Antisocial Behavior Corpus for Harmful Language Detection

Myriam Munezero[1] Maxim Mozgovoy[2] Tuomo Kakkonen[1] Vitaly Klyuev[2] Erkki Sutinen[1]

[1]School of Computing
University of Eastern Finland
P.O.Box 111, FI-80101
Joensuu, Finland
Email: {mmunez, tkakkone, sutinen}@cs.joensuu.fi

[2]The University of Aizu
Tsuruga, Ikki-machi,
Aizu-Wakamatsu, Fukushima
965-8580 Japan
Email: {mozgovoy, vkluev}@u-aizu.ac.jp

Abstract— **We report on experiments that demonstrate the relevance of our *AntiSocial Behavior* (ASB) corpus as a machine learning resource to detect antisocial behavior from text. We first describe the corpus and then, by using the corpus for training machine learning algorithms, we build a set of binary classifiers. Experimental evaluations revealed that classifiers built based on the ASB corpus produce reliable classification results with up to 98% accuracy. We believe that the dataset will be valuable to researchers and practitioners working in preventing, controlling and diagnosing antisocial behavior and related problems.**

I. Introduction

’What is said’ is important and can reveal a lot about a person’s thoughts, emotions and behavior. It particularly is important, when what is said, expresses feelings or thoughts of harming another. As Biber [1] points out, a writer’s thoughts, opinions and attitudes about a topic can be explicitly or implicitly expressed through the choice of word and grammatical constructions. Due to the proliferation of the Internet and Web 2.0, written information on how people feel and their plans and interests is more readily available to researchers studying natural language.

The feelings and actions of harming other human beings can be considered as manifestations of *antisocial behavior* (ASB). ASB is broadly defined as any unconsidered action taken against individuals or groups of individuals that may cause harm or distress to society [2]. Often individuals involved in ASB have disclosed in advance their emotions and plans through oral or written language [3]. Reputedly, the Internet has been used as the outlet for the expression of such emotional states and / or plans of violent acts through the use of blogs or video sites [4]. Moreover, online communication is often used as a way of shouting out their intentions before engaging in their acts of violence [5].

The wealth of antisocial and criminal activity taking place on the Web has resulted in a surge of research interest in the automatic detection of this negative and destructive content. Being able to automatically detect negative content is beneficial, for instance, to managers of websites that allow users to post content or as part of an early warning system to authorities on possible threats to public safety. The automatic detection of ASB could also give rise to self-awareness systems for the individuals that are expressing thoughts or emotions related to ASB.

Identifying the individuals who pose danger to a community involves collecting and analyzing information pertaining to their attitude, thoughts on violence, descriptions of criminal activity and threats among others, including information about homicidal or suicidal ideation [6]. However this information is often difficult to obtain. Reasons such as privacy, legality of the often sensitive information, affect its availability to researchers for analysis

Hence, albeit the problems antisocial behavior causes, there still does not exist a publicly available corpus of ASB texts. However, research projects that focus on ASB and that have been motivated, for instance by the occurrence of school shootings, require a domain relevant corpus for learning linguistic features that may be used for recognizing future risks of antisocial and destructive behavior from texts.

In this paper, we present such a collection of documents, aimed to remedy the situation. To our knowledge, this is the first attempt to build a corpus with a wide variety of types of antisocial, criminal and extremist content; the previous works have concentrated on a single type of antisocial content such as cyberbullying [7] or forms of extremism [8].

Furthermore, we use our corpus to address the problem of detecting ASB for texts by applying *machine learning* (ML) and text mining techniques. We train ML algorithms with positive examples obtained from the ASB corpus, and with negative examples of antisocial behavior collected from the ISEAR [9], Movie reviews [10] and Wikipedia [11] corpora.

Our experimental results show that classification based on content features discriminates ASB texts from non-ASB texts with accuracy up to 98%. Thus we demon-

This work was supported by: 1) “Detecting and visualizing emotions and their changes in Text” project, No.14166, funded by the Academy of Finland; 2) JSPS KAKENHI Grant Number 25330410.

strate that the ASB corpus can serve as a valuable resource for an ongoing antisocial behavior research.

II. Related Work

While the detection of spam in e-mail messages and web content dates back to the early days of the Internet, detection of antisocial content is a new and emerging area of research interest. The methods applied in the detection of antisocial content draw from the ones developed for detecting spam. In discussing related work, as no previous general models for detecting antisocial behavior from text exist, we provide an overview of the work done in the context of detecting cyberbullying, terrorism and criminal behavior.

Perhaps the most notable related work has been carried out in a research project entitled “Intelligent information system supporting observation, searching and detection for security of citizens in urban environment” (INDECT) [12]. The project aimed at automatic detection of terroristic threats and recognition of serious criminal behavior or violence based on multi-media content. Within the context of INDECT, criminal behavior as “behavior related to terrorist acts, serious criminal activities or criminal activities in the Internet”.

Our work differs from the one done in the INDECT project in the focus of the research. While INDECT aims at using the analysis of images, video, and text, our focus is on the analysis of text data.

In their cyberbullying study, Dinakar et al. [7] made use of YouTube comments that involved sensitive topics related to race and culture, sexuality and intelligence. Moreover, Yin et al. [13] in their research made use of online forums for detecting online harassment. Bogdanova et al. [14] in their cyberpedophilia research made use of online perverted journal texts on which to learn models to discriminated pedophiles from non-pedophiles.

Thus, although the corpora used in the studies reported above contain negative behaviors, no corpus has yet addressed the more broad antisocial behavior characterized by covert and overt hostility and intentional aggression toward others [15].

III. Corpora

Textual data is required for analyzing what is said, thought or felt in texts. Unfortunately, when it comes to analyzing antisocial behavior, a suitable text collection is difficult to find. Many of the document collections, for example, those from YouTube and MySpace are generic collections and need to be filtered according to the research area.

It was because of the difficulty and lack of a domain-relevant corpus that we sought to create our own. The corpus can further drive the study of linguistic patterns and emotional content present in ASB texts.

The following subsections describe each corpus used in the experimental study. As we are firstly concerned with the binary classification analysis (that is either a document is deemed as being antisocial behavior or it is not), we therefore collected both positive (Subsection A) and negative (Subsections B-D) examples of non-antisocial behavior texts. To obtain the negative examples of antisocial behavioral, we used popular sentiment corpus (movie reviews [10]), emotion annotated corpus (ISEAR [9]) and factual Wikipedia texts extracts [11]. Table 1 summarizes the documents collected.

A. Antisocial Behavior Corpus

As part of a bigger project that involves detecting antisocial behavior from text, we have created a corpus of aggressive, violent, and hostile texts[1]. Two researchers searched online content in order to collect the documents from various blog posts and news-websites which they could conclusively identify as being ASB. In total 148 documents were identified as ASB. The collection is all English texts, having topics such as: serial killer manifestos, antisocial texts, terrorism, violence-based texts, and suicide notes.

Importantly, the messages in these documents are reflective of the author's thoughts and emotions. The corpus was collected specifically for the purpose of detecting antisocial behavior, conflict, crime and violence behavior from text documents. The collection is based on the research on antisocial behavior that has shown that aggression, violence, hostility, and lack of empathy are among the traits that are most directly associated with ASB [16], [17]. Antisocial behavior also has strong links to negative emotions, such as anger, frustration, arrogance, shame, anxiety, depression, sadness and fear [18]. The link of emotions to antisocial behavior will guide our future research.

B. International Survey on Emotion Antecedents and Reactions (ISEAR)

The ISEAR corpus is a collection of student reports on situations in which the respondents felt any of the seven major emotions: joy, fear, anger, sadness, disgust, shame, and guilt. The responses include descriptions of how they appraised the situation and how they reacted [9].

C. Movie Reviews

This collection consists of 2000 movie reviews. They are labeled in respect to their polarity: negative and positive. The corpus was first used in [10], and now is often applied in sentiment analysis and opinion mining research as a standard development and test set.

[1] The current work-in-progress version of the corpus is available upon request.

D. Wikipedia Text Extracts

We searched and collected Wikipedia articles by using similar concepts such as those we found to be characteristic ASB: killing, terror, violence, aggression, and frustration. The aim of including these texts was to observe how well our classification algorithms could distinguish between antisocial behavior texts and informative texts containing similar keywords.

TABLE I
CORPORA DESCRIPTION WITH SOURCE, NUMBER OF FILES AND AVERAGE FILES SIZE

Corpus	Source	Documents	Avg. File Size (characters)
ASB	blog posts	148	680
ISEAR	[9]	265	110
Movie reviews	[19]	178	390
Wikipedia extracts	[11]	212	680
Total		803	

IV. EXPERIMENTAL SETUP

In order to test the corpus, we approached the ASB detection problem as a classification task. We performed the step-by-step process outlined in Figure 1.

A. Preprocessing Data

We processed each collected online entry or blog post as a whole. That is we assigned the whole text or message as being antisocial behavior or not. From the corpora, we have two fields; a text field consisting of the message and a binary class label (1 = antisocial behavior, 0 = non-antisocial behavior).

The message field needs to be preprocessed because it contains unstructured text. We applied further preprocessing using WEKA utility (StringToWordVector) that performs tokenization, stemming, and stop/frequent word removal.

B. Machine Learning-Based Classification

For classifying the documents into the two classes, we experimented with three supervised ML classifiers: Naïve Bayes Multinomial, SMO for the implementation of Support Vector Machines, and J48 for Decision Trees. The three selected algorithms have shown to be effective in various text classification studies. We made use of the WEKA tool for the above classifiers.

As a first experiment with the corpus, we used a Vector Space Model approach so as to consider the words as independent entities. The model makes an implicit assumption that the order of words in document does not matter, also called the Bag-of-Words (BoW) assumption [20]. The approach is sufficient for the classification task, as the collection of words appearing in the document (in any order) is usually sufficient to differentiate between semantic concepts [20]. Each document in the corpora was represented as a feature vector composed of binary attributes for each word that occurs in the file.

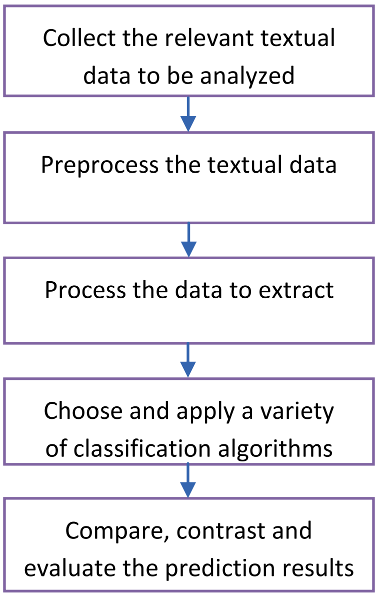
Collect the relevant textual data to be analyzed
Preprocess the textual data
Process the data to extract
Choose and apply a variety of classification algorithms
Compare, contrast and evaluate the prediction results

Fig 1. Process map. (Adapted from [20])

Let $\{f_1, \ldots, f_m\}$ be a predefined set of m features that can appear in a document. Let $n_i(d)$ be the number of times f_i occurs in a document d. Then each document d is represented by the document vector $d:=(n_1(d), n_2(d),\ldots,n_m(d))$ [10]. If a word appears in a given file, its corresponding attribute is set to 1, otherwise it is set to 0. Generally, the BoW approach works well for text classification. However, it does not take into consideration any semantic and contextual information.

Moreover, in order to reduce the number of words in the BOW representation we used the LovinsStemmer in order to replace each word by its stem.

We experimented with the three classifiers:

Multinomial Näive Bayes (NBM). With the Näive Bayes classifier, the input is assumed to be independent. The NB classifier, given the data estimates the probability of a class which is proportional to the probability of the class times the probability the data given the class [20]. In other words, the NB classifier assigns a given document d the class $c^* = \text{argmax}_c\, P(c|d)$ [10]. We used the Multinomial Näive Bayes classifier implemented in WEKA, which uses a multinomial distribution for each of the features.

Support Vector Machine (SVM). The classification method of SVM is based on the maximum margin hyperplane rather than probabilities as the Naïve Bayes [20]. In particular, the SVM classifier in a binary classification case aims to find a hyperplane, represented by a vector that maximally separates the document vectors in one class from those in the other [10].

J48 Decision Tree (J48). This classifier is an implementation of the C4.5 decision tree in WEKA. Decision trees are predictive machine models that are used for classification tasks by starting at the root of tree and moving through it until a leaf is encountered [21]. The decision tree is built from the input training data using the property of information gain or entropy to build and divide nodes of the decision tree in a manner that best represents the training data and the feature vector [7].

The evaluation of the classifiers is discussed in the next section.

V. RESULTS

For an exploratory purpose, we conducted four experiments using the ASB corpus for classifying emotional sentences.

We made use of three corpora as negative examples of ASB: ISEAR, Movie reviews, and Wikipedia extracts as described in Subsection B together with the positive examples of ASB to train supervised ML algorithms. In the first experiment, binary classifiers using the three algorithms were trained on ASB+ISEAR, in the second on ASB+Movie reviews, and in the third on ASB+Wikipedia extracts. Finally, all the corpora were combined.

TABLE II.
RESULTS FOR THE ASB CORPORA USING THE ACCURACY RATE (%)

Corpora	Classifier	Accuracy	Precision	Recall	F-measure
ASB + ISEAR	NBM	**94.91**	**0.95**	**0.94**	**0.95**
	SMO	93.94	0.94	0.93	0.93
	J48	87.89	0.87	0.87	0.87
	Baseline	64.16	0.41	0.64	0.50
ASB + Movie Review	NBM	**98.61**	**0.98**	**0.98**	**0.98**
	SMO	95.83	0.95	0.95	0.95
	J48	90.27	0.90	0.90	0.90
	Baseline	58.88	0.34	0.58	0.43
ASB + Wikipedia	NBM	95.15	0.95	0.95	0,95
	SMO	**95.64**	**0.95**	**0.95**	**0.95**
	J48	88.13	0.88	0.88	0.88
	Baseline	64.16	0.41	0.64	0.50
All	NBM	94.82	0.81	0.93	0.87
	SMO	**96.46**	**0.96**	**0.96**	**0.96**
	J48	92.92	0.92	0.92	0.92
	Baseline	81.31	0.66	0.81	0.72

The performance of the classifiers was then compared in terms of accuracy, precision, recall and F-measure. For baseline values, we made use of the ZeroR classifier from WEKA which classifies data into the most frequent class in the training set. We made use of ten-fold cross validation whereby samples of data are randomly drawn for analysis and the classification algorithm then computes predicted values [20]. Table 2 displays the average of the ten-fold cross validation results on the corpora for each of the ML techniques.

Based on the results shown in Table 2, the ML algorithms NBM, SMO, and J48 clearly surpass the baseline performance. They further show that for our experiments the NBM and SMO algorithms have the highest accuracy rates. The use of the global corpus (All) also resulted in high accuracy results, as it contains heterogeneous data, however, the difference between the SMO accuracy results and the baseline is much lower. With the global corpus, SMO is statistically better than the next-best classifier (NBM) with a confidence level of about 96% based on the accuracy rate. Its F-measure (0.96), a function of both precision and recall, further indicates a high accuracy.

The experimental results illustrate that from our collected corpus, we can successfully classify antisocial behavior type of texts.

VI. CONCLUSION AND FUTURE WORK

In this paper, we applied text classification techniques for the detection of antisocial behavior. In order to accomplish our task we applied various classification algorithms.

Our experimental results show that the task can be successfully accomplished. Experiments show that we achieve high accuracy using Naïve Bayes Multinomial and SMO.

In this paper we have used individual words as features without any additional syntactic or semantic knowledge. In future we are planning to incorporate emotion related information that may positively affect the accuracy of the task.

Ideally, text mining techniques are applied to corpora containing thousands or even millions of documents. In this case, fewer than 200 records were used that could be confidently identified as antisocial behavior. For further linguistic pattern analysis, a larger corpus will need to be attained. In order to attain a larger corpus, we will incorporate semi-automated methods that will ensure that each topic in the corpus is sufficiently represented.

With the larger corpus, researchers can identify features such presence of emotions, causal events or linguistic patterns that pertain to ASB which can in turn be used to train machine learning algorithms. The main purpose of the corpus is for it to be used as a machine learning resource

However, despite these limitations, the created corpus proved to be effective in training ML algorithms.

References

[1] D. Biber, *University language: A corpus-based study of spoken and written registers*: John Benjamins Publishing Company, 2006.

[2] R. Card and R. Ward, *The Crime and Disorder Act 1998*. Bristol, England: Jordans, 1998.

[3] M. E. O'Toole, *The school shooter: A threat assessment perspective*. Quantico, Va: Critical Incident Response Group (CIRG), National Center for the Analysis of Violent Crime (NCAVC), FBI Academy, 2000.

[4] S. Crowley, *Finland shocked at fatal shooting.* Available: http://news.bbc.co.uk/1/hi/world/europe/7084045.stm (2013, Mar. 15).

[5] N. Böckler, T. Seeger, P. Sitzer, and W. Heitmeyer, *School Shootings: International Research, Case Studies, and Concepts for Prevention*. Dordrecht: Springer, 2012.

[6] M. Logan, *Case Study: No More Bagpipes. The Threat of the Psychopath.* Available: http://www.fbi.gov/stats-services/publications/law-enforcement-bulletin/july-2012/case-study (2013, Mar. 15).

[7] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *International Conference on Weblog and Social Media-Social Mobile Web Workshop*, 2011.

[8] A. Abbasi, "Affect intensity analysis of dark web forums," in *Intelligence and Security Informatics, 2007 IEEE*: IEEE, 2007, pp. 282–288.

[9] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *Journal of personality and social psychology*, vol. 66, p. 310, 1994.

[10] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*: Association for Computational Linguistics, 2002, pp. 79–86.

[11] Wikimedia Foundation, *Wikipedia: The Free Encyclopedia* (2013, May. 08).

[12] The INDECT Consortium, *XML Data Corpus: Report on Methodology for Collection, Cleaning and Unified Representation of Large Textual Data from Various Sources: News Reports Weblogs Chat.* Available: http://www.indect-project.eu/files/deliverables/public/INDECT_Deliverable_4.1_v20090630a.pdf (2010, Dec. 10).

[13] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," *Proceedings of the Content Analysis in the WEB*, vol. 2, 2009.

[14] D. Bogdanova, P. Rosso, and T. Solorio, "Modelling fixated discourse in chats with cyberpedophiles," in *Proceedings of the Workshop on Computational Approaches to Deception Detection*: Association for Computational Linguistics, 2012, pp. 86–90.

[15] C. Hanrahan, "Antisocial Behavior," in *The Gale encyclopedia of children's health: Infancy through adolescence*, K. M. Krapp and J. Wilson, Eds, Detroit: Thomson Gale, 2005.

[16] D. Clarke, *Pro-Social andAnti-Social Behaviour*. Abingdon, UK: Taylor & Francis, 2003.

[17] W. G. Parrott, *Emotions in social psychology: Essential readings*. Philadelphia: Psychology Press, 2001.

[18] L. J. Cohen, "Neurobiology of Antisociality," in *Neurobiology of exceptionality*, C. Stough, Ed, New York: Kluwer Academic/Plenum Publishers, 2005.

[19] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*: Association for Computational Linguistics, 2004, p. 271.

[20] G. Miner, J. Elder, T. Hill, R. Nisbet, D. Delen, and A. Fast, *Practical text mining and statistical analysis for non-structured text data applications*, 1st ed. Waltham, MA: Academic Press, 2012.

[21] J. R. Quinlan, *C4.5: Programs for machine learning*. San Mateo, Calif: Morgan Kaufmann Publishers, 1993.