Towards automatic detection of antisocial behavior from texts

Myriam Munezero
School of Computing
University of Eastern Finland
mmunez@cs.joensuu.fi

Tuomo Kakkonen
School of Computing
University of Eastern Finland
tuomo.kakkonen@uef.fi

Calkin S. Montero
School of Computing
University of Eastern Finland
calkinm@gmail.com

Abstract

The automatic analysis of emotional content of text has become pervasive and has been applied in many fields of research. The work reported in this paper is in particular interested in modeling antisocial behavior and the emotional states that define it. We introduce the *antisocial behavior detection* (ASBD) model for portraying the emotions pertaining to antisocial behavior. In addition to describing negative affective states, our model uses the concepts of action tendencies and evidences in order to predict possible acts of antisocial behavior based on input texts. We outline a design for an antisocial behavior detection system based on the ASBD model.

1 Introduction

Emotions connect individuals to the social world and, hence, are the triggers of many social psychological phenomena, such as altruism, antisocial behavior, and aggression (Parrot, 2001). To be able to identify and classify a behavior, one has to understand the behavior itself and the emotional states (e.g. happiness, sadness and anger) that pertain to it. This paper focuses on modeling the emotional states that characterize antisocial behavior.

We define antisocial behavior as any unconsidered action against others that may cause harm or distress to society. Antisocial behavior has been linked to disruptive and impulsive behaviors, bullying, and in extreme cases, school shootings (Flory et al, 2007; Sutton et al, 1999; Borum et al, 2010).

Upon reviewing the available data about extreme antisocial behavior, O'Toole (2000) reported that often the individuals involved have disclosed in advance their plans orally or in written form. In particular, the Internet has been used as the outlet for the expression of their emotional states through the use of blogs or video sites (Crowley, 2007). In many cases these troubled people have written and publicly distributed documents over the web in the form of manifestos as a way of shouting out their intentions before engaging in their acts of violence (Web search, Dec 2010). Interestingly, little research has been done regarding the automatic analysis of the media in order to warn the pertinent authorities of the threat.

The aim of our research is the automatic analysis of texts in order to uncover emotions and possible behavioral traits related to antisocial behavior. By analyzing and identifying these specific traits in writings we seek to determine hints of antisocial behavior while the possible acts of violence that may follow are still in their planning stages.

As a cornerstone, this paper introduces our proposed model of emotions for the detection of antisocial behavior from text sources. Section 2 reviews the related previous and ongoing research on antisocial behavior and briefly introduces the circumplex model of emotions. Section 3 outlines our proposed model of emotions and its connection to antisocial behavior. Design of a system for detecting antisocial behavior based on the ASBD model is outlined in Section 4. Conclusions and directions for future work are given in Section 5.

2 Background Work

2.1 Research on antisocial behavior and associated emotions

Antisocial behavior has been substantially researched in the fields of psychology and education (Borum et al, 2010). It can manifest itself or be expressed in different ways; it can range from aggression to verbal abuse, from conduct disord-

Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 20–27,
Chiang Mai, Thailand, November 13, 2011.

er to delinquencies (Foster, 2005). In our work we are interested in the emotional traits of antisocial behavior that can be perceived linguistically in people's writings.

Notably, aggression is the behavioral state that is most directly associated with antisocial behavior (Clarke, 2003). Other types of behavioral states also associated with antisocial behavior include violence, hostility, and lack of empathy. Behavioral states in this paper are considered as a result of emotions. For example, hostile and aggressive inclinations are a result of depression and anger (Parrot, 2001).

Antisocial behavior has also been linked to several negative emotions. Some of these emotions include anger, frustration, arrogance, shame, anxiety, depression, sadness, low levels of fear, and lack of guilt (Cohen, 2005). Many of these emotions have been shown detectable in writings (Gill et al, 2008).

2.2 Previous work on automatic detection of antisocial behavior

Sentiment analysis and opinion mining are established areas of study within the NLP research community and both have received a raising amount of attention over the last decade. Although negative emotions like anger and sadness have been identified in writings, the detection of antisocial behavior from text per se is a new area of research interest. The analyses of texts written by terrorist groups and the automatic detection of criminal behavior have received some attention from the NLP community. While terrorism and crime might be regarded as extreme forms of antisocial behavior, they form a rather narrow sub-part of the whole issue we are dealing with in this work. Nonetheless, as no previous general models for detecting antisocial behavior from text exist, we provide an overview of the work done in the context of terrorism and criminal behavior since they are also a result of negative emotions.

Perhaps the most notable related work is carried out in a research project entitled "Intelligent information system supporting observation, searching and detection for security of citizens in urban environment" (INDECT) (The INDECT consortium, 2009). The project aims at "automatic detection of terroristic threats and recognition of serious criminal ("abnormal") behavior or violence" based on multi-media content. Within context of INDECT, such abnormal behavior is defined as "criminal behavior", and specifically as "behavior related to terrorist acts, serious criminal activities or criminal activities in the Internet".

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

2.3 Circumplex model

While most of the work on sentiment analysis has been done based on the theories of basic emotions, our work however starts from a different view - the circumplex model. Whereas the basic emotions based models (e.g. Ekman, 1992) divide all human emotions into a limited set of discrete and independent categories (such as fear, anger), the circumplex model, first proposed by Russell (1980), asserts that emotions can be characterized in a two-dimensional space: pleasure-displeasure and arousal-sleep. In this model, emotions are seen as a linear combination of the two dimensions rather than judged belonging or not belonging into a specific basic emotion category. This allows for a "fuzzy" characterization of emotions.

Posner et al. (2005), for example, stated that the fact that people have difficulties in assessing their own emotions implies that "individuals do not experience, or recognize, emotions as isolated, discrete entities, but that they rather recognize emotions as ambiguous and overlapping experiences". The circumplex model provides a starting point for the development of our model of emotions. For a full description of the circumplex model see (Russell, 1980; Posner et al, 2005).

3 Model for Detection of Antisocial Behavior from Texts

Based on the relevant literature on antisocial behavior (see Section 2), we developed the ASBD model (Figure 1) that takes into consideration *negative emotional states* (Section 3.1), *action tendencies* (Section 3.2) and *evidence* (Section 3.3) that may lead into those *behavioral states* that are associated with antisocial behavior.

3.1 Circumplex-based model of emotions related to antisocial behavior

The left-hand side of our model, shown in Figure 1, illustrates 14 interplaying emotions (not an exhaustive list) that may lead to antisocial behavior. These discrete emotions are seen in a two dimensional space within the unpleasantness and

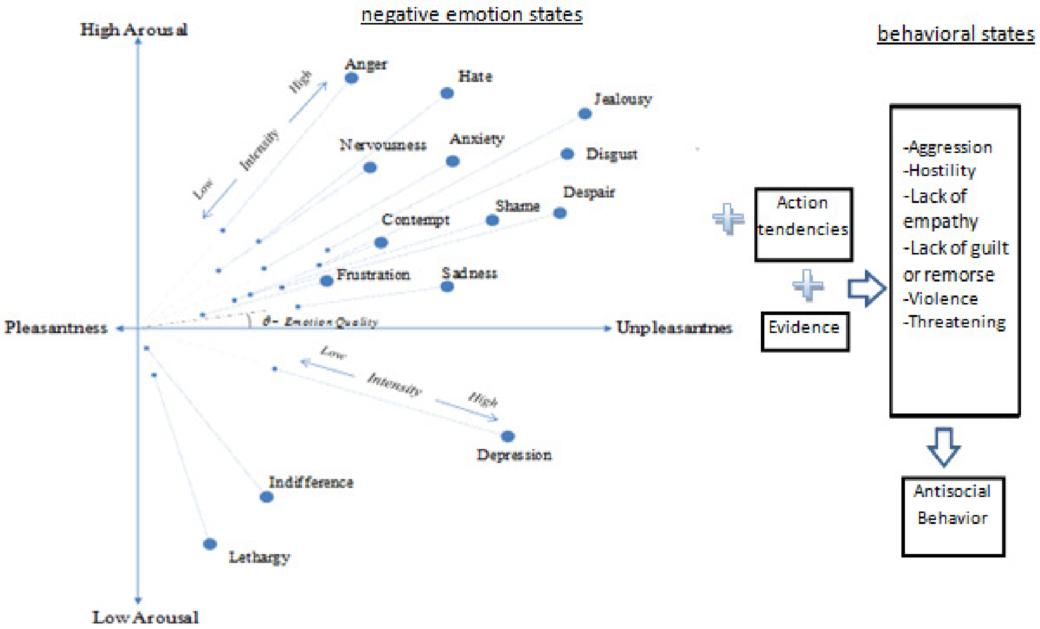

Figure 1. Model of antisocial behavior

arousal dimensions. As previously stated, this spatial representation has been adopted from Russell's 1980) circumplex model of emotions. In our proposed model, each emotion type is placed within the two dimensional space according to its subjective proportion of unpleasantness and arousal as given by Reisenzein (1994) and Russell (1980).

The proportion of unpleasantness and arousal of a specific emotion determines its *emotion quality,* which is represented by the angle between the emotion type and the unpleasantness axis. The conceptual *emotion intensities* used in our model are taken from the works of Reisenzein (1994). Reisenzein demonstrated that the intensity of an emotion can be represented by the distance from a subjective neutral point (hedonic neutrality and medium arousal level) to a point in the space symbolizing that emotion. The subjective neutral point of the space corresponds to a neutral emotional state; a state in which there is no emotion present. The minimum intensity for an emotion (denoted in Figure 1 by a small dot along the line towards the neutral point) is the neutral state for that particular emotion.

3.2 Connecting emotions to behaviors

While the way in which emotions and behaviors are connected has been heavily debated in psychological and social science literature (see, for example Green, 1970; Lyons, 1978; Baumeister et al, 2009) there does not, however, appear to be a lack of consensus that such a connection exists.

Table 1 gives samples of the types of connections that have been reported in research literature between specific emotions and behaviors.

Emotion	Associated behavior	Source
hurt feelings, shame leading to rise of anger	Aggression and conflict escalation	Maiese, 2005
anger, resentment, hatred	Give rise to cycle of violence	
feeling agitated, angry, fearful	"…cause disputes to escalate and sometimes even cause negotiations to break down."	
chronic anger	Endorsement of aggressive solutions, and identification with delinquent peers".	Granic and Butler, 1998
anger	Provides sufficient impetus for the formation of the intention to correct what is perceived as a problem.	Cho and Walton, 2009
frustration	Increased aggression	Verona and Curtin, 2006 Clarke, 2003

Table 1. Emotions and associated behaviors

Regardless of the precise way in which the connection between emotion and behavior occurs in the human brain, in our model we adopt the notion of emotions as influencing or participating in shaping the mind's processes, including those which activate behavior (Russell, 2003).

What can be concluded based on Table 1 is that emotions do influence the motivational state of a person to carry out an action or behavior. We use the notions of action tendency and evidence to model this connection.

Action Tendency

In order to link the emotion to a possible action outcome, the ASBD model supplements the circumplex representation of emotions with Frijda's (1986) concept of *action tendencies* (ATs).

Frijda (1986)' emotion theory associates emotions to a small set of action tendencies (see Table 2.), which are defined as "states of readiness to execute a given kind of action [which] is defined by its end result aimed at or achieved". For example, in the case of negative emotions, reaching the corresponding end state should mitigate its experience (e.g., anger subsides once one believes the object of one's anger has been removed) Frijda (1986). Table 2, provides some examples of ATs.

Table 2, tells us that, for example, in the case of anger, a person is in a "state of readiness" to remove their obstruction. However, ATs should not be mistaken for intentions, while intensions are goal-directed, ATs are stimulus driven (Frijda, 1986). Hence, a person's intentions or manner in which they are planning to carry out the action is only revealed to us through additional information (evidence) in the text.

Emotion	**Function**	**Action tendency**	**End state**
desire	consume	approach	access
anxiety	caution	inhibition	absence of response
anger	control	agonistic	obstruction removed
fear	protection	avoidance	own inaccessibility
disgust	protection	rejecting	object removed

Table 2. Classification of some action tendencies. Adapted from Frijda (1986) (p.88)

Evidence

In addition to using the concept of ATs, we draw from Green's (1970) concept of *evidence* to describe indications of antisocial behavior linked to negative emotions. Green describes evidence as the actions or reactions that a person ordinarily carries out or has when they experience a particular emotion within "appropriate circumstances". Thus, instead of directly connecting antisocial behavior to specific negative emotions we describe those behaviors that a certain emotion might evoke under specific conditions: 'B is the behavior a person is likely to engage in when, among other things, they feel emotion E in C circumstances'.

While various types of circumstances leading to anti-social behavior have been suggested (see for instance the references given in Table 1), we do not believe that it is possible to reliably detect all of them based on a piece of text. Taking Maiese (2005) as an example, she suggests that being angry and fearful may cause disputes to escalate when "a person feels that their interests are threatened". It would be impossible to make judgments about such a condition without having access to detailed information about the situation the author is facing. Such background information is almost impossible to obtain based solely on analyzing few text fragments written by an author.

Thus we confine ourselves to a subset of these circumstances in which we believe we can obtain from text fragments and user profiles. The subset, though not exhaustive includes the following:

- Age: Moffitt (1993) has noted that antisocial behavior is almost ten times more common among adolescents than other age groups.
- Gender: It is commonly accepted that males are more prone to extreme forms of antisocial behavior, such as violence, delinquency and physical aggression (Björkqvist et al, 1992).
- Presence of frustration: Research has revealed a strong link between frustration and antisocial behavior, showing that frustration can lead to extreme manifestations of antisocial behavior such as aggression (Clerk, 2005).

In addition, we expand this concept of evidence to include keywords, such as 'kill', shoot', 'gun', abuse', etc, that are commonly expressed by people exhibiting violent and antisocial behavior.

4 Design of an antisocial behavior detection system based on the proposed model

The ASBD model serves as the backbone for the design of an antisocial behavior system. The system will function as follows. When it receives an input text, it first detects the emotions (quality and intensity) in the text. Next, it resolves the ATs corresponding to each detected emotion. Thirdly, the system identifies the available evidence (both in the input text and from external sources, for instance, user profiles). Finally, the system uses all the collected information to predict the behavioral state connected with the input. Figure 2 illustrates the system design.

As shown in Figure 2, the system consists of three components. The Emotor component combines the circumplex-based model for detecting emotions along with their corresponding ATs.

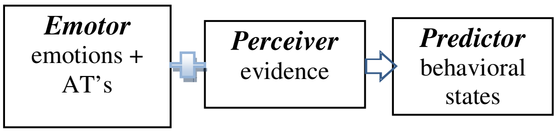

Figure 2. Architecture of an antisocial behavior detection system based on the ASBD model. Adapted from (Frijda and Moffat, 1994)

The Perceiver component collects the pieces of evidence from the input document and other sources. The Predictor component finally combines the information collected by the Emotor and Perceiver in order to predict which behavioral state associated to antisocial behavior might occur.

Figure 3 gives an example of the process of detecting potential anti-social behavior with the proposed system. As illustrated in Figure 3, the Emotor component first automatically detects and analyses the emotion qualities and intensities present in the two input sentences (*s1* and *s2*) based on the circumplex model. It makes use of a supervised classification algorithm developed through a human-annotated corpus to resolve the emotion quality and intensity. Thus s*1* is resolved to be near the emotion 'disgust' and *s2* near the emotion 'anger'. The Emotor component then identifies the ATs connected to these two emotions. s*1* is defined to have the AT 'rejecting' and *s2* 'agonistic' (see Table 2).

The Perceiver collects sets of evidence, such as the writer's gender and age if they are available in the text or the user profile. In addition, it is able to detect the keywords related to various forms of antisocial behavior such as violence, racism and crime. To that end, we are developing an ontology and an ontology-based information extraction tool. The antisocial behavior, conflict and violence (ABCV) ontology currently consists of a 19-class classification system for terms related to antisocial behavior and is capable of detecting a total of 340 terms related to these classes. The predictor finally collects the analysis results from the Emotor and Perceiver, and based on a statistical classification algorithm, it resolves that *s1* could indicate potential hostile behavior and *s2* is showing signs of threatening behavior.

4.1 Data model

A key design issue related to the implementation of the proposed system architecture was the data model. Our main aim was to come up with an extensible data model that is based on a standard. We therefore opted for the *EmotionML* (Emotion Mark-up Language) that was recently introduced by W3C as a working draft standard for representing emotions in text (Baggia et al, 2011).

EmotionML is an XML-based mark-up language that provides a standard interface between components. It defines a set of vocabularies for representing emotion-related states (Schröder and Pelachaud, 2011). EmotionML comes with the vocabulary definition for Frijda's (1986) ATs. This pre-defined set, however, does not support the circumplex-based representation of emotions. EmotionML is flexible enough to allow us to define our own vocabularies depending on the needs of our model and system.

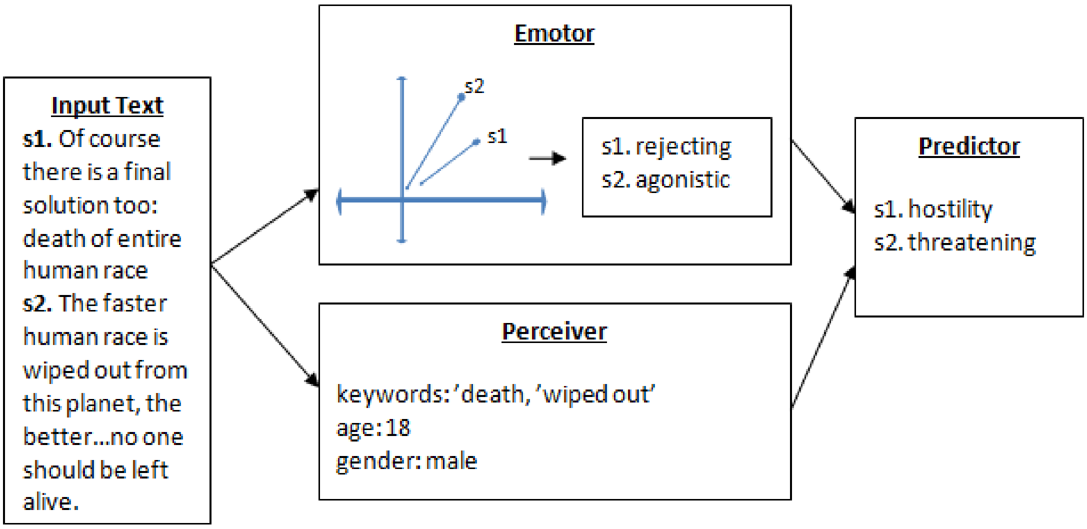
Input Text
s1. Of course there is a final solution too: death of entire human race
s2. The faster human race is wiped out from this planet, the better...no one should be left alive.
Emotor
s2
s1
s1. rejecting
s2. agonistic
Perceiver
keywords: 'death, 'wiped out'
age: 18
gender: male
Predictor
s1. hostility
s2. threatening

Figure 3. Example of antisocial behavior detection process. s1 and s2 are two sample input sentences. Example sentences cited from (OddCulture, 2011).

4.2 Vocabulary definition of emotions and action tendencies

As the Emotor component is representing emotions the circumplex-based model, we need to define a new vocabulary in accordance to the structure of EmotionML. As described above, our emotional model has two values for representing an emotion: quality and intensity. These can be defined in EmotionML as follows:

```
<vocabulary type = "dimension"
        id="cplex">
        <item name="quality" />
        <item name="intensity" />
</vocabulary>
```

Let us assume that we want to describe the emotion 'anger' which has an emotion quality of 81 degrees (when taking the unpleasant axis as 0degrees) and an intensity value of 0.5. The definition for 'anger' would appear in an EmotionML document inside <emotion> tags as:

```
<emotion dimension-set="#cplex">
    <dimension name = "quality"
       val ue="81degrees"/>
    <dimension name = "intensity"
       value = "0.5"/>
</emotion>
```

In addition, the Emotor module annotates the text with the available default set of ATs. The AT for the emotion 'anger' would appear in an EmotionML document as (Ashimura et al, 2011):

```
<emotion dimension-set="#cplex" ac-
tion-tendency-set="#frijda-subset">
    <dimension name = "quality"
       value="81degrees"/>
    <dimension name = "intensity"
       value = "0.5"/>
    <action-tendency name="agonistic"
value="0.9"/>
</emotion>
```

Furthermore, in our XML-based document representation, the Perceptor module annotates the evidence in the text with <evidence> tags.

Vocabulary definition for behavioral states

Whenever the Predictor component receives information from the Emotor and Perceiver component, it analyses and computes the value of a behavioral state. The Predictor component also outputs in EmotionML format. The vocabulary of the behavioral states is defined as follows:

```
<vocabulary type="behavior-state"
id="antisocial-subset">
        <item name="violence"/>
        <item name="aggression"/>
        <item name="hostility"/>
        <item name="threats"/>
</vocabulary>
```

Let us consider an example:

<emotion behavior-state="#antisocial-subset ">
< behavior-state name = "violence" value="0.3"/>
< behavior-state name = aggression" value = "0.5"/>
</emotion>

In addition to the above representations, EmotionML allows us to provide reference information regarding the resolved behavior. For example, if the behavior resolved is ‘hostility’ we can reference the following values:

- Who expressed the behavior (experiencedBy)
- To whom the behavior is directed at (targetedAt).

The <reference> element may occur as a child of the <emotion> element (Baggia et al, 2011).

5 Conclusion and Future Work

We have reviewed the previous research work on antisocial behavior, its defining emotions and automatic detection from texts. We also proposed ASBD, a combined model of negative affect states, ATs, evidence and behavioral states that have been shown to lead to antisocial behavior.

In addition, the paper outlined the architecture of antisocial behavior detection system based on the ASBD model. The system design consists of three modules that communicate with each other using the standard EmotionML markup language. We defined new EmotionML vocabulary sets which pertain to the purpose of our system.

Our future work involves the implementation of the outlined system. The next steps in this work include collecting and annotating a corpus with the proposed annotations and running sentiment detection experiments by applying the circumplex model.

Acknowledgments

This work was supported by the “Detecting and visualizing emotions and their changes in text” project, funded by the Academy of Finland.

References

Kazuyuki Ashimura, Paolo Baggia, Felix Burkhardt, Alessandro Oltramari, Christian Peter, and Enrico Zovato. 2011. Vocabularies for EmotionML. W3C Working Draft 7 April 2011. [Cited: 01.06.2011] www.w3.org/TR/2011/WD-emotion-voc-20110407/

Paolo Baggia, Felix Burkhardt, Catherine Pelachaud, Christian Peter and Enrico Zovato. 2011. Emotion Markup Language (EmotionML) 1.0. W3C Working Draft 7 April 2011. [Cited: 01.06.2011] www.w3.org/TR/2011/WD-emotionml-20110407/

Roy F. Baumeister, Nathan C. DeWall, Kathleen D. Vohs and Jessica L. Alquist. 2009. Does Emotion Cause Behavior (Apart from Making People Do Stupid, Destructive Things)? In Christopher R. Agnew, Donald E. Carlston, William G. Graziano, and Janice R. Kelly (eds.). *Then a Miracle Occurs: Focusing on Behavior in Social Psychological Theory and Research*. 119-137. Oxford University Press, New York, USA.

Kaj Björkqvist, Kirsti M. J. Lagerspetz and Ari Kaukiainen. 1992. Do Girls Manipulate and Boys Fight? Developmental Trends in Regard to Direct and Indirect Aggression. *Aggressive Behavior*, 18(2):117-127.

Randy Borum, Dewey G. Cornell, William Modzeleski and Shane R. Jimerson. 2010. What Can Be Done About School Shooting? *A Review of the Evidence. Educational Researcher*, 39(1):27-37.

Seungho Cho and Laura R. Walton. 2009. Integrating Emotion and the Theory of Planned Behavior to Explain Consumers’ Activism in the Internet. *Institute for Public Relations*, Gainesville, Florida, USA.

David Clarke. 2003. *Pro-social and Anti-social Behavior*. Routledge. New York, USA.

Lisa J. Cohen. 2005. Neurobiology of Antisociality In: C. Stough. (ed.). *Neurobiology of Exceptionality*. Kluver Academic/Plenum Publishers, New York, USA. 107-124.

Sean Crowley. 2007. Finland Shocked at Fatal Shooting. BBC News. [Cited: 10.12.2010] http://news.bbc.co.uk/1/hi/world/europe/7084045.stm

Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition and Emotion*, 6:169-200.

Janine D. Flory, Jeffrey H. Newcorn, Carlin Miller, Seth Harty and Jeffrey M. Halperin. 2007. Serotonergic Function in Children with Attention-Deficit Hyperactivity Disorder: Relationship to Later Antisocial Personality Disorder. *The British Journal of Psychiatry*. 190:410-414.

Sharon L. Foster. 2005, Aggression and Antisocial Behavior in Girls. In Debora Bell, Sharon L. Foster, Eric J. Mash (eds). *Handbook of Behavioral and Emotional Problems in Girls. Issues on Clinical Child Psychology.* Springer, New York, USA. 149-180.

Nico H. Frijda. 1986. *The emotions. Studies in emotion and social interaction.* Cambridge University Press, Cambridge, UK.

Nico H. Frijda, David Moffat. 1994. Modeling emotion. *Cognitive Studies*, 1(2):5-15.

Alastair J. Gill, Robert M. French, Darren Gergle and Jon Oberlander. 2008. The Language of Emotion in Short Blog Texts. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work.* ACM, New York, USA. 299-302.

Isabela Granic and Stephen Butler. 1998. The Relation Between Anger And Antisocial Beliefs In Young Offenders. *Personal Individual Difference*. 24(6):759-765.

Harvey O. Green. 1970. The Expression of Emotion. *Mind*, 79(316):551-568.

William E. Lyons. 1978. Emotions and Behavior. *Philosophy and Phenomenological Research*. 38(3):410-418.

Michelle Maiese. 2005. "Emotions." Beyond Intractability. In Guy Burgess and Heidi Burgess (eds.). *Conflict Research Consortium*, University of Colorado, Boulder.

Terrie E. Moffitt. 1993. Antisocial Behavior: A Developmental Taxonomy. *Psychological Review*, 100(4): 674-701.

OddCulture. The Pekka Eric Auvinen Manifesto. [Cited: 02.06.2011] www.oddculture/weird-news-stories/the-pekka-eric-auvinen-manifesto/

Marry Ellen O'Toole. 2000. School Shooter: A Threat Assessment Perspective. National Center for the Analysis of Violent Crime, Federal Bureau of Investigation, Quantico, Virginia, USA.

Gerrot W. Parrot. 2001. *Emotions in Social Psychology*. Taylor & Francis. Philadelphia, Pennsylvania, USA.

Jonathan Posner, James A. Russell and Bradley S. Peterson. 2005. The Circumplex Model of Affect: An Integrative Approaches to Affective Neuroscience, Cognitive Development and Psychopathology. *Development and Psychopathology*. 17:715-734.

Rainer Reisenzein. 1994. Pleasure-Arousal Theory and the Intensity of Emotions. *Journal of Personality and Social Psychology*, 67(3):525-539.

James A. Russell. 1980. A Circumplex Model of Affect. *Personal and Social Psychology*. 39(6):1161-1178.

James A. Russell. 2003. Core Affect and the Psychological Construction of Emotion. *Psychological Review*. 110(1):145-172.

Jon Sutton, Peter K. Smith and John Swettenham. 1999. Social Cognition and Bullying: Social Inadequacy or Skilled Manipulation? *British Journal of Developmental Psychology*, 17: 435-450.

The INDECT Consortium. 2009. XML Data Corpus: Report on Methodology for Collection, Cleaning and Unified Representation of Large Textual Data from Various Sources: News Reports, Weblogs, Chat. [Cited: 10.12.2010] http://www.indect-project.eu/files/deliverables/public/INDECT_Deliverable_4.1_v20090630a.pdf/view.

Edelyn Verona and John J. Curtin. 2006. Gender Differences in the Negative Affective Priming of Aggressive Behavior. *Emotion*, 6:115-124.

Web Search. "school shooting manifesto". [Cited: 10.12. 2010.].