this dataset. According to our tests, in the YouTube case, it is possible to observe a more efficient clustering of no bully posts, with respect to the FormSpring case. Clustering results for the bully posts have a good recall value, but a lower precision result respect to the FormSpring test.

C. Twitter dataset

Twitter allows users to write a maximum of 140 characters, so people are forced to share only essential information. This limitation (short documents) represents a big challenge in the text analysis since the most common adopted techniques perform better with long documents. Twitter users commonly adopt an informal language in posting messages, with many slang words and acronyms, also due to the limitation imposed by the platform. We tested our unsupervised model with the Twitter dataset, adopted in [12], where authors, using a Naive Bayes classifier, reached an accuracy of 67.3% (see table III). Here weak results in recall and F1 score can be attributed to the dataset origin, that is much more related to general sentiment analysis tasks rather than cyber bullying.

For the Twitter case we also conducted a qualitative test. We reused a previous trained GHSOM, and tested it to see which are the performances of the proposed model in classifying a real Twitter stream of data. We collect 1.000 tweets during the summer of 2015, without any filtering, except for the language (English). In figure 4 we can see the smoothed data histogram of the network, showing the density (clearer areas) of BMU (Best Matching Unit). The real test dataset is unbalanced (statistically the percentage of cyber bullying posts in a twitter stream is between the 4% and 7% of the whole dataset). In order to visually see which are the identified clusters discovered by our model, we can observe the density map in fig. 4.

It's possible to verify how the tweets listed in table 3 can be considered possible candidates to bully traces. We expect

TABLE IV
EXCERPT OF TWEETS BELONGING TO CLUSTER IDENTIFIED IN FIG. 4.

Tweets sample in identified cluster (real test stream)
@*** wtf u literally post pics of u s**tting
@*** you're so fu**ing annoying lmao
@*** I'm out, I'm done wasting time with yo bit** ass

at least an error rate proportional to the results obtained in the training stage.

VIII. CONCLUSIONS

We proposed to adopt an unsupervised approach to detect cyber bully traces over social networks, based on Growing Hierarchical Self Organizing Map. Our model comprises several hand crafted features that are used to catch semantic and syntactic communicational behavior of potential cyber bullies. We conducted some experiments on datasets taken from literature, like those coming from FormSpring and YouTube platforms, and also on a real data stream, collected from Twitter. Results indicate that our model achieves reasonable performance and could be usefully applied to build concrete monitoring applications to mitigate the heavy social problem of cyber bullying. Indeed, there is plenty of room for improvement on these techniques (as sarcasm identification) in order to achieve better results.

REFERENCES

[1] Patchin, J., Hinduja, S. "Bullies move beyond the schoolyard; a preliminary look at cyberbullying.". Youth violence and juvenile justice. 4:2 (2006). 148-169.

[2] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0". In Proceedings of CAW2.0 Workshop, 2009.

[3] A. Rauber, D. Merkl, and M. Dittenbach: "The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data". IEEE Transactions on Neural Networks, IEEE 2002.

[4] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying,". MIT. International Conference on Weblog and Social Media. Barcelona, Spain, 2011.

[5] Xu, Jun-Ming; Kwang-Sung Jun; Xiaojin Zhu; and Amy Bellmore. Learning from bullying traces in social media. In Proceedings of the NAACL HLT Conference, Montreal, Canada, 2012, pp.656-666.

[6] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards. 2013. "Detecting cyberbullying: query terms and techniques". In Proceedings of the 5th WebSci 2013. ACM, New York.

[7] M. Dadvar and F. de Jong. 2012. "Cyberbullying detection: a step toward a safer internet yard". In Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion). ACM, New York, NY, USA, 121-126.

[8] Kohonen, T. 1995. Self-organizing maps. Series in Information Science, Vol. 30 Springer-Verlag, Berlin.

[9] M. Dadvar, R.B. Trieschnigg and F.M.G. de Jong. "Experts and Machines Against Bullies: A Hybrid Approach to Detect Cyberbullies". In 27th Canadian Conference on Artificial Intelligence, University of Waterloo, Montral, Canada, 2014.

[10] D. Terrana, A. Augello, G. Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream" in Semantic Computing (ICSC), 2014 IEEE International Conference, pp.128-134, 16-18 June 2014.

[11] Reynolds, K. and Kontostathis, A. and Edwards, L., "Using Machine Learning to Detect Cyberbullying", in Proceedings of International Conference on Machine Learning and Applications and Workshops, ICMLA, pp. 241-244,2011.

[12] H. Sanchez, S. Kumar. "Twitter bullying detection". ser. NSDI 12, 1515, 2011.