PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND

Dissertations in Forestry and Natural Sciences

MYRIAM DOUCE MUNEZERO

LEVERAGING EMOTION AND WORD-BASED FEATURES FOR ANTISOCIAL BEHAVIOR DETECTION IN USER-GENERATED CONTENT

UNIVERSITY OF
EASTERN FINLAND

PUBLICATIONS OF THE UNIVERSITY OF EASTERN FINLAND
DISSERTATIONS IN FORESTRY AND NATURAL SCIENCES

N:o 264

Myriam Douce Munezero

LEVERAGING EMOTION AND WORD-BASED FEATURES FOR ANTISOCIAL BEHAVIOR DETECTION IN USER-GENERATED CONTENT

ACADEMIC DISSERTATION

To be presented by the permission of the Faculty of Science and Forestry for public examination in the Louhela Auditorium in Science Park at the University of Eastern Finland, Joensuu, on April 12th, 2017, at 12 o'clock.

University of Eastern Finland
School of Computing
Joensuu 2017

Grano Oy
Jyväskylä, 2017
Editor: Prof. Matti Tedre

Distribution:
University of Eastern Finland Library / Sales of publications
julkaisumyynti@uef.fi
http://www.uef.fi/kirjasto

ISBN: 978-952-61-2463-6 (print)
ISSNL: 1798-5668
ISSN: 1798-5668
ISBN: 978-952-61-2464-3 (pdf)
ISSNL: 1798-5668
ISSN: 1798-5676

Author's address:
University of Eastern Finland
School of Computing
P.O.Box 111
FI-80101 JOENSUU
FINLAND
email: myriam.munezero@uef.fi

Supervisors:
Professor Markku Tukiainen, Ph.D.
University of Eastern Finland
School of Computing
P.O.Box 111
FI-80101 JOENSUU
FINLAND
email: markku.tukiainen@uef.fi

Professor Erkki Sutinen, Ph.D.
University of Turku
Department of Information Technology
FI-20014 TURKU
FINLAND
email: erkki.sutinen@utu.fi

Docent Tuomo Kakkonen, Ph.D.
University of Eastern Finland
School of Computing
P.O.Box 111
FI-80101 JOENSUU
FINLAND
email: tuomo.kakkonen@uef.fi

Calkin Suero Montero, Ph.D.
University of Eastern Finland
School of Computing
P.O.Box 111
FI-80101 JOENSUU
FINLAND
email: calkin.montero@uef.fi

Reviewers:
Associate Professor Tapio Pahikkala, Ph.D.
University of Turku
Department of Information Technology
FI-20014 TURKU
FINLAND
email: tapio.pahikkala@utu.fi

Docent Joel Brynielsson, Ph.D.
Royal Institute of Technology
School of Computer Science and Communication
SE-100 44 STOCKHOLM
SWEDEN
email: joel@kth.se

Opponent: Professor Diana Inkpen
University of Ottawa
School of Electrical Engineering and Computer Science
800 King Edward, Ottawa
K1N 6N5 ONTARIO
CANADA
email: diana@site.uottawa.ca

Myriam Douce Munezero
Leveraging Emotion and Word-based Features for Antisocial Behavior Detection in User-Generated Content
Joensuu: University of Eastern Finland, 2017
Publications of the University of Eastern Finland
Dissertations in Forestry and Natural Sciences

ABSTRACT

Online platforms increasingly provide opportunities to publish user-generated content that expresses emotions, thoughts, and intentions. Some of this content may include emotions, thoughts, and intentions related to antisocial behavior. With the vast amount of user content generated each day, it has become overwhelming and impossible to manually monitor and detect incidents of antisocial behavior. While antisocial behavior, which is harmful to society, has often been studied from educational, social, and psychological points of view, little research has been conducted on computational linguistics and natural language processing regarding antisocial behavior.

To address this gap, this work investigated whether recent advances in natural language processing methods and tools can be used to automatically detect potential instances of antisocial behavior.

This dissertation presents a framework that can be used for the automatic detection of antisocial behavior in text. The framework is based on the emotion and language theories, which provide a comprehensive understanding of antisocial behavior and explain how antisocial behavior, word usage, writing styles, and emotions are connected and are represented in text. The framework leverages word-based and emotion features for the detection of antisocial behavior in user-generated content. Using the features, supervised machine learning models were developed for the automatic detection of antisocial behavior. The effectiveness of the developed models was evaluated based on a set of corpora, one of which - the antisocial behavior corpus - was created by the author of this research and her colleagues. Several of the developed models showed high accuracies of over 90% in their detection of antisocial behavior in text.

In addition, the distinguishing features of antisocial behavior texts, such as a high use of swearing, insults, and negative emotions, including anger, were identified and analyzed, thus allowing for an improved understanding of antisocial behavior in user-generated content.

In sum, this study has demonstrated the potential of utilizing natural language processing techniques for antisocial behavior detection. Continued research on the relationships between natural language use and public security concerns as well as multidisciplinary efforts to develop models that can accurately predict harmful behavior are required to extend this research.

Universal Decimal Classification: *004.773, 004.85, 159.942, 316.613.43, 316.624.3, 81'322, 81'322.2*
Library of Congress Subject Headings: *Natural language processing (Computer science), Computational linguistics, Artificial intelligence, Machine learning, Emotions – Analysis, Emotions – Detection, Emotion recognition, Social media*

Free keywords: *antisocial behavior, emotion analysis*
Yleinen suomalainen asiasanasto: *tietokonelingvistiikka, tekoäly, koneoppiminen, luonnollinen kieli, prosessointi, sosiaalinen media, tunteet, tunnistaminen, analyysi, havainnointi, poikkeava käyttäytyminen, käyttäytymishäiriöt*

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Professor Erkki Sutinen, Dr Tuomo Kakkonen, and Dr Calkin Suero Montero for taking a chance on me to work in the project "Detecting and visualizing changes in emotions in text" funded by the Academy of Finland. This project marked the beginning of the research work. Many thanks Erkki for your constant support, belief in me, and for providing an interesting and innovative environment for discussions and idea generation. Thank you Tuomo for your guidance and expert advice in what was a new field for me. Many thanks Calkin for your mentorship, guidance, and inspiration along this rather long journey. Thank you Professor Markku Tukianen for being my supervisor and helping me to finish the dissertation. To all my supervisors, your feedback, suggestions, and advice was always appreciated.

I would like to also thank my dissertation committee members, Professor Tapio Pahikkala and Dr Joel Brynielsson for taking the time to read my thesis and provide insightful and constructive feedback, and thank you Professor Diana Inkpen for agreeing to act as my opponent.

Moreover, this work could not have started nor been completed without the generous funding from the Academy of Finland during the project "Detecting and visualizing emotions and their changes in text", project No. 14166. I would also like to thank the JSPS KAKENHI Grant Number 25330410, for their contribution towards my research visit to the University of Aizu, Japan, which led to great collaboration and co-authorship opportunities.

I would also like to thank all the people who have helped me ideate on parts of my dissertation and those who have directly worked with me on publishing and implementing those ideas. These include Maxim, Vitaly, John, Carolina, Harri, Tommi, Victor, Ehsan, and Tersia. In addition, I would also like to thank Jarkko Suhonen, Eva, and Tarja who have made many opportunities possible for me at UEF.

Finally, through the many years of this research, this work would not have come to fruition without the support, belief, and encouragement from my colleagues, friends, and dear family. A special thank you goes to my husband. Thank you love, for your patience, wise words, and encouragement.

"If it doesn't challenge you, it won't change you"

Joensuu, April 12, 2017

Myriam Douce Munezero

TABLE OF CONTENTS

1 INTRODUCTION

Emotions are what make people special and give us a reason for living. - Rousseau

With the rise of Web 2.0 and social media, the Internet has become a vast repository of textual content, particularly, *user-generated content* (UGC). Due to the simplicity and low cost of publication, the Internet has made it easier for users to share their personal stories, life experiences, and opinions regarding various events, thus making the shared content more personalized and subjective. Most UGC on the Internet is benevolent or harmless; the only intention is to share something the author believes to be interesting to other people, such as an opinion regarding a product or a movie.

While UGC has certainly enriched online communication, it has also introduced challenges and threats. Like all Internet users, persons with emotional or social problems can easily locate one another online. Web 2.0 and social media encourage group formations based on common interests and behavioral patterns, which could lead to positive outcomes or could pose threats to the well-being of other persons as well as society in general. For example, this is illustrated by the emergence of online discussions that are characterized by aggressive or otherwise strongly emotional textual content [1,2]. Thus, UGC can be malevolent, harmful, and in extreme cases, criminal. In its worst form, the content can be characterized as *antisocial behavior* (ASB) - behaviors "which cause, or are likely to cause harassment, alarm or distress" [3] or "carried out with the proximate (immediate) intent to cause harm" [4] - and may consist of direct threats of violence, such as in the case of the tragic school shootings that have taken place in Finland [2]. The perpetrators had posted messages to popular online forums before committing acts of violence.

The word choices and the linguistic style of UGC reveals information about people's preferences, thoughts, emotions, and behaviors. Especially, emotions connect individuals to the social world and hence trigger several types of social and psychological phenomena, such as altruism, as well as negative actions, such as ASB. Moreover, thoughts, opinions, and attitudes can be explicitly or implicitly expressed through the choice of words and grammatical constructions [5]. As Tausczik and Pennebaker [6] pointed out, language is the "most common and reliable way for people to translate their internal thoughts and emotions into a form that others can understand." An analysis of language use thus provides the ability to understand communication and the psychology of human beings [6].

ASB has been extensively researched in the fields of psychology and education [7–11]; however, the computational linguistic analysis of ASB and its associated emotions as they appear in written language has not received much attention, which makes applying *natural language processing* (NLP), i.e., the automatic analysis of language, to ASB detection a significant research challenge.

To identify and classify a behavior, the behavior itself and its characteristics including the emotions (e.g., happiness, sadness, and anger) that pertain to it must be understood. Hence, to detect harmful content published on the Internet and to assess the potential for ASB that may result, the behavior's characteristics or features must be detected and analyzed, and the information must then be used to predict

ASB. By analyzing and identifying these specific features in writing, indication of ASB can be detected while potential acts of violence that may follow are still in their planning stages.

An exploratory and interdisciplinary approach that combined computer science (algorithms), information retrieval (text comparison methods), NLP (automatic syntactic and semantic analysis of texts), linguistics (linguistic expressions used for expressing emotions), and psychology (mental models of emotions and behavior) was required to achieve the objectives of this research. As a cornerstone, this doctoral dissertation presents novel research in the detection of ASB based on textual content by utilizing various NLP technologies.

1.1 MOTIVATION

ASB incidents, such as school shootings, have often included written materials, such as journals, poems, and school assignments, that display negative emotions, such as anger, depression, and distress, or that share unpleasant experiences, such as victimization, bullying, thoughts of suicide, and homicide. For instance, Luke Woodham, who killed two and injured seven people on October 1, 1997, in Pearl, Mississippi, left a last will and testament and journal entries, including this excerpt:

> "I kill because people like me are mistreated every day. I do this to show society - push us and we will push back." [12] "I am not insane. I am angry. I am not spoiled or lazy, for murder is not weak and slow witted. Murder is gutsy and daring. I killed people because people like me are mistreated every day... I am malicious because I am miserable." [13]

Another incident involved 14-year old Barry Loukaitis, who killed his teacher and two students and wounded another student. In a poem he wrote in his English class, he stated, "I am at my point of no return." He also wrote a poem hinting at murder that ended with: "I look at his body on the floor, Killing a bastard that deserves to die, Ain't nothing like it in the world, But he sure did bleed a lot." [14].

There are several similar cases, such as the Columbine school shooters of 1999, who also left journal entries, as well as similar examples in Finland such as the Jokela shooter, Pekka-Eric Auvinen [15]. Although these cases might be considered extreme, aggression and ASB do have the potential to escalate into criminal behaviors when unresolved or ignored.

The study of ASB has attracted a diverse set of researchers, including social and clinical psychologists, criminologists, and biologists [16]. Considerable efforts have also been directed towards detecting and preventing physical manifestations of ASB in communities (e.g., the Home Office in the United Kingdom[1]); however, this issue has not been researched from the perspective of using natural language processing for its early detection. This approach is needed, as manually monitoring online content for ASB is unfeasible due to the massive quantity of UGC that is published each day.

[1] http://www.homeoffice.gov.uk/crime/anti-social-behaviour/

NLP techniques have been shown to be useful in identifying similar harmful behaviors, such as cyberbullying, harassment, extremism, and terrorism in text, all with varying levels of accuracy; however, few research address the broader ASB, which is characterized by covert and overt hostility and intentional aggression toward others, as Hanrahan [17] explained.

Thus, there is an immediate need and multiple practical applications for the current research, particularly in the security field. For instance, law enforcement authorities need automated solutions for the detection of harmful content with potential threats for violent acts because it is impossible to screen all online content manually due to a lack of manpower. In addition, the research results can be beneficial to offices such as the Finnish Office of the Prosecutor General that are interested in stopping hate-oriented writing [18]. Similar interests are shared by several other countries in the world. In a broader context, the tools developed for the current study could be used to research radicalization and marginalization.

Due to the various types of applicability of the research, the results will likely play an integral role in future ASB prevention systems.

1.2 PROBLEM DEFINITION

The overall research problem of this dissertation was: *How can ASB automatically be detected in written language?* Language is a powerful tool used in communication that often shapes cognition and allows for classifying and categorizing the world using words. Words carry information and various implications, such as the words "America" and "Islam." In addition, the syntax of language also allows for identifying the subject, action, or object of a sentence [19]. Moreover, the semantic features of language allow for categorizing and grading experiences, including emotional experiences. For instance, there are more than 30 words in English for gradations of fear (fear, panic, anxiety, worry, trepidation, consternation, etc.) [20]. Polanyi and Zaenen [21] stated that clues about an author's attitude can be obtained from the choice of lexical items as well as from the organization of the text; however, written language has not yet been analyzed in the context of ASB. Consequently, NLP techniques were used for the detection and analysis of ASB in written language. NLP has been successfully applied in analyzing the content of written documents for several decades.

More specifically, the detection of ASB was defined as a *text classification* (TC) problem in which text was determined to either contain or not contain ASB. It was hypothesized that lexical and semantic features could be used to identify ASB in text. To this end, lexical and syntactic text characteristics were combined with semantic information. As such, the emphasis was on the way in which word choice can reveal important features about written ASB.

1.3 RESEARCH QUESTIONS

To achieve the research objectives, three primary research questions were formulated:

RQ1: How can emotions in text be identified and automatically analyzed effectively?

To answer research question 1, an analytical approach was utilized to study and compare the implications of different emotion theories for the identification of emotions in text. During the analysis, a dimensional model was developed to represent emotions in text. The results of this task included an annotated emotion resource. The analysis further served to illustrate the complexity and the challenges of detecting emotions in language.

RQ2: Which features are most beneficial for the detection of ASB in text?

The purpose of the second research question was to understand the nature of ASB and how it is represented in text. Here, the accurate detection of ASB was the focus, and the most reliable features that contributed most to accurate detection were investigated. Documents that could be classified as ASB were collected to investigate the language and emotion characteristics of ASB. To the author's knowledge, the collected data is the first corpus that exists that focuses on ASB. The corpus was used to develop the *Machine Learning* (ML) detection models for ASB.

RQ3: How can research on the language and emotion features of ASB further improve the understanding of ASB?

The final research question focused on the broader meaning and the impact of the research results towards a greater understanding of ASB in text. The findings will be useful in informing prevention and intervention approaches.

1.4 RESEARCH CONTRIBUTION

This work is the first to detect and analyze ASB in text. The current research has resulted in the following contributions, which can be summarized as follows:

- An in-depth analysis performed by the researcher of the literature on emotion, ASB, written language, and the way in which the three are connected.
- A freely available corpus of ASB for fellow researchers and practitioners to use in ASB-related research. The corpus was compiled by the researcher and Dr Tuomo Kakkonen.
- A freely available novel and fine-grained resource containing words that are annotated along a dimensional model of emotions. The resource was designed and developed by the researcher in collaboration with Dr Tuomo Kakkonen and Dr Calkin Suero Montero.
- ASB detection models that detect ASB in text with accuracies over 90%. These models were developed by the researcher and are its main results. The models can easily be integrated into other systems.

1.5 ORGANIZATION

The dissertation consists of two main parts. Part I presents the theoretical basis of the research from the perspective of emotion analysis, written language, and ASB as the central concepts of this study. Chapter 2 provides an extensive review of the literature of emotions and the associated theories. Chapter 3 reviews the connection

between emotions and language. Chapter 4 discusses ASB and the emotional features of ASB. Chapter 5 describes the existing techniques and methods that could be used to address the research problem.

Part II discusses the development of solutions to the research problem. Chapter 6 presents the framework used to develop the solutions. The development of the ASB detection models using emotion and word-based features is described in Chapter 7 and the results are presented. Chapter 8 discusses and interprets the study results as well as their implications. The limitations of the study and the potential application areas are also discussed in Chapter 8. Chapter 9 concludes the dissertation by providing a summary of the contributions and discusses remaining questions that provide directions for future research.

Although several research papers have been published by the author, the following conference papers and journal articles contain the most direct results of the research conducted for the current dissertation. For each paper, the researcher's contributions and that of the co-authors are outlined.

I M. Munezero, S.C. Montero, E. Sutinen and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Transactions on Affective Computing* **5** 101–111 (2014).

The researcher performed an in-depth analysis of the existing definitions and usage of subjective terms in NLP in order to clarify, differentiate between them, and identify ways they can be detected in text.

II M. Munezero, S.C. Montero, M. Mozgovoy, T. Kakkonen, V. Klyuev and E. Sutinen, "Automatic detection of antisocial behaviour in Texts," *Informatica* **38** 3–10 (2014).

The researcher compared the performance of word usage and emotional features for the detection of ASB in text. The emotional features used were obtained from an emotional ontology built by the second and fourth authors.

III M. Munezero, M. Mozgovoy, T. Kakkonen, V. Klyuev and E. Sutinen, "Antisocial behavior corpus for harmful language detection," *The Federated Conference on Computer Science and Information Systems*, Kraków, Poland, (2013), pp. 261–265.

The paper presented an ASB corpus developed by the the researcher and the third author. The corpus was a resource to build ML detection models of ASB in text.

IV M. Munezero, T. Kakkonen and C.S. Montero, "Towards automatic detection of antisocial behavior from texts," *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP*, Chian Mai, Thailand, (2011), pp. 20–27.

The paper presented an ASB detection model developed in cooperation with all the three authors.

Papers I and IV are a result of the work done in Part I of the dissertation. Paper I is a result of the literature analysis on emotions conducted in Chapter 2 and its related terms. Paper IV is a result of literature analysis of ASB. Papers II and III are a result of the computational implementations presented in Part II of the dissertation.

Part I – Theoretical Foundations

2 EMOTION ANALYSIS

Emotions are a primary idiom for defining and negotiating social relations of the self in a moral order. – Luts and White (1986)

2.1 INTRODUCTION

Emotions are a complex concept that encompass a mixture of expressions, embodiments, cognitive appraisals, neural activities, and social constructs [22], all of which have made it difficult to reach a consensus on a definition of emotions. There have been several attempts to define emotions. For instance, the psychologist Kenneth Strongman [23] defined emotions as follows:

> "Emotion permeates life, it is there as a subtext to everything we do and say, it is reflected in physiology, expression and behavior; it interweaves with cognition; it fills the spaces between people, interpersonally and culturally. Above all, emotion is centered internally, in subjective feelings. Like physical pain, emotion provides us with personal information that is integral to our well-being or in the extreme to our survival."

Philosopher Ben-Ze'ev [24] described emotions as typically occurring when we "perceive positive or negative significant changes in our personal situation." The *significant change perspective* is supported by Lazarus [25], an influential psychologist in emotion research, who describes emotions as involving a "provocation, which is an event that signifies a change in the person-environment relationship for better or worse."

In addition, psychologist Hillman [26] quoted the following from Drever's [27] Dictionary of Psychology about emotion:

> "Emotion: differently described and explained by different psychologists, but all agree that it is a complex state of the organism, involving bodily changes of a widespread character - in breathing, pulse, gland secretion, etc. - and, on the mental side, a state of excitement or perturbation, marked by strong feeling, and usually an impulse toward a definitive form of behavior. If the emotion is intense there is some disturbance of the intellectual functions, a measure of dissociation, and a tendency towards action."

Based on the definitions of Strongman [23] and Drever [27], it is apparent that emotions are multiclass phenomena involving cognition, sociocultural factors, behavior, and subjective feelings. It is also clear that precisely defining the term "emotion" is a complex task, which in turn makes it challenging to measure and quantify. Naturally, something that is difficult to define must be difficult to quantify.

Determining a precise definition of emotion is also further complicated by the commonplace language use involving words related to emotions. For example, the statement "I am afraid I can't give you the job" does not mean the person is experiencing fear.

Consequently, due to the inherent complexity of emotions, the research community still lacks a standard definition of the term. According to the psychologists Smith and Lazarus [28], the most common solution has been to base the definition on descriptive characteristics of the general reactions of emotions. In this chapter, the various descriptions of emotions are examined (Section 2.2), and in light of the findings, from reviewing the existing literature, a sufficient working description of emotion is presented.

2.2 DEFINING EMOTIONS

Emotions are highly complex phenomena that activate various neural, cognitive, and motoric processes [29]. Definitions of emotions have been proposed for decades, especially in the discipline of psychology [30]. Hence many of the existing definitions of emotions have been proposed by philosophers, theorists, and psychologists. For instance, the psychologist Scherer's [31] working definition of emotion was "episodes of coordinated changes in several components (including at least neurophysiological activation, motor expression, and subjective feeling[1] but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism." Dolan [33] believed that from a psychological perspective, emotions have three characteristics: "First, unlike most psychological states emotions are embodied and manifest in uniquely recognizable, and stereotyped, behavioral patterns of facial expression, comportment, and autonomic arousal. Second, they are less susceptible to our intentions than other psychological states insofar as they are often triggered, in the words of James, "in advance of, and often in direct opposition of our deliberate reason concerning them" [34]. Finally, and most importantly, emotions are less encapsulated than other psychological states as evident in their global effects on virtually all aspects of cognition."

Moreover, in their work on understanding interpersonal communication, West and Turner [35] defined emotions as "the critical internal structure that orients us to and engages us with what matters in our lives: our feelings about ourselves and others. Emotion encompasses both the internal feelings of one person as well as feelings that can be experienced only in a relationship (for instance, jealousy or competitiveness)." In addition, the therapeutic school of thought views emotions as "a complex synthesis of expressive motor, schematic, and conceptual information that provides organisms with information about their responses to situations that helps them orient adaptively in the environment" [36].

In an attempt to establish a description or definition of emotion, Kleinginna and Kleinginna [37] compiled a list of 92 definitions and nine skeptical statements from the emotion research literature. Based on this compilation, they suggested a formal definition of emotion:

> "A complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; b) generate cognitive processes such as emotionally relevant perceptual affect, appraisals, labeling processes; c) active widespread physiological adjustments to the arousing conditions and d) lead to behavior that is often, but not always expressive, goal-directed and adaptive".

[1] i.e., the subjective experience of an emotional episode [32].

Hence, based on these definitions of emotion, there seems to be a reasonable agreement concerning the relevant factors that determine emotional experiences, though there is some disagreement regarding the relative importance of these factors [38]; however, it is apparent that a description of emotion depends on the following components:

- Appraisal of a stimulus, situation, or event.
- Physiological reactions of the body, such as increased heartbeat, sweating, etc.
- Subjective feeling.
- Expressive behaviors, such as facial expressions, bodily expressions, and speech, including verbal and non-verbal aspects.
- Readiness to behave in a particular way (also called action tendencies by Frijda [39]).

The components of emotion are not independent of each other. That is, changes in one factor can directly lead to corresponding changes in others; however, which component precedes another component or whether all components are necessary for an emotional episode is unclear [40]. To illustrate, the following simplified scenario of an emotional episode that can be labeled as that of anger might take place as follows [41]:

- First, there is an eliciting event, or a stimulus; for instance, someone hits another person.
- Then, there is an appraisal of the eliciting event as hurtful and as being incompatible with the victim's desires for social interaction or well being.
- Thus, appropriately so, a negative feeling occurs.
- Then, there is an impulsive response to take a stand and oppose. Physiological arousal with a possibility to counter-aggress occurs.
- An angry facial expression and verbal counter abuse may then occur.
- Finally, the awareness of any or all of these possibilities constitutes the subject's emotion of anger.

Frijda [39] added a seventh sequence to the scenario known as regulation, where a person chooses ways to deal with the emotion and action taken. For instance, the person might feel justified and seek approval from others. Notably, all of these components might be triggered at relatively the same time.

Naturally, the manner in which the reactions and responses in the example are expressed culturally and socially dependent or might even be a causality of a specific culture and its society (the role of culture and society in emotions is discussed in the next section along with the theories of emotion). Regardless, for this thesis, the description of emotion in terms of the five components was adopted (Figure 2.1). This decision was also based on Frijda's [42] assertion that describing emotions as a componential system allows for examining emotions in terms of their descriptions, regardless of the names applied to these descriptions.

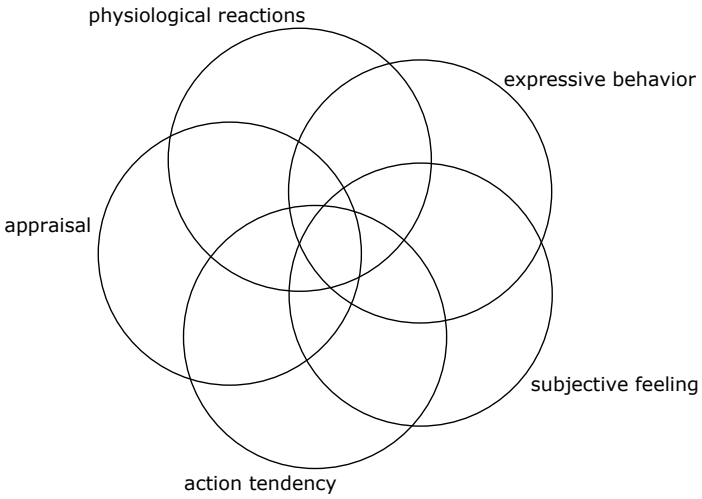

Figure 2.1: Venn diagram illustrating the five emotion components (adapted from Roesch et al. [43].

2.3 THEORIES OF EMOTION

Based on his review of the relevant literature, Strongman [23] found 150 different theories of emotion. His research led to the conclusion that there is no unified or generally accepted theory of emotions because the theories differ regarding the necessary and sufficient factors that elicit an emotion. What precedes what? Is it a thought first, or is it a physiological or neurological aspect? What specifically leads to an emotional experience?

Each of the theories has an impact on the theoretical and empirical assumptions that can be made when studying emotions. In this section, six theories of emotions that have had a significant impact on the field of emotion research are discussed. Each of the six theories presents a different perspective, which has implications regarding the approach that can be used in the study of emotions pertaining to ASB in text. Thus, after reviewing each of the six theories, the assumptions each theory contributes to the current research are summarized, and whether the theoretical and empirical assumptions are applicable in the context of analyzing emotions in written texts is discussed.

2.3.1 Darwin's evolutionary theory

The first theory of emotion can be traced to Charles Darwin's evolution theory of emotion in 1872 [44][2]. Darwin's theory focuses on the nature of emotion expression, and it states that non-verbal communication, such as body language, movements, and facial expressions, are not only used to communicate meaning but have also been genetically retained because they were useful to ancestors.

[2]The first edition was published in 1872.

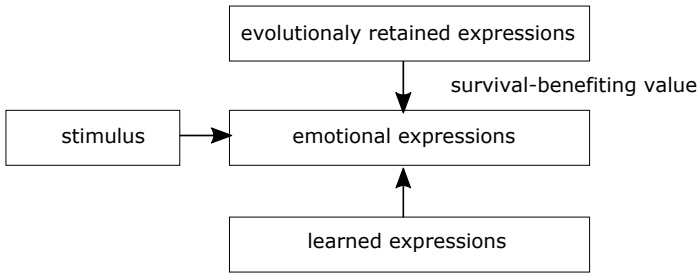

Figure 2.2: Sequence of Darwin's theory emphasizing the survival factor leading to emotion expression.

Darwin also suggested that emotional expressions are initially learned behaviors, but the main emphasis of the theory is the "survival value" of emotions, particularly their universal similarity across races and cultures (illustrated in Figure 2.2). Darwin's theory was the first to propose the universality of emotions through facial expressions.

Although Darwin's study had limitations, Ekman and Friesen [45] sought to prove the theory and developed the concept of basic emotions - a small number of emotions that are shaped by evolution to fulfill survival-benefiting functions [46]. In addition to facial expressions, other followers of Darwin's theory, such as Frijda [39] believed that the nervous system of every living species is a bundle of predispositions (expressive responses, visceral changes, or instrumental behaviors) that are triggered when contact is made with relevant features of the environment [47]. Frijda [39] referred to these predispositions as "action tendencies" or states of readiness to act in a particular way when confronted with an emotional stimulus.

2.3.2 James-Lange theory

In the 1880s, during his research on emotions, William James focused on emotions as the subsequent effect of a person perceiving his or her bodily state change. James' focus was on the nature of emotional experience. In his 1884 article, "What is an Emotion?", James argued that bodily changes must come first and that it would be impossible to have emotions without these bodily changes [48]. Similarly, Carl Lange, a Danish professor, emphasized the influence of vasomotor[3] changes to emotional experiences [49]. Because the two scientists similarly emphasized that physiological arousal precedes emotions, their two theories were combined to form one theory, which came to be known as the *James-Lange* (J-L) theory. The theory states that physiological arousal occurs first, and when this arousal is perceived or interpreted, emotion is experienced (illustrated in Figure 2.3). In other words, a stimulus triggers physiological changes in a person's body, and a person's brain interprets these physical changes into the appropriate emotion [34].

Similar to Darwin's theory, the J-L theory adopts the perspective that emotions are environmental adaptations that serve a survival purpose. According to James [48], "the nervous system of every living thing is but a bundle of predispositions to react in particular ways upon contact of particular features of the environ-

[3]Vasomotor – causing or relating to the constriction or dilation of blood vessels (vasomotor: Oxford Dictionary, 2014)

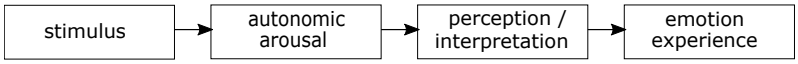

Figure 2.3: Sequence of J-L theory emphasizing the importance of physiological arousal for an emotion experience (adapted from Walsh [50]).

ment." James considered all physical responses as potential sources of the feelings that he equated to emotions [47]. Hence, when we feel sorry, it is because we cry, angry because we strike, and afraid because we tremble [48]. Damasion [51] provided evidence to support the J-L theory when he found that patients suffering from anosognosia (defined by Antoine et al. [52] as: "an impaired ability to recognize the presence or appreciate the severity of deficits in sensory, perceptual, motor, affective, or cognitive functioning") were unable to experience emotions simply because they cannot feel their bodies.

As influential as the J-L theory has been in psychology, the implications that each emotion must have its own unique constellation of physiological changes that accompany it and that an emotion cannot be expressed unless accompanied by *sympathetic nervous system* activity were heavily challenged by many researchers, especially by Walter Cannon [53].

2.3.3 Cannon-Bard theory

The Cannon-Bard theory argues against the J-L theory and states that physiological arousal, such as sweating and trembling, occurs simultaneously with emotions. The theory argues that the thalamus[4] is a necessity for experiencing emotion. According to the theory, the thalamus sends messages to the cortex for an interpretation of the emotion, which then generates the subjective feeling of emotion, and simultaneously sends them to the sympathetic nervous system for the appropriate physiological responses, thus producing arousal at the same time [54, 55] (illustrated in Figure 2.4).

Walter Cannon was first to advance this theoretical perspective, and Phillip Bard, an associate of Cannon's, later extended it. Bard also argued that emotion occurs even if there are no bodily responses to transmit feedback to the brain [55]. Bard conducted an experiment in which he severed the neural connections to the cortexes of cats. When the cats were provoked, they still exhibited the emotional behaviors commonly associated with rage and aggression, i.e., erect hair, growling, and the bared teeth. Bard called the behavior "sham rage" because according to the J-L theory, emotional behavior could not have occurred without connections to the brain.

2.3.4 Schachter-Singer theory

Like the Cannon-Bard theory, the Schachter-Singer theory [56] acknowledges that the same pattern of physiological arousal can occur for different types of emotions. Similar to the J-L theory, the Schachter-Singer theory also states that physiological arousal occurs first and provides important feedback for interpretation; however, rather than simply perceiving or interpreting the arousal, Schachter-Singer's theory

[4]Thalamus – Either of the two masses of grey matter lying between the cerebral hemispheres on either side of the third ventricle, relaying sensory information and acting as a center for pain perception (Oxford Dictionary, 2014).

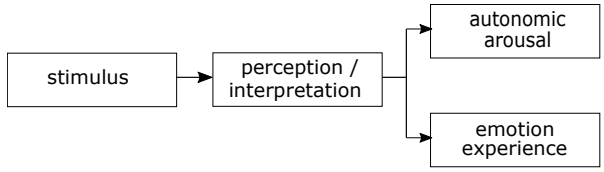

Figure 2.4: Sequence of the Cannon-Bard theory emphasizing the simultaneous occurrence of physical arousal and emotion (adapted from Walsh [50]).

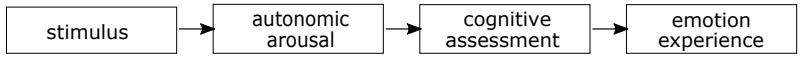

Figure 2.5: Sequence of the Schachter-Singer theory emphasizing physical arousal and cognitive labeling of the arousal to experience an emotion (adapted from Walsh [50]).

suggests that a reason for the arousal must be identified before being able to experience an emotion (see Figure 5). Thus, Schachter and Singer developed a theory of emotion that integrated cognitive activity (perception, reasoning, and memory) to appraise the stimuli and understand the reason for the arousal.

2.3.5 Cognitive-Appraisal theory

Richard Lazarus [25], one of the main proponents of cognition in emotion theories, asserted that thought precedes emotion or physiological arousal. The focus of the cognitive-appraisal theory is that thought and emotion are inseparable [47]. According to this theory, to experience an emotion and respond to it, one must think about the situation they are in.

The cognitive-appraisal theory has been widely adapted. One of the main reasons for the wide acceptance is that the theory is often believed to provide the missing link that explains the *interpretation* or *perception* in the J-L theory. Interpretation is thus explained by cognition, more particularly by *appraisal*, a term coined by Arnold [57] to represent *sense judgments*, which are "direct, immediate, nonreflective, nonintellectual, [and] automatic." Theorists of this perspective, such as Lazarus [25], Oatley and Johnson-Laird [58], and Scherer [59], pointed out that depending on the significance for the individual, the appraisal of a situation will automatically trigger an emotion and physiological response as an appropriate response to the stimuli, which can either be immediate, imagined, or remembered [60][5] (illustrated in Figure 2.6).

2.3.6 Social constructivist theory

Social constructivists view emotions and their meanings as products of culture and learned social rules. James Averill [62], a leading advocate of this theory, specified that "emotions are not just remnants of our phylogenetic past, nor can they be explained in strictly physiological terms. Rather, they are social constructions, and

[5] *The Handbook of Cognition and Emotion* provides an extensive review of the cognitive appraisal theory [61]

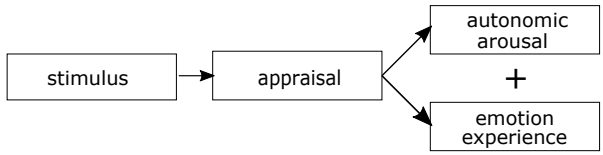

Figure 2.6: Sequence of the Cognitive-Appraisal theory emphasizing the importance of cognition (appraisal) before emotion experience and response (adapted from Walsh [50]).

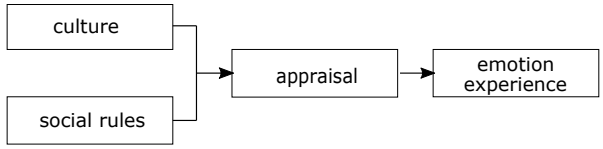

Figure 2.7: Sequence of the Social Constructivist theory emphasizing the importance of culture and social norms in emotion expression.

they can be fully understood only on a social level of analysis." The theory focuses on the systems of culturally specific rules that govern how, when, and by whom particular emotions are to be experienced and expressed [47] (see Figure 2.7).

From this perspective, emotions fulfill a social purpose by regulating interactions between individuals. Although the theory differs from Darwin's and James' theories, those who support Darwin's theory have acknowledged the role of culture in regulating emotional displays [63]. Similarly, proponents of the social perspective do not deny the psychobiological reaction components of emotions; they simply consider them secondary [31].

2.3.7 Discussion of theoretical assumptions

Having reviewed the theories of emotion, it is important to determine which theories and assumptions provide the most suitable basis for describing emotions in the context of the current doctoral dissertation topic.

The theories of emotion discussed can be grouped into five main categories: evolutionary, physiological, neurological, cognitive-appraisal, and social constructivism [64]. Although the theories seem to contradict one another, they actually focus on different perspectives of emotions, which Cornelius [64] summarized as follows:

> "Neurophysiologists are interested - almost by definition - in the neural organization of emotion, Darwinians are interested in the evolutionary organization of emotion, Jamesians [those following the James-Lange theory] are interested in the bodily organization of emotion (for want of a better term), cognitive-[appraisal] emotion theorists are interested in the psychological organization of emotion, and social constructivists are interested in the social-psychological and sociological organization of emotion."

Apart from having different foci, these perspectives have also been researched in different modalities, such as in visual signals [65, 66]; audio signals [67–69]; bio signals such as galvanic skin response to arousal [70]; electromyography - frequency of muscle of tension [71]; heart rate and respiration rate [70, 71]; brain activity measurements, i.e. amygdala [72]; and spoken or written language [73]. Research on multi modalities has also been conducted[6]

The psychological theories further provide theoretical and empirical assumptions that guide and/or limit studies on emotion. Keeping the purpose of the current dissertation research in mind - the use of the characteristics of emotions for the detection of ASB in text - the assumptions and limitations that each of the five theories impose on this purpose are discussed.

If the Darwinian approach were to be followed, which focuses on emotions as expressions and identifies the facial correlates of emotions [22], it can be assumed that there are universally recognized facial expressions present when experiencing emotions. It can also be assumed that there are body movements or reactions that are primary in every living being [39]. From this perspective, and considering a textual environment, evidence of facial expressions would be present in texts in descriptions of the face, i.e., "I have sad face" or "she is smiling at me," or in the use of symbols, e.g., emoticons. Otherwise, there would be no evidence of facial expressions in a textual environment, and thus the emotion could not be identified in accordance with the theory. Hence, the Darwinian theory is more suited to studies that have access to visual signals.

Moreover, this theory is limited by the number of basic emotions that have been identified to correspond to facial expressions. The theory can only explain a certain number of emotional behaviors and expressions [22], which would be insufficient for the purpose of this dissertation. Ekman [76], a researcher who supported Darwin's theory, acknowledged that there are emotions for which no facial signals exist or have not been identified, such as awe, guilt, and shame, and that there are different emotions that share the same facial signal, i.e., several positive emotions are expressed by a smile. Ekman [77] agreed with the possibility that any aspect of emotion might be linked to cognition, i.e., appraisal of prototypic situational events. Thus, following the Darwinian perspective does not invalidate cognitive structures; however, it does not state when or how appraisal is involved in the process of facial expressions and emotions. The advantage and usefulness of this theory lies in the assumption that there are certain innate behaviors and reactions activated by emotions [39].

From the perspective of the J-L theory, it can be assumed that an emotion experience is accompanied by unique patterns of physiological activities, i.e., changes in the autonomic nervous system. In the textual environment, expressive and descriptive[7] words may provide evidence of physiological arousal, as it is not possible to use equipment to measure physiological signals. For example, descriptive phrases such as "my palms are sweating" in reference to nervousness or "I am finding it hard to breathe" in reference to panic provide self-physiological activity descriptions; however, the theory is limited to only subjective feelings which is the component of emotion that the J-L theory focused on [32]. Moreover, it assumes that each feeling that is equated to an emotion has a unique set of autonomic nervous system activities, and thus to detect the emotions, these activities associated with

[6] Pantic and Rothkranz [74] and Sebe et al. [75] provide reviews of the multimodality research.

[7] For example, "ouch" is expressive and "I am in pain" is descriptive [78].

each emotion must be detected. The problem with this theory is that it is unable to adequately differentiate between emotions; for example, "palms sweating" could be due to nervousness or fear [53]. Still, the fact that individuals easily recognize and report their feelings and physiological arousals confirms their importance in the emotion process. Thus, the ways these feelings are reported in text will also be examined.

Moreover, based on the neurophysiological perspective, which was represented by the Cannon-Bard theory in the review, it can be assumed that there are activities in the nervous system that cause some of the emotion experiences and the accompanying physiological arousals. In particular, the emotion experience can occur without an awareness of bodily changes. That is, people can react to the emotional significance of a stimulus before fully understanding the stimulus [79]. The theory also assumes that there are neural circuits that have developed evolutionarily [31]. From a text perspective, this theory makes it more difficult to determine how the nervous system's activities make assessments of a stimuli because it would be overly complicated to use circuit models as a means to differentiate one emotion from another in text.

From a cognitive-appraisal perspective, which includes the Schachter-Singer theory, it can be assumed that each emotion experience has its own corresponding and unique pattern of appraisal, thought, and mental activity. More specifically, following from the Schachter-Singer theory, it can be assumed that an emotion has been labeled and recognized by an author of a piece of text. As cognitive-appraisals arise from personal conceptions of a situation, identifying an emotion experience becomes a complex task to perform because the uniqueness of each pattern makes it difficult to evaluate it across different people [22]. This challenge is particularly difficult in a textual environment because detailed information is not often available.

Cognitive-appraisal researchers have argued that bodily changes do not provide much information about the emotion unless something is known about how the object of the emotion is seen by the person experiencing it [22]. Although whether cognition or emotion occurs first or at the same time has been debated, it was acknowledged by Ellsworth and Scherer [60] that thinking and feeling are inextricably interrelated most of the time. Certain ways of interpreting one's environment are inherently emotional, few thoughts are entirely free of feelings, and emotions influence thinking [60]. Thus, cognitive appraisal is not always seen as a cause of an emotional experience, as noted by Scherer [80], but it can also be a consequence of emotions. In addition, it is accepted that cognitive appraisals do not occur in a social vacuum [81]. The cognitive-appraisal theory presents assumptions that are important in explaining the mental appraisal processes that either take or should take place.

Furthermore, from the social constructive perspective, it can be assumed that there are cultural and social factors in play during emotion experiences and expressions. Social processes and cultural norms play significant roles in specifying when emotions are felt and how emotions are expressed [22]. Some emotions are directed towards other people and arise from interactions with them. Although it is apparent that social and cultural norms do affect emotion expression and that there is a need to study emotions in a social context, it is a challenging task in a textual environment because the textual environment might not offer enough background information to obtain accurate results when adopting this perspective.

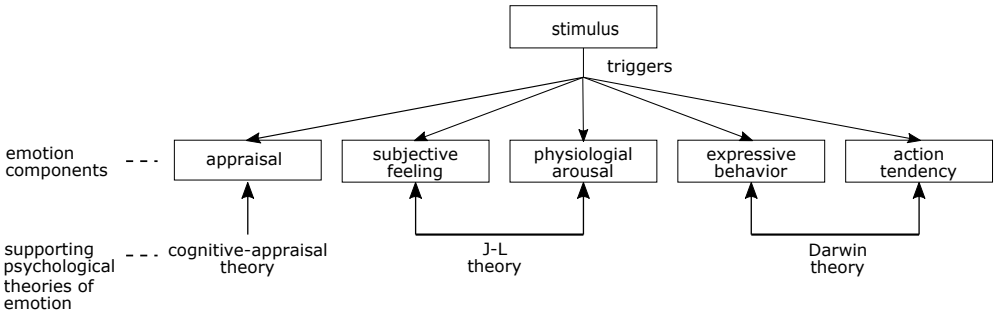

Figure 2.8: Relationships between emotion components and the theories that support them.

2.3.8 Theories convergence

All theories discussed in Section 2.3 are supported by ample empirical evidence. The aim of this chapter is not to refute them but rather to obtain theoretical support from the scientific theories for the presence of the five components of emotion. The components that precede other components cannot be accurately determined in a textual environment. Hence, the mere presence of these components can provide evidence of whether an emotion is expressed and which emotion is expressed.

Based on the review of psychological theories, the theories that provide support for the five components are: Darwin's theory (facial expressions and action tendencies), the J-L theory (subjective feelings and physiological arousal), and the cognitive-appraisal theory (appraisals). Each theory agrees that emotions are triggered by a stimulus or event (external or internal) that is deemed important to the organism. Figure 2.8 illustrates the relationships between the components and theories.

Now that the fundamental components present in an emotional experience have been identified, a way to differentiate the occurrences of one emotion from another must be determined. The differentiation of emotions is typically accomplished using descriptive frameworks that capture various aspects of emotions according to a given psychological theory. These frameworks are discussed in the next section in the context of the three supporting psychological theories identified (see Figure 2.8).

2.4 MODELING EMOTIONS

While psychological theories of emotion allow for understanding emotions, they do not provide a means to model or differentiate between emotions. In contrast, a model of emotion is a schematic description of an emotion that allows for distinguishing one emotion from another by capturing the various aspects of emotions. Unfortunately, no general model has been universally supported by the emotion research community [46]. Thus, this section reviews the available options, and the models that are most appropriate for this research are chosen, particularly for modeling the three emotion theories selected in the previous section. The principle that guided the selection was the ability to differentiate emotions based on the identified components of emotion (see Figure 2.8).

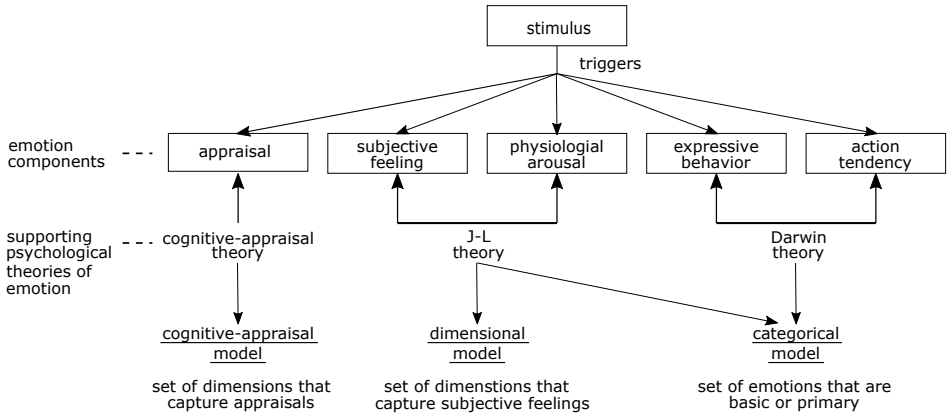

Figure 2.9: Relationships between emotion models and the theories they are based on.

There are three dominant models for differentiating between emotions: categorical, dimensional, and appraisal. These three models differ significantly with respect to the number of emotions they explain as well as the principles that are evoked [31]. In addition, they differ regarding the emotion theories that serve as the foundations of their structures. Figure 2.9 illustrates the relations between the theories described in Section 2.3 and the models of emotion that are described in this section.

2.4.1 Categorical model

The categorical model represents emotions as a small set of emotions that are considered basic, hardwired in the brain, and recognized universally [45]. The categorical model argues that each emotion has unique experiential, physiological, and behavioral correlates [82]. These correlates are captured in discrete categories, such as happiness, anger, and fear. There are no "fuzzy" emotional states described in these models. For the differentiation of human emotions, the categorical framework has been the most commonly adopted approach, particularly for differentiating between nonverbal emotion expressions [83].

Various categorical models differ regarding the number of emotional categories. Generally, the categories are described using emotion-denoting words or category labels derived from everyday language [84]; however, using emotion-bearing words as category labels may lead to a large set of emotions that would be difficult to implement. For example, Whissell [85] identified 107 adjectives, and Averill's [86] *Semantic Atlas of Emotion Concepts* contains 558 words. Hence, this approach leads to a rather large number of categories. To reduce the number of categories to a manageable level, a subset of the selected emotion-bearing words is usually used to denote the categories [46].

Some researchers, particularly, those discussed in Section 2.3 who supported the Darwin and J-L theories, base their subset requirements on the notion that some emotions are basic or primary. Thus, their subsets consist of emotions that are deemed biologically basic and stable across people and cultures. For instance, Ekman et al. [63] demonstrated the existence of six basic emotions: anger, disgust, fear, joy, sadness, and surprise; however, as the number of basic emotions might be too

limited in some cases to explain the variety of emotional states that are described by laymen [31], other researchers have based their requirements on superordinate emotions - emotions that include others [87–90] - essential everyday emotion terms [91], or category labels as agreed upon by the researchers of a study [46].

The categorical approach has been widely adapted due to its simplicity and usefulness in modeling emotions that are maximally different from each other [46]. In addition, the categories reflect that happiness, anger, and fear are basic feeling states and are easily recognizable by most people [60]; however, modeling emotions by using a set of discrete categories does fail to capture the apparent similarities between emotions. In addition, this approach does not illustrate the divergence of emotions [92]. Scholsberg [93] believed that divergence would be better revealed by modeling emotions on a continuum scale rather than by an indefinite number of categories.

Some researchers [65, 94] have also found it difficult to believe that a single label or small number of discrete classes would be adequate for differentiating between human emotions due to their complexity. Due to the limitations of the categorical model, there have been advocates for a dimensional model of emotions that is based on a continuous representation that relates emotions to each other rather than separating them into categories (e.g. [40, 65, 69, 95]. The dimensional approach to modeling emotions is discussed in the next subsection.

2.4.2 Dimensional model

The complexity of describing emotions can be simplified by representing them along dimensions. A dimensional representation aims to capture the most conceptual features of a phenomenon and provides a way for measuring similarities [46]. To capture emotions, dimensional models have been developed to represent the consciously accessible subjective feelings that do not have a specific direction, such as the feeling of pleasure or displeasure, tension or calmness, and depression or elation, all of which might be caused by external or internal situations, e.g., feeling sad when a particular song is played [96].

Izard [29] stated that the first researcher to represent emotions as dimensions or states of consciousness was Spencer [97] in 1890. After Spencer, Wundt [98] expanded on Spencer's work and identified three dimensions: pleasantness - unpleasantness, excitement - depression, and tension - relaxation.

Unfortunately, Wundt was not able to provide experimental evidence to support the three dimensions; however, in his desire to arrange facial expressions on a continuum scale, Scholsberg [93] obtained fundamental results to support the basic dimensions of emotions, which he identified as pleasantness-unpleasantness and attention-rejection. The scale was also identified as circular. Still, Scholsberg acknowledged that the two dimensions could only provide a limited account and that additional dimensions would be necessary to capture a more complete description of all possible facial expressions.

Subsequent studies have focused on determining the appropriate number of dimensions; however, as Fontaine et al. [99] pointed out, establishing the underlying set of dimensions that most economically account for the similarities and differences of emotional experiences has been a difficult process. To this day, there is considerable disagreement regarding the number and nature of the underlying dimensions that provide an optimal framework for studying emotions.

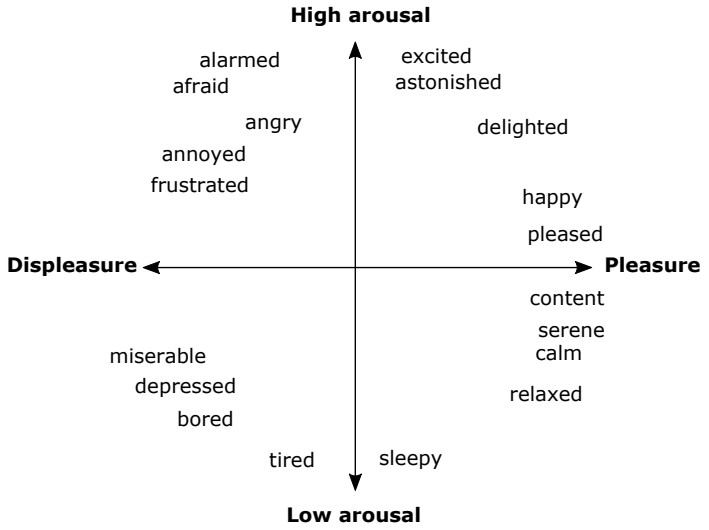

Figure 2.10: Russell's [65] Circumplex model. Russell used statistical techniques to produce a two-dimensional "map" of the mental space of emotions starting from surveyed ratings of similarities between emotions (adapted from Russell [65]).

Theorists have suggested that each emotion corresponds to a unique region in their proposed uni- or multi-dimensional spaces [60]. In addition, dimensional theorists have mainly focused on modeling the subjective feeling component of emotion (see Section 2.2). The two-dimensional model of valence and arousal has been identified as being able to efficiently differentiate between subjective feelings (Barrett & Russell, 1999). *Valence* can be viewed as an intuitive dimension. It not only captures what is generally observed as the most important dimension of feeling but also reflects the two fundamental behavioral orientations of approach and avoidance [31]. *Arousal* (also referred to as excitation or activation) refers to a sense of mobilization or energy [96]. The two dimensions valence and arousal were called *core-affect* by Russell and Barrett [96]. Core-affect refers to the "consciously accessible elemental processes of pleasure and activation." Ellsworth and Scherer [60] stated that the two dimensions provide a sufficient basis for representing similarities and differences among emotions and that they can account for a great number of emotional experiences. In addition, the two dimensions make it easier to visualize similarities and differences between emotions in terms of how close or far the emotions are in the dimensional space [31]. See Figure 2.10 and Figure 2.11 for two examples of visualizing emotions along two and three dimensions, respectively.

The dimensional model provides a sufficient definition for various emotional experiences, including a simplified description of several aspects of the emotion phenomena, namely action tendencies, expressive behavior, and cognitive appraisal [46]; however, the model has limitations in that it only provides a limited account of emotions [101] and does not include important measures of emotions, such as eliciting conditions and specific action tendencies [25], which are part of the components of emotion according to the analysis in Section 2.2.

Moreover, the dimensional model has evoked criticism because in contrast to the categorical model that groups emotions in clearly separated classes, numerous emotions lie close in their dimensional spaces, thus making them not clearly dis-

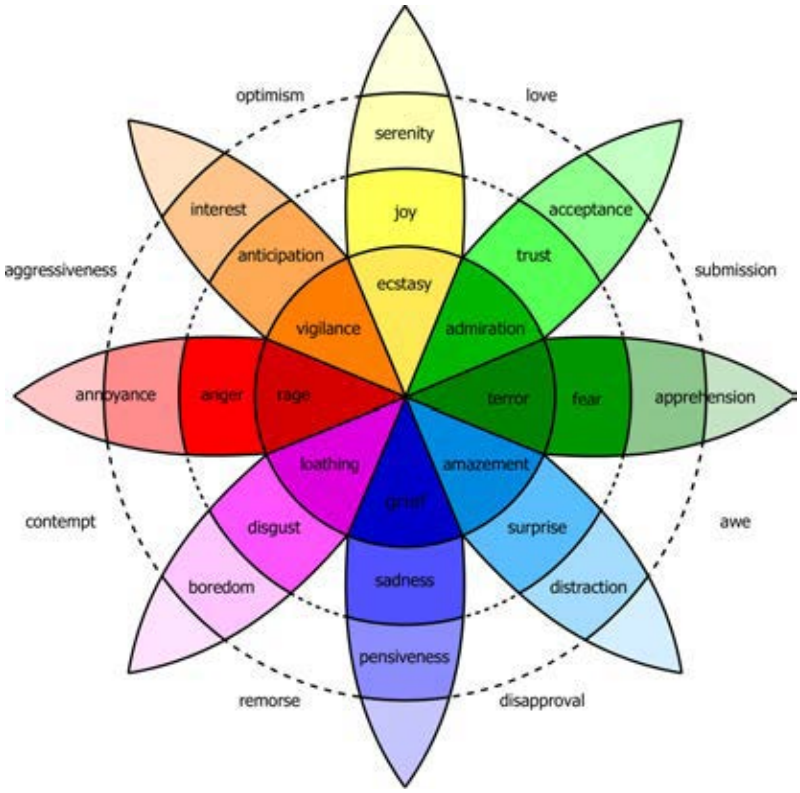

Figure 2.11: Plutchik's [100] three-dimensional model. The cone's vertical dimension represents intensity and the circle represents degrees of similarity among the emotions. The eight sectors are designed to indicate that there are eight primary emotion dimensions as defined by the theory and these are arranged as four pairs of opposites. The emotions in the blank spaces are the primary dyads-emotions that are mixtures of two of the primary emotions (source: Wikimedia Commons).

tinguishable [102]. For example, the valence and arousal values for envy are close to those of disappointment and shame. Thus, a valence-arousal dimensional model is not very discriminatory to certain emotions. This can be resolved by including causal information in the form of appraisals. As Reisenzein [103] pointed out, the emotions of disappointment, envy, and shame, which lie close together, could be "defined as subtypes of displeasure-activation emotions" and could be differentiated from each other based on their appraisal causes. Reisenzein not only called for more dimensions in the dimensional model (as that would befit the purpose of representing subjective feelings using low-dimensional representation) but also for the pleasure and arousal values to reflect the appraisal of eliciting events.

To examine the event that elicits the emotion, several researchers have added a third dimension to their core-affect based models. Examples of the third dimension include potency or dominance [104], locus of causation [105], affiliativeness [106], certainty-uncertainty [107], and intensity [103]. Fontaine et al. [99] used four di-

mensions, core-affect (valence and arousal), potency, and unpredictability. Eyben et al. [108] used five dimensions , which included Fontaine et al.'s [99] four dimensions as well as an intensity dimension.

Intensity was strongly advocated by Reisenzein [103] and Frijda et al. [109] as an important dimension for modeling the subjective feeling component of emotions. It is natural that when people describe their emotions, they do not only say they are angry, sad, or happy; they also say that they are a little, somewhat, or very angry, that they feel nothing, or that they feel little or great sadness or moderate or intense happiness [103]. Russell and Barrett [96] also acknowledged this by stating that "core affect feelings vary in intensity." Reisenzein [103] provided evidence of ways intensity can be included in the core-affect dimensions.

2.4.3 Cognitive-appraisal model

Cognitive-appraisal models are used to differentiate between emotions as well as to evaluate eliciting situations (i.e., a situation that may trigger an emotion if a person thinks that it affects one of their goals) [25]. There has been a consensus among cognitive-appraisal theorists that appraisals are fundamental to emotion experiences, not only determining whether or not emotions are experienced, but also strongly influencing the precise emotion experienced [38]. Appraisal models specify that there is a set of mental activities or patterns that are unique in both form and function that trigger an emotion [110]. The conscious and unconscious evaluations of stimuli or situations that the brain constantly performs trigger corresponding emotional expressions in accordance with the significance of the stimuli to the individual.

The appraisals of situations that trigger an emotion vary from person to person as well as across cultures, with culture most often providing the context of the appraisals that generate emotions [47]; however, appraisal models show that there are basic responses that are consistent with each emotion [25,39,57]. That is, certain patterns of appraisals may be common across cultures [60].

Each appraisal model has its own criteria and processes for evaluating a perceived stimulus and determining the emotion reactions triggered by the appraisal. These cognitive-appraisal processes may occur rapidly, automatically, and even unconsciously [111]. Scherer [80] categorized four approaches used to develop appraisal processes and structures:

1. Using a fixed set of dimensions or criteria to evaluate the significance of events [25,39,59,112,113].
2. Focusing on the nature of the causal attribution involved in emotion-antecedent appraisal [114,115].
3. Evaluating the goal-relatedness of an event by applying specific patterns or themes [58].
4. Analyzing the semantic field of emotion-denoting natural language [116].

The appraisal theories that emerged based on the four approaches listed have often differed regarding which appraisal dimensions are viewed as most important in differentiating between emotions and analyzing the elicitation situation (described in Table 2.1); however, in general, are similarities between the dimensions that have been proposed [60,117].

Table 2.1: Appraisal dimensions and their descriptions [60, 118]

Dimension	Description
Novelty	Something in the environment (physical, social, or mental) changes and an organism's attention is attracted.
Intrinsic pleasantness/ unpleasantness	Inherent pleasantness or unpleasantness of a stimulus, which causes the organism to experience pleasure or distress.
Certainty or predictability	Evaluation of whether what is occurring is understandable to the organism.
Goal significance	Evaluation of the goal relevance of the stimulus.
Agency	Evaluation of what caused a situation to occur.
Coping potential	Evaluation of the extent to which the organism is capable of coping with the stimulus in terms of its goal/plan structure.
Compatibility with social or personal standards	Evaluation of one's own actions or the actions of others and their results, with social norms and various aspects of self-concept.

Unfortunately, in the field of cognitive-appraisal, a direct comparison of models is a difficult task, as it would involve identifying shared and differing factors related to the causes of an emotion and attempting to measure them independently to compare their roles in identifying an emotion [113]. In addition, while there are similarities in the dimensions, there is no standard set of dimensions [113]. Hence, it is the task of each researcher to choose an appraisal theory that best fits the purpose of the study.

Many of the measures of the appraisal dimensions were obtained through self-reporting in which participants were asked to recall an event when they felt a certain emotion and then were asked to rate the event along several dimensions (e.g., [111]); however, analyzing text as a medium of communicating emotions poses a challenge to the cognitive-appraisal model in that it is not always possible to explicitly or implicitly determine the identified appraisal dimensions. Most appraisal dimensions cannot be automatically extracted from texts, as no or not enough information pertaining to the dimensions is present [119].

2.4.4 Discussion of the models of emotion

This section has reviewed the most widely accepted models that represent emotions. These models are essential in detecting emotions pertaining to ASB, as they allow for characterizing and investigating emotions. The three models discussed have different foci and respective explanations for various aspects of the emotion process. This is mostly because the models attempt to capture emotions based on psychological theories, which differ in their theoretical underpinnings, as explained in Section 2.3. The diversity thus creates a methodological challenge in that there is no model that is capable of capturing all the emotion components discussed in Section 2.2. Of the three models, the dimensional model comes closest to capturing the emotion com-

ponents, i.e., a combination of dimensions could be used to capture the components and make distinctions between emotions. For example, a dimension from the Darwinian model could be "primitiveness" and it would capture the extent to which an emotion is culturally or socially adopted. A dimension "approach-withdrawal" from action tendency, could capture flight, flee, or value-relevance. The dimensions "valence" and "arousal" could capture subjective feelings, and the dimensions "potency" and "intensity" could capture physiological arousal, and, "unpredictability," "uncertainty" or "unexpectedness" could capture emotions reflecting urgent reactions to novel stimulus or unfamiliar situations. Russell [105] however, pointed out that dimensions beyond valence and arousal were limited to subsets of emotion-related words.

It is beyond the scope of this research to investigate the number of dimensions that would be sufficient to completely capture the emotion components, but this type of study is encouraged. Identifying dimensions that capture the components of emotions could serve as a rationale for ways research on emotions can be conducted to obtain a unified model of emotions for the practical use of detecting and differentiating between emotions. Other approaches could then be ignored, or alternatives to this model could exist to process emotions based on different levels of descriptions or explanations.

As the focus was capturing emotions expressed in text, the dimensional model was selected for this work, specifically the two-dimensional model with valence and arousal. The dimension of intensity was a fundamental addition to this representation. The review of the models yielded results similar to those of Russell and Barrett [96] who argued that two dimensions are sufficient to capture subjective emotional experiences. The dimensional model not only captures subjective feelings and physiological arousal but also provides a simplified description of several aspects of the emotion phenomena, namely action tendencies, expressive behavior, and cognitive appraisal [46].

2.5 CONCLUSION

This chapter began by recognizing the lack of a standard, unified definition for the concept of emotion and showed that emotion is an integration of multiple factors. Some of the theories of emotions were reviewed, and a working description of emotion was presented that includes five components: appraisal, subjective feeling, physiological arousal, expressive behavior, and action tendencies. It was also concluded that these factors are not independent of each other.

Psychological theories of emotion that explain and provide support for the five components identified were then reviewed. The theories that provide sufficient support for this study on emotions in text include the: cognitive-appraisal, J-L, and Darwinian theories.

In addition, the five components have been modeled and differentiated using different theoretical models. An analytical approach was adopted to compare the implications of the different emotion theories. Based on this analysis, it was found that emotions in text can be captured by the dimensional model along the dimensions of arousal, valence, and intensity.

3 LANGUAGE AND EMOTION

Mistrust the person who finds everything good; the person who finds everything evil; and still more the person who is indifferent to everything. - Johann Kaspar Lavater

3.1 INTRODUCTION

As explained in the previous chapter, an emotion experience causes various reactions that include facial expressions, physiological arousal, neural circuitry, and cognitive processes. As indicated by Derks et al. [120], individuals need to share their emotional experiences, especially as soon as they happen, with the exception of shame[1] [120]. Wilce [122] pointed out that written communication is inherently emotional. The fact that emotion experiences can be lexically encoded, and more specifically, can be expressed by using words, was central to the aims of this dissertation [73].

In addition to the use of emotion-bearing words, an author of a text is equipped with a variety of linguistic markers that convey modulations of emotion [73]. In this chapter, ways that an emotion experience can be communicated and expressed using natural language are explored.

Several researchers have affirmed that emotion is expressible in language as well as that language itself contributes to the emotion experience [59, 123–125]. In accordance with the aim of the current dissertation, this chapter specifically focuses on the ways emotions are expressed in written language.

Reilly and Seibert [73] described the connection between emotion and language as two systems that are continuously involved in the daily interactions of members of all cultures. While differences have been observed in the ways cultures and languages express emotions, regardless of culture, any linguistic utterance can be produced to convey emotion and can also be interpreted in an emotional context [62, 73, 126]. Kövecses [127] also indicated that language does indeed shape the expression of emotions both in writing and in speech. Thus, by studying the language used by an individual, a great deal can be learned about their emotions, and in this case, information about the emotion experience of an individual as evidence of the presence of ASB can be obtained.

3.2 DESCRIPTION OF LANGUAGE

A primary function of language use is communication. Language can be expressed in different mediums, such as speech, writing, and other symbolic or gestural systems [128]. As Tausczik and Pennbaker [6] pointed out, language is the most common and reliable way for people to convey and share their thoughts and emotions in a form that others can understand. Rus [129] defined language as "a collection of symbolic expressions used through innate or learned convention by a group of communicators to construct knowledge, while interacting with their environment

[1] This is because "shame" makes one want to withdraw and hide from social interaction rather than talk about it [121].

and to communicate with each other." Based on Rus [129], Stavrianou [130], Huddleston [131], and Quirk et al.'s [132] descriptions of language, language can be broken down into several components (Table 3.1).

Table 3.1: Components of language. The components that are most relevant to this work are highlighted.

Content	Description
Alphabet	The alphabet is the set of symbols used by a language. Alphabets form words, which are the linguistic units that enable people to talk about object, actions, and quality such as 'door,' 'happy,' 'president.' 'dig,' etc. These units (words) have a meaning and structure which connects them to the world outside of language as well as to other words within language (e.g., 'car,' 'drive,' and 'driver').
Lexicon	Referred to also as a dictionary. A lexicon deals with the lexical items of the language - the lexemes, words, and so on, that make up its vocabulary.
Morphology	Morphology is concerned with the forms of words.
Syntax	The order and combinations in which words are combined to form sentences as governed by rules and conventions.
Grammar	A grammar encompasses the complex set of rules and conventions specifying the combination of words into larger units, phrases, and sentences.
Semantics	Semantics deals with the meaning relations created by words and phrases.
Pragmatics	Pragmatics is concerned with the meaning of linguistic expressions when uttered within particular types of situations.
Phonology	Deals with the sound pattern of language.

By identifying the components of language, it can be concluded that there are two groups of language users: the *Speaker*, who produces language expressions, and the *Listener*, who interprets language expressions. The meaning created by words and sentences might be different for the speaker and the listener. Regardless, a speaker and listener are able to efficiently communicate with each other through the use of consistent language processing tools that: (i) allow speakers to generate incrementally valid language expressions that represent objects of their meaning and (ii) facilitate auditors to interpret language expressions produced by the speakers in the meaning relations created by the speaker [129].

All the components listed in Table 3.1, excluding phonology, comprise the structure of texts. Most notably, as Tausczik and Pennebaker [6] explained, words can be used to understand communication and the psychology of human beings.

Thus, for this dissertation, the words and phrases used to communicate ASB and the emotions present in the words and phrases were the focus. Hence, from Table 3.1, the following three components of language were included: lexicon, syntax, and semantics. Including pragmatics would be essential for this type of study; however, this is a proposed focus for future work for several reasons. As explained by Hickey [133], pragmatics are "directly interested, not in language, but in what people do with language: its uses and users." It is concerned with knowing what language-users mean rather than what their language means. In other words, it involves distinguishing between what speakers or writers actually mean from what their utterances mean. An example is sarcasm, such as the expression: "Yeah, right. That was really great!" For this phrase, the interpretation of the meaning depends on observing the non-verbal communication of the speaker. The accurate interpretation, assumptions, purposes, or goals of the meaning rely on both social and cultural norms [134] and are varied within communities. This information is typically not available or identifiable in text.

For the purpose of the current research, which was to determine whether certain utterances can be used to detect ASB, it is beyond the scope of this study to delve into identifying the meanings attached to language. Moreover, handling, modeling, and using contexts are still poorly understood and are one of the more difficult aspects of NLP [135]. Current NLP techniques do not process pragmatics effectively; however, including pragmatics would be an ideal continuation of this work.

3.3 EMOTIONS IN WRITTEN LANGUAGE

Expressing emotional experiences in face-to-face communication is very different from the way they are conveyed in written communication. During face-to-face discussions, in addition to the words uttered, it is possible to infer emotions from non-verbal communications, such as gestures, body posture, facial expressions, and the tone and the pitch of the speaker's voice. For example, people might express anger by raising their voices, squinting their eyes, and frowning while pounding their fists in their palms.

In written communication, emotions must be inferred only from the words that are expressed and the style of writing. For example, anger might be expressed using insulting words, an angry face emoticon, capitalized words, and exclamation marks. In textual communication, tone and style patterns can be conveyed with the use of words, punctuation, and capitalization. For instance, capitalization can be used to convey shouting similar to shouting during face-to-face communication. An example would be: "GET OUT OF HERE! WHAT DO YOU WANT?" [136]. This is the reason that writing an email or text message in all capital letters is discouraged. Capitalization can also be used to convey a heightened intensity of emotions, such as happiness (e.g., "IT WAS SO GOOD!, Couldn't help but eat all of it.") [137]. Punctuation can also convey tone. For instance, a question mark raises the pitch at which readers mentally hear a sentence, while an exclamation point causes readers to mentally raise the volume of the phrase or sentence.

Therefore, the textual environment provides its own advantages and limitations in effectively expressing an emotion experience.

3.3.1 Written language in communicating emotions

There are several limitations imposed by written language when analyzing emotions. For instance, it is impossible to observe when someone is trying to refrain from crying after hearing bad news or trying to manage anger. The lack of non-verbal cues in text is the principle reason that it is difficult for readers to determine the true emotions felt by writers [138]. The lack of non-verbal cues has led to the use of emoticons and symbols in text to communicate non-verbal cues. For example, if someone writes a text message that says "You are being silly" and then adds an emoticon such as ;), this is indicative of playfulness; however, if that same person had just written a text message that says "You are being silly," whether the intention is to scold or to be playful is unclear [139]. The smiley face in the first example makes the intended meaning clearer [136]; however, emoticons can also be used to express feigned emotions, i.e., to display sarcasm. Thus, emoticons do not necessarily affirm that an individual is experiencing a specific emotion.

Derks et al. [120] showed that people are more willing to express their true emotions in text-based communication than in face-to-face communication. In their comparison of face-to-face and computer-mediated communication, Derks et al. [120] found that textual-based communication has the same potential of expressing positive emotions as face-to-face communication but that intense negative emotions are more overtly expressed in text. This is because the text environment lacks a physical presence, which affects emotion expression. Similarly, Siegel et al. [140] concluded that it is easier to express negative emotions and to show more antagonistic displays towards others in text. The reasons for this include the lack of physical feedback from the other person and the absence of other people, which makes one less aware of the social effects of expressions. For example, in face-to-face interactions, a person would normally be careful in how they express anger or sadness to other people because their expressions might evoke negative consequences [120]. The observations of Derks et al. [120] and Siegel et al. [140] supported the motivation to focus on text communication, as it serves as an adequate modality for detecting the negative emotions that pertain to ASB

Communication within textual environments also leads to reduced spontaneity. This means that individuals undergoing emotional experiences have time-lags before they share the experiences. Sharing emotions in text allows for time to think, re-read, reflect, and modify the way emotions are expressed. Thus, the environment provides more control of the expression and extent to which a person chooses to share emotions with others. It also inhibits impulses in comparison to face-to-face communication [120]. Due to this time-lag, the automatic detection of emotions in text provides an interesting opportunity for self-reflection applications. For instance, applications embedded in email, instant messaging, and Facebook could include self-reflection tools to confirm that the writer wants to send a message containing the detected emotional information.

3.3.2 Challenges in detecting emotions in written language

Identifying emotions in text has challenges, many of which are discussed in this dissertation. Unfortunately, due to time constraints, it was beyond the scope of the current research to resolve them, thus presenting future research avenues.

Ambiguity

Ambiguity in language makes it difficult to identify emotions. While words such as "miserable" and "painful" are fairly unambiguous in their emotional meanings, there are words that are capable of involving different emotions depending on the context in which they appear [141]. For instance, the word "unhappy" may involve angry or sad emotions depending on the situation in which it is expressed [142]. In addition, phrases such as "shed tears" do not always imply grief but can imply happiness as well. Thus, in some situations, it is difficult to clarify the emotion, as a writer might write a sentence with a double meaning or there might be multiple possible meanings. One solution would be to examine the sentences before and after the sentence to consider the context in which it appears. Other approaches such as word sense disambiguation can be utilized, which is a technique used to determine the most probable meaning of a polysemous word by considering the context in which the word appears [143].

Different entities in text

In some instances, writing may express emotions of other entities instead of those of the author. By merely examining words and not their referents, it is possible that emotions might be erroneously attributed to the author, for example: "(I) had such a great time; however, my (mother) was annoyed with me." In this example, the mother was annoyed rather than the author.

Context sensitivity and domain dependency

Often, a certain set of words may not convey emotional information until they are put into a certain context, for example: "I don't understand why people tell lies!" This sentence does not contain words that could be considered emotional; however, a feeling of sadness and anger are conveyed by the sentence. In other words, the meaning of the sentence as a whole is more important than the meaning of the individual words separately [144]. Interestingly, words with the same connotation may evoke different emotions depending on the context. For instance, the word "forgive" in the following sentences evokes two different emotions based only on the pronouns [141]:

s1 - Tell him to forgive me if I ever treated him badly. (guilt)
s2 - Tell him I forgive him for all my heart aches. (forgiveness)

Contextual emotion valence shifters

Valence shifters [21] include terms that change the valence or intensity values of expressed emotions. Valence shifters are important concerns in analyzing emotions. *Negation* is a common type of linguistic construction that affects valence. Negation provides various challenges in emotion analyses because it can appear in different forms: at a local level, i.e., "not good," longer level, i.e., "does not look very good," or negation of the subject level, i.e., "no one thinks that it is good." It includes words such as never, none, nobody, nowhere, nothing, and neither. In addition, there are other words such as "avoid" and "doubt" that inherently invert the valence, i.e., "I doubt he can do it" [145]. As Wiegand et al. [146] noted, a negation in a sentence with multiple clauses does not invert the valence of all the clauses, so the scope of the negation must be taken into account. Consider the following example: "I do not like the design of the new Nokia model, but it contains some intriguing new functions." The negation only affects the preference of the design.

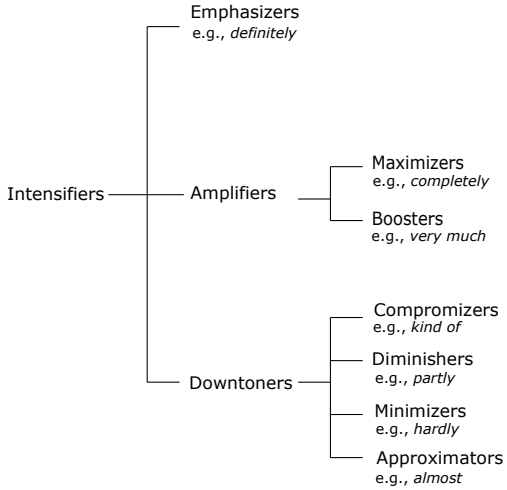

Figure 3.1: Ontology of intensifiers (adapted from Osherenko [145]).

It is also possible to use negation to intensify rather than change valence, i.e., "not only good but amazing" [147]. Thus, negation does not always invert the valence of words or phrases. More specifically, when a negation is combined with a negative word, it can make the valence positive, e.g., "He is not crazy." Furthermore, negation does not affect neutral statements, such as "I am going to the shop today" or "I am not going to the shop today." Both sentences are still neutral. For a review of approaches in analyzing negation in text, see Wiegand et al. [146].

Other valence shifters include *intensifiers* such as "deeply" and "very," which strengthen the base valence of the modified terms, i.e., "deeply in love." Others weaken the valence, such as "slightly" and "rather," i.e., "rather selfish" [148]. These valence shifters can belong to several lexical classes, such as adverbs, quantifiers such as "few" and "most," and nouns such as "lack (of)" [21]. Osherenko [145] illustrated the ontology of the classes of intensifiers, as shown in Figure 3.1.

Modality is another linguistic feature that can affect the valence of texts. An example is the sentence "Tom is a terrible person. He is rude to his friends." "Terrible" and "rude" are negative valence terms; however, in the sentence "If Tom were a terrible person, he would be rude to his friends," the "if" and "would" are the indicators that Tom is not rude to his friends nor necessarily a terrible person. Therefore, in this case, the modal operators neutralize the sentence, which would have otherwise been classified as unpleasant based on the words "terrible" and "rude" [21]. In other instances, modalities can also inverse the valence, i.e., "in theory the phone *should* have worked under water" [146].

3.4 LANGUAGE FEATURES OF EMOTIONS

Frijda et al. [41] highlighted two perspectives on viewing emotion words in language. The first perspective involves emotion words that "dictate the way things are seen," and the second involves things that "are given names and thus have words assigned to them." These emotion words are further classified as expressive or descriptive. Kövecses [127] explained that expressive words serve to express emotions, such

as "shit!" when angry, "wow!" when enthusiastic or impressed, and "yuk!" when disgusted. While descriptive words serve to describe the emotions they signify, words such as anger, angry, joy, happy, sad, anxiety, terror, apprehension, etc.

There are other forms of descriptive words that more or less serve the purpose of denoting other aspects of emotion, such as intensity, cause, and control, e.g., "boiling with anger," "to be on cloud nine," or "burning with desire." These descriptive words use language features such as metaphors[2] or metonymys[3] to express emotions, e.g., "to have cold feet" – as a way to express fear [127].

Unfortunately, focusing solely on expressive and descriptive words will not capture the many indistinct ways units of language can interact to express an emotion [127]. As Derks et al. [120] pointed out, emotions are often hidden behind combinations of words, which makes it difficult to accurately analyze the emotions present in a text. In addition, the existence of humor, sarcasm, and puns makes it even more challenging to interpret the emotional content of text. In addition to words, there are other linguistic features in text that also serve particular communicative purposes, including stylistic features such as explicit punctuation, capitalization, or formatting [149], as explained in the previous section.

Irvine [150] provided characterizations of some features that can be used to express and infer emotions in language. In a study of the way people expressed emotions on Facebook and Twitter, Parkins [137] also identified linguistic cues. In addition, Osherenko [145] compiled a list of the linguistic features of expressing emotions. Based on the findings of Irvine, Parkins, and Osherenko, a list of linguistic features used to express emotions was compiled and is listed below[4]; however, this is not a complete list, as combinations of words can be used to communicate an emotion:

- Exclamation and punctuation marks (such as !! and ??).
- Capitalization.
- Interjections ("Oh! what a lovely morning!").
- Morphological and syntactic devices for emphasis, e.g., vowel lengthening in deictic determinants ("oohhhh, coool"), reduplication, and repetition of own's or of another's utterance.
- Lexical devices.
  - graded sets (nothing, absolutely nothing, a lot, too much, a whole lot).
  - "loaded" terms (hoity-toity).
  - ideophones and hyperbolic expressions, e.g., "hehehee" and "I laughed to death."
  - lexical encoding of emotion and emotion-bearing speech acts, e.g., "I'm angry" and "I'm ashamed."
  - lexical descriptions, e.g., "you're good" and "you're a pig."

[2]Metaphor: a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness between them (Merriam-Webster dictionary, 2015).

[3]Metonymy: a figure of speech consisting of the use of the name of one thing for that of another of which it is an attribute or with which it is associated (Merriam-Webster dictionary, 2015).

[4]These linguistic cues are considered in the context of the English language.

  - address forms: kinship terms, praise-names, and nicknames.

- Phonological and prosodic devices
  - lengthening of long vowels and continuants and perhaps geminates: "No!!" and "Nothing!"
  - intonational phenomena: pitch, volume, speed of speech, and intonational contour.

- Nonverbal devices, such as facial expressions, gestures, and stance, especially degree of animation (amount and energy of movement, e.g., emoticons and other symbols).

- Discourse and interactional devices
  - sequencing, pauses, and repetitions. (e.g. ....., "this car is too, too expensive")

The language features listed and discussed highlight some of the ways in which emotional experiences can be communicated in written form.

3.5 EMOTION COMPONENTS IN TEXT

In Chapter 2, emotions were described based on five components. During an emotion experience, appraisal, physiological reactions, subjective feelings, expressive behavior, and action tendencies are all triggered. The question investigated in this section is: To what extent can all these components be communicated in text as evidence of an emotion experience?

Let us revisit the example in Chapter 2 of someone being punched with a fist, which triggered the reaction of the person feeling angry. If someone were to describe the story in an email or chat, the description of the experience might proceed as follows:

> "Hi, John. Remember my friend Luke? Can you believe that he punched me on Friday at John's party? I can't believe my friend would hit me! He was acting crazy! Never felt so insulted in my life. How could he do that??? Just made me lose my temper. I wanted to just hit him right back, too, and could feel the adrenaline rushing in. Started swearing at him and pointing a certain finger at him. You should have been there. Wanted to kick the guy's butt, but I thought it was not worth it, you know."

From the story, the following emotion components can be observed:

- Appraisal – "I can't believe my friend hit me! He was crazy."
  - This expression is assessed as negative and undesirable to one's social interaction and well-being. Appraisals are evident through the use of evaluative language.

- Physiological reaction – "Adrenaline rushing in."

  - Certain phrases can be used to express the inner physiological arousal, such as references to heart rate (my heart skipped a beat), temperature (getting cold feet), breathing (finding it hard to breath), and so on.

- A (negative) feeling occurs – "How could he do that???", "Made me lose my temper."
  - Feelings are made clear using descriptive or expressive language. They can be explicitly stated with emotion labels such as "angry" or "furious", or they could be expressed implicitly, e.g., "made me lose my temper."

- Expressive display – "swearing at him and pointing a certain finger at him."
  - Expressive behavior, such as facial expressions, can be found in text in the use of symbols, such as emoticons (e.g., :)), and with references to gestures and postures (e.g., shaking one's head in disapproval, fiddling, clearing of throat, quivering, and grin stretching from ear-toear).

- Readiness to behave in a particular way (action tendency) – "wanted to just hit him right back."
  - Behaviors are expressed using verbs (e.g., "hit").

The story serves to illustrate that it is indeed possible to have all five components of emotion present in written language. Unfortunately, in most cases, not all five components are communicated in text; however, it is also not clear whether all components occur during an emotion experience. For instance, Russell (1980) reported emotional experiences in which no appraisal occurred, e.g., someone waking up and saying "I woke up very happy today." This was also one of the main reasons for choosing the dimensional model to capture emotions, as it captures all emotion components even if at a simplified level (see discussion in Section 2.4).

It is acknowledged that the above example is well-structured, grammatically correct, and does not contain any spelling mistakes. This certainly is not always the case when dealing with real-life UGC from social media, discussion forums, blogs, emails, etc. These relatively new communication environments exhibit a variety of linguistic and stylistic conventions depending on the target audience: the use of slang, short hand, over use of word elongation and punctuation, and emoticons. As a result, emotions may also be expressed differently in these environments in comparison to the manner described in the example.

For instance, the example story could appear as follows, along with several other variations, with the use of slang words, emoticons, and short hand.

> "Wassup J, remember Luke? Can u believe that guy took a swing at me the other day??? He has lost his mind that one. I mean, where does he get the balls to do that, TO ME EVEN :)!!! Ohhhh, and then I lost it. I wanted to suck it to him after that, so I just gave him the finger. You had to be there to see it. Could have beaten his sorry ass, but I chickened out."

3.6 CONCLUSION

This chapter has clearly shown that there is no doubt that written language is a vehicle for expressing emotions. Although the textual environment has limitations (e.g., lack of visual cues), it is still an adequate medium that can be used to investigate the communication of emotions. Thus, based on the words and phrases presented in text, ASB will be investigated in written language.

In addition, the analysis of the components of human language presented in Section 3.1 showed that there is an author and a reader. This provides two perspectives from which emotions in text can be analyzed. Written words can evoke or trigger emotions in those who read the text and text can also reflect or express the emotional state of the author. This dissertation focuses on the latter.

Moreover, it is important to note that people display emotions according to what they deem to be appropriate behavior. Ekman [151] and Ekman and Friesen [45] identified five rules that govern the display of emotions. The following five actions are carried out by people when expressing emotions:

- Intensify – a person expresses more emotions than actually felt
- De-intensify – a person attempts to express less emotion than actually felt
- Simulate – a person acts as if he or she feels an emotion that he or she does not actually feel
- Inhibit – a person acts as if he or she does not feel an emotion actually felt
- Mask - a person acts as if an emotion is felt that is very different from the emotion actually felt

For example, employees might become frustrated and angry with their bosses, but since it is in their best interest to avoid emotional outbursts of anger, they might mask or inhibit the action tendency of lashing out due to the socially acceptable or socially desired way of displaying anger [152].

4 ANTISOCIAL BEHAVIOR

Emotions are what makes people special and gives us a reason for living - Rouseeau

4.1 INTRODUCTION

People respond emotionally to events that affect them in some way. These emotional responses are filtered through various cognitive activities and result in a wide-ranging set of behaviors. The choice of behaviors that are undertaken are influenced by factors such as prevailing social and cultural norms regarding what is appropriate, personal beliefs, and stances. This chapter first describes ASB and then identifies the ways in which emotion is involved in ASB.

4.2 DEFINING ANTISOCIAL BEHAVIOR

ASB has been difficult to define because it includes a wide variety of different types of behaviors [153]. These behaviors have been specified by various authorities ranging from the police to housing agencies, thus making ASB a subjective and relative concept that can vary over time, context, and location [153]. Piotrowska et al. [8] described it as a "heterogeneous concept that encompasses a variety of behaviors such as physical fighting, vandalism, stealing, status violation, and disobedience to adults." Hanrahan [17] described ASB as "disruptive acts characterized by covert and overt hostility and intentional aggression toward others," which includes a number of behaviors, such as repeated violations of social rules and standards, defiance and disobedience of authority and of the rights of others, deceitfulness, theft, reckless disregard for self and others, abusing drugs and alcohol, and acts of aggression and hostility that harm others.

Furthermore, Bandura [154] stated that ASB consists of both physical manifestations, such as violence towards others, and non-physical acts, such as rejection, exclusion, humiliation, and verbal abuse. ASB is also described as a "subcategory of aggressive behavior characterized by imbalance of power and continuous intention to inflict injury or discomfort" [155]. In addition to aggressive behaviors, ASB can also be expressed in the form of subtler, passive, or dismissal acts, such as ignoring, refusal to cooperate or help, hoarding, and withdrawal [9]. Both expressions (aggressive and dismissal) of ASB share the common intention to cause the recipient physical or emotional pain; in other words, there is an intent to harm someone [9].

A more formal definition of ASB was provided by the Crime and Disorder Act [3] of the parliament of the United Kingdom, which defined ASB as "acting in a manner that caused or was likely to cause harassment, alarm, or distress to one or more persons not of the same household as [the defendant]." This definition describes the consequences of the behaviors rather than defining the behavior itself and is therefore open to interpretation and inclusion of behaviors, such as the behaviors mentioned [10]. Thus, for specificity and measurability, in the current research,

aggression was the focus because in most of the descriptions of ASB, it is predominantly associated with ASB. Aggression, which is discussed in more detail in Section 4.4.1, was also the principle guide when determining whether a text contained ASB.

Admittedly, there is no simple answer to the question regarding what leads to ASB. There are several reasons that people exhibit ASB, and their origins are complex. The reasons that an individual would have the intent to cause harm are varied, intricate, and intertwined. The reasons further vary from individual to individual and are based on the circumstances. As O'Toole [156] pointed out, an individual comes to an environment, either at work or school, with a collective life experience, which can be both positive and negative, shaped by the environments of family, school, peers, communities, and cultures. Bundled values result from the collective experiences, such as prejudices, biases, emotions, and an individual's responses to stress and authority. Thus, all these factors come into play in an individual's behavior in a certain environment.

No single factor can be identified as the decisive factor of ASB. Furthermore, none of the factors is completely without effect, and they can all contribute to the resulting behavior [156]. There are cases in which the cause of ASB is a psychological defect, such as a conduct or personality disorder. Moffit [157], a psychologist, also indicated that there are individual differences in ASB. Moffit added that many people exhibit ASB at some point, but the behavior is temporary and situational, especially in adolescents; however, for some people, the display of ASB is stable and persistent. Moreover, Clarke [158] stated that individuals are more likely to engage in ASB when they think they cannot be identified.

In summary, ASB is a concern for society at large, with its main intention of "harm" being felt at the individual, school, community, or even national level. Considerable effort has been directed towards detecting and reducing physical manifestations of ASB in communities; however, ASB can also take place in non-physical forms, such as verbal or written forms. The written form of ASB was the focus of this research.

4.3 LINKING EMOTIONS AND BEHAVIORS

Emotions motivate social behaviors, such as seeking attention, withdrawing from people, or aggressively confronting people. They act as relevance detectors that determine the functional response in respect to the evaluation of stimuli [39]. When a situation is appraised, the appropriate emotional response is determined, and as André et al. [159] pointed out, this emotion response then influences the subsequent behavior. Instigating behavior has been considered one of the functions of emotion among theorists such as Frijda [39] and Izard and Ackerman [160]. Unfortunately, little information is available regarding the exact link between emotion and behavior. While some theorists assume a direct causal relationship [161], others view the link as indirect [162]. In this section, these approaches are analyzed.

Naturally, the choice of behavior might also be influenced by culture and social norms or personality traits. For example, some may express anger with fighting, and others may express it with scolding or silence. Hostility may be expressed by sulking and withdrawal, while grief may be overt or silent [117].

4.3.1 Direct relationship

Theorists who support a direct causality of behavior explain actions by citing the emotional state that triggered the action. For instance, someone did something "because he was angry" or "because she was happy." Based on this view, the emotion state itself is sufficient to account for the generation of the behavior [163]. The direct causal supporters usually focus on the evolution function of expressive behavior and action tendencies discussed in Chapter 2 (e.g., [164–166]).

Furthermore, Russell [167] highlighted that although emotions alone do not produce instrumental behaviors (flight, fight, etc.), they may still affect action preparation and the choice of behavior. Zajonc [168] provided an example of this: a person can choose a behavior based on feelings that are different conceptually from what he or she is thinking. For instance, when a person decides to make an investment and thinks that it might be risky but the rewards might be worth it, he or she might experience a negative feeling and decide not to invest even though there is no evidence to validate the feeling.

Emotions can affect the decision-making process and the ability to choose between the behavioral options. Zhu and Thagard [169] indicated that daily experiences and ordinary psychological practices are heavily impacted by the decision-making process; however, when people experience emotions such as anger, joy, fear, jealousy, embarrassment, shame, or depression, the decision-making process may be different from the process under calm deliberation [169]. More specifically, when experiencing intense negative emotions, a person is more likely to make poor decisions. Under these circumstances, Leith and Baumeister [170] found that people are more likely to disregard the high probability of negative outcomes and focus only on the best possible outcome, which in some cases has only a 2% likelihood of materializing. These intense emotions tend to cloud one's judgment and could increase one's willingness to partake in crime and aggression [11]. Baumeister et al. [163] found that when a behavior occurs based on current emotional experiences, the behavior is often "problematic and suboptimal, and indeed sometimes downright maladaptive." Baumeister and Lobbestael [11] added that the purpose of the emotion system is to facilitate positive social connections, and when this system fails to work, ASB can result.

4.3.2 Indirect relationship

Proponents of the notion of an indirect causal relationship focus on factors other than emotion that might influence the process of behavioral choice, such as cognitive mechanisms (attention, judgment, decision making, and memory). Cognition is recognized by theorists such as Anderson and Anderson [171] to be more directly influenced by emotions rather than emotions directly priming specific behaviors.

Baumeister et al. [163] viewed the link between emotions and behavior as an inner feedback mechanism. They separated emotions into *affective responses* and full-blown *conscious emotions*[1]. Full-blown conscious emotions tend to occur after a behavior and function as a type of inner feedback system that prompts the person to reflect on the behavior and its consequences and possibly to learn lessons that could be useful in guiding future behaviors. On the other hand, affective responses may

[1] Affective responses are automatic affective reactions (such as liking and disliking something). They are simple and rapid. Full-blown consciously experienced emotional reactions are complete with physiological arousal. They may be too slow and complex to instigate behavior in the same way [163].

directly inform "cognition and behavioral choice and thereby help guide current behavior" [163]. Baumeister et al. [163] further advocated that people choose their actions based on anticipated emotional outcomes. Manucia et al. [172] also showed that, for instance people experiencing sad emotions move to help others when they think that the helping will improve their emotions. Whereas some researchers (e.g., [173]) have argued that sadness directly leads to helping, Manucia et al. [172] found that the helping behavior was based on the pursuit of emotional gains. Similarly, angry people only "aggress when they anticipate that aggressive action could change their emotional state, presumably making them feel better" [11].

Baumeister and Lobbestael's [11] view was also shared by Baron and Byrne [174], who stated that people behave pro-socially to experience the joy the positive action will bring [158]. Thus, from this perspective, the choice of behavior is influenced by "mental simulations of action and their anticipated emotional consequences" [175]; however, the current emotion experience also has an influence in that a person would not be motivated to seek happiness unless he or she were in a state of sadness.

Tice et al. [176] stated that currently felt emotions can contribute to ASB in several ways other than by a direct causal influence. Current emotional distress may cause people to pursue actions that they think might make them feel better, and this perception can overpower normal self-regulation, which can then produce antisocial and self-defeating outcomes. A common example of this is when revenge is pursued in the hope of feeling better (revenge is sweet).

4.3.3 Summary

Admittedly, relatively little is known about the exact way in which the human brain organizes the cognitive and control mechanisms that allow for the crucial shift from reaction to action or about the role of emotion in this shift [177]. Much must be learned before a clear picture of how an organism can precede from reaction to action is obtained [169]; however, it is clear that emotions play a role in behavior, whether direct or indirect. When emotion does play a more direct role in priming behaviors, it results in less than optimal behaviors. This is because when experiencing intense negative emotions, the self-regulation of behavior may be impaired, which can further undermine control against violent impulses [178], thus making a person more susceptible to ASB. For this dissertation, the aim was to investigate and identify the association between emotions and written texts containing ASB. If a positive association existed, the identified emotions were used as "evidence" for detecting a state of readiness to engage in maladaptive behaviors such as ASB.

4.4 FACTORS LEADING TO ANTISOCIAL BEHAVIOR

This section specifically focuses on the factors that play a large role in incidents of ASB according to the literature, which are aggression and negative emotions.

4.4.1 Aggression

As stated in Section 4.2, aggression is the most prominent factor associated with ASB. Bushman and Anderson [4] defined human aggression as "any behavior directed toward another individual that is carried out with the proximate (immediate) intent to cause harm." Based on this definition, aggression excludes accidental harm because the harm is not intentional [4]. The intent to harm is indicated by Wood et

al. [179] as being generated through "thoughtful deliberation or relatively superficial processes." Green [180] also pointed out that negative emotions are necessary for the rational generation of the intent to harm.

Zillman's [181] theory of aggression indicates that high general arousal is associated with increased levels of aggression. Berkowitz [182] also recognized that high arousal energizes even the dominant behavioral tendencies, resulting in a person who is provoked under high arousal to become aggressive. When experiencing high arousal or intense emotions, people cannot effectively make thoughtful decisions, which can result in unplanned attacks [183]. Under high arousal, people are not always in control of what they should feel, think, or do. Thus, Berkowitz [183] cautioned that these negative, agitated emotions can result in aggressive impulsive reactions and can also instigate thoughts that facilitate these responses.

In addition to high arousal emotions (e.g., rage), low arousal negative emotions (e.g., depression) have been associated with ASB and aggression. According to Anderson and Huesmann [16], individuals experiencing low arousal may engage in ASB and aggressive behavior as a means to increase their arousal or to seek sensation-producing stimuli and situations that also tend to increase aggression. Thus, aggressive behavior may result during both low and hight arousal.

Aggression can occur at different levels, with the higher end of the aggression spectrum consisting of physical aggression, such as violence, murder, and aggravated assault, and the lower spectrum consisting of verbal aggression, such as insults and spreading rumors [16]. Both physical and verbal aggression can occur either directly or indirectly [184]. Direct aggression targets the instigator, while indirect aggression occurs outside the presence of the instigator and usually involves acts such as spreading rumors or telling lies. Interestingly, aggression is sometimes directed towards subjects other than the instigators. Substantial evidence has been provided that indicates that females predominantly partake in indirect aggression, while males are more likely to engage in direct aggression (even more prone to extreme forms of ASB, such as violence, delinquency, and physical aggression) [185]. Both genders equally engage in verbal aggression [16,184,186,187].

Age has also been identified as a factor that influences the type of aggression. While aggression occurs in all age groups from children to adults, it manifests differently [16]. Loeber and Hay [188] observed that at the preschool age, there is more physical aggression in boys than girls, although girls do partake in physical aggression as well. In the elementary years, the gender differences are more noticeable, with girls on average engaging more often in indirect aggression, while males engage more often in physical aggression.

4.4.2 Negative emotions

The occurrences of ASB and aggression have been found to originate from negative emotions. Berkowitz [189] [182] argued that increases in negative feelings will in turn increase the likelihood of aggression. Berkowitz [190] explained that individuals are aggressively inclined when they experience strong negative emotions, including frustration, when an individual's goals are interrupted. The same emotions that are associated with ASB and aggression should be detected in the automatic analysis of emotions pertaining to ASB.

Anger in particular is considered a natural instigator of aggressive outbursts. Depending on the intensity of anger or any of its associated emotions (e.g., rage, irritation, and exasperation), individuals may experience a burst of emotional flood-

ing and have difficulty staying calm and making rational decisions [191]. Anger, although common, is a negative emotion in terms of both the subjective experience and the social evaluation [87]. After ptototyping anger, Shaver et al. [88] concluded that the cognitive antecedents of the anger process involve something or someone interfering with an individual's execution of plans or attainment of goals, which also leads to frustration or the belief that someone is attempting to harm him or her. Anger is especially instigated when an individual believes that the harm is uncalled for, or illegitimate [192]. Anger has been associated with the desire to "move against" someone or an obstacle by fighting or harming it [193, 194]. In addition, anger has been found by researchers [195, 196] to have a particularly strong relationship with aggressive ASB compared with other negative emotions.

Frustration is another prominent instigator of aggressive behavior. As O'Hair et al. [197] explained, individuals experiencing frustration are likely to respond with both nonverbal and verbal displays. The goal setting theory [198] proposes that all human behavior is intentional in that "behavior is directed toward achieving goals and purposes as opposed to being random or reflexive," and when an individual is prevented from achieving a goal, frustration results. Not surprisingly, an unexpected failure to obtain a desired goal is more unpleasant than an expected failure, and the greater displeasure in the former case is more likely to arouse aggression.

Shame is another emotion that has been correlated with aggression. Tangney [199] found that shame encourages the occurrence of anger and hostility because it is a painful emotion that results from individuals negatively judging or evaluating themselves, and when the pain of shame occurs with loss of self-esteem and self-efficacy, it may lead to unfocused anger and hostility.

Not all unpleasant emotions contribute to aggression or ASB. Among the negative emotions, guilt has been found to instigate prosocial behavior, or empathy. This is because "guilt involves a negative evaluation of specific behavior somewhat apart from the global self. The experience of guilt is less threatening and therefore less likely to invoke defensive maneuvers akin to externalization of blame and other directed anger" [87].

4.5 CONCLUSION

While the way in which emotions and behaviors are connected has been heavily debated in the psychological and social science literature, there does not appear to be a lack of consensus that such a connection exists. In this chapter, it has been shown that emotions play a role in influencing or contribute to shaping other mental processes to activate behavior [167], often very rapidly.

When emotion has a more direct link to the resulting behavior, the behavior produced is counterproductive, such as aggression and other ASB behaviors. This is because when individuals experience intense negative emotions, decision making is impaired, often resulting in less than optimal behaviors. This more direct link provides a basis for the early detection of these intense negative emotions. It should be noted that by focusing on negative emotions, there are many aspects of behavior that were not considered. For instance, a deeper analysis of emotional processing and reaction modulation for regulating behavior [183] could be investigated; however these aspects depend on several factors, such as biology, psychology, culture, and social factors, which might be difficult to fully access in written texts.

It is also acknowledged that emotional states can be temporary, whereby a person can vent feelings as a result of factors such as alcohol or drugs, a romantic breakup, failing grades, or conflict with a parent or co-worker, and after a person has had time to reflect and absorb the effects of the situation, the motivation of aggression can begin to diminish [156]. Thus, the controlled processes of behavior are beyond the scope of this thesis. Furthermore, individuals may have predispositions towards aggression or anger caused by personality traits, substance abuse, or chemical imbalances [197]. Thus, it is acknowledged that the true motivation for harming another can never be determined with complete certainty. There are cases in which negative feelings such as frustration as well as love, desire, and even excitement can instigate aggressive acts.

In conclusion, it was found that aggression and negative emotions such as anger, frustration, and shame as well as both high and low arousal are factors that may cause an individual to engage in ASB. In the automatic analysis of ASB, some of these factors were expected to be identified, and using the dimensional model described in Chapter 2, several ASB texts were expected to be found in the unpleasant dimension with both high and low arousal.

5 COMPUTATIONAL APPROACHES FOR ANTISOCIAL BEHAVIOR DETECTION IN TEXT

5.1 INTRODUCTION

Developing a system that can analyze text is prone to challenges that are inherent to written natural language. This is in part because written language captures limited tones and contexts and can provide little indication of the author's intended meaning. For example, the statement "You are a real friend" could have either a sarcastic or serious meaning [200]. In addition, computers are not capable of "understanding" text, as they lack the common knowledge that human beings have in connecting concepts and creating meaning. This challenge has also led rise to the creation and use of commonsense databases (see [201] and [202] for a review).

Still, it is possible to rely on the structure of sentences, as they follow the rules of grammar [203], as well as syntactic rules that capture key patterns in the language structure [204]. These language constructs played a crucial role in the formulation of the ASB detection problem.

In this chapter, the NLP techniques commonly used for detection tasks are reviewed, and the formulated problem of detecting ASB as a text classification problem is described (see Section 1.2). In addition, relevant related work is reviewed, and gaps that the current research helps fill are discussed.

5.2 A DEFINITION OF TEXT CLASSIFICATION

Text classification (TC) is the analytical process of assigning textual documents to a predefined set of classes or categories. As mentioned, manually analyzing and classifying documents is a formidable task in terms of the time and effort required. Thus, fully or semi-automated methods that can perform this task in a matter of seconds are essential to save time and effort.

An automatic classification task requires a set of categories $C = c_1, c_2, \ldots, c_n$, where each c_j represents a class, and n is the total number of classes. The TC task requires a collection of documents, $D = d_1, d_2, \ldots, d_m$, where m is the total number of documents in the collection. In NLP terms, D is also called a *corpus* (plural form *corpora*). The automatic TC task is then to assign an appropriate class to a document with respect to the class set C. The output is then a set of pairs for which each document d_i is assigned to category c_j, where d_i, c_j is an element in $D \times C$ [205] (illustrated in Figure 5.1). Henceforth, the unit of analysis will be referred to as a text or document, though it can refer to any unit of text: a blog, a Twitter or Facebook post, a sentence or paragraph, etc.

TC tasks are differentiated by the number of classes a document can belong to. In a *binary* classification problem, there are only two classes to which a document can be assigned (e.g., ASB or not-ASB). *Multi-class* classification includes more than two categories (e.g., ASB, not-ASB, or maybe), and each document can only be assigned to one category. There are cases in which a document may be associated with more

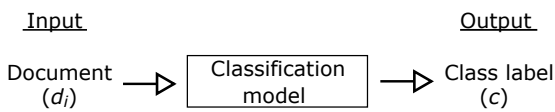

Figure 5.1: TC as the task of assigning an input document into its class label c.

than one category, (e.g., a suspense thriller book can be assigned to categories such as suspense, thriller, and crime). In this case, the TC problem is referred to as a *multi-label* classification problem. Based on the description of TC problems, the research problem (i.e., identifying whether or not a document is ASB) was categorized as a binary TC problem.

There are two common approaches to addressing the TC problems: the rule-based and the ML (statistical-based) approaches. Both approaches consider the document's characteristics or features, such as title, length, word use, or the presence or absence of certain features, to assign a document to the appropriate class. Depending on the effort invested in tuning these approaches, both rule-based and ML approaches have the capability of performing at near-human levels [206]. These two approaches are further discussed in the next subsection.

Rule-based text classification approach

The rule-based approach (also known as the knowledge engineering approach) defines a set of manually constructed conditional rules that are applied to the text to assign it to a certain category. A rule consists of an antecedent, which is a pattern in the document, and an associated consequent, which is the class. As Sebastiani [205] explained, the antecedent is a disjunctive normal form formula, and a document is assigned to a consequent *iff* it satisfies the formula. The antecedent and the consequent are then connected by an "if-then" relation [207]:

$$antecedent \rightarrow consequent$$

The conditional rules can combine lexicons, word shapes (capitalization, hyphenation, etc.), and grammatical features present in the antecedent to associate a document with a consequent. A simple rule, for example, might stipulate that if a word in the antecedent matches that in a lexicon, then the matched word, sentence, or document is assigned to a certain class [206]. Rules can also be made to manage special conditions, such as the presence of negation, intensifier, and diminisher words [208], which were discussed in Section 3.3.2. The set of rules is developed using a combination of automated methods and hand-tuning by experts [206]. In particular, to achieve high performance, the rules are typically tuned manually depending on the corpus. This can be done by either changing the weights of certain rules or by adjusting the conditions within the rules.

Unfortunately, the tuning process can require significant time from both linguistic and corpus domain experts, especially when the rules must be applied to a new domain or updates to the set of classes must be made [205, 206]. In addition, obtaining a good performance using the rules requires significant effort and linguistic capabilities to take all special cases that commonly occur into account when analyzing the text. One advantage is that, rule-based approaches do not require the antecedents to have been pre-associated to a consequent to develop the rules [206], which is the case for the ML approach discussed next.

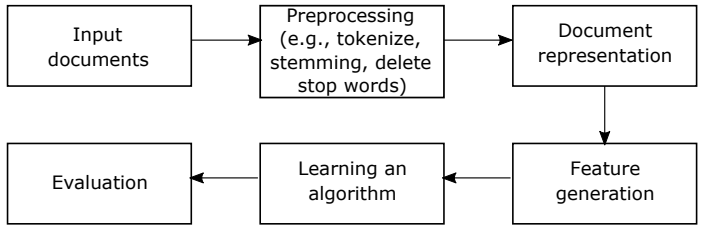

Figure 5.2: Text classifier creation process (adapted from Ikonomakis et al. [209]).

Machine learning text classification approach

Since the 1990s, the ML approach has been preferred over the rule-based approach in the research community. The ML approach is a statistical-based approach that learns classification algorithms called classifiers from examples of documents that have been pre-labeled, or associated with a category. These learned classifiers are then used to automatically assign labels to new instances of documents. This type of learning is called *supervised*, as the learning process is "supervised" by the knowledge of the categories and of the labeled document examples that belong to them [205].

For the learning process, a set of labeled documents are used to train a classifier. Using patterns and relationships in the labeled document, the classifier is then able to predict the category label for a new document. For the classifier to have a good performance, a large amount of training data (pre-labeled documents) is required. This training data can be a challenge to collect at times, and labeling each of the documents can also be a laborious task. In addition, the labeling process might require some domain experts who might not be readily available or might not offer affordable services. Despite these challenges, ML classifiers are more flexible than the rule-based approach and can be applied more quickly to new collections of text [206]. In addition, compared to the rule-based approach, ML requires less expert man power because intervention from either knowledge engineers or domain experts is not needed for the construction of the classifier itself [205].

Hence, due to the flexibility and robustness of the ML approach, the supervised ML approach was adopted to address the research problem.

5.3 SUPERVISED MACHINE LEARNING APPROACH TO TEXT CLASSIFICATION

As mentioned, the supervised ML approach to TC relies on a collection of documents that have been pre-labeled. That is, each document in the corpus is assigned a class. Using these documents, the aim of the approach is to train a classifier so that when it is applied to a new set of documents, the classifier can predict the label that a document belongs to. The process of the supervised ML approach for the creation of a classifier can be illustrated as follows (Figure 5.2):

All process components in Figure 5.2 are discussed in more detail in the next subsections.

5.3.1 Preprocessing and document representation

One of the challenges with analyzing text is that it is unstructured and several classification algorithms are designed to work with structured data. Thus, to perform TC, the textual documents must be organized in a form that ML classification algorithms can process. This is a fundamental task, and the most widely used representation for structuring the text is the *Vector Space Model* (VSM) [210]. Using the VSM, a document is described as a vector whose dimension is the number of text features, which represent the content or sometimes the structure of a document. Each feature typically encodes the presence of specific words, word n-grams, or syntactically or semantically tagged phrases or words [206]. This vector can formally be represented as:

Let $f_i, \ldots, f_m$ be a predefined set of m features that 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 $d := (n_1(d), n_2(d), \ldots, n_m(d)$ [211].

Each feature is assigned a weight, which shows the degree of information represented by that feature's occurrence in a document. The simplest way to represent a document is for each word in the document to be represented as a feature. That is, if a word appears in a given document, its corresponding attribute value $n_i(d)$ is set to 1; otherwise, it is set to 0. This approach for representing features is called the bag-of-words (BOW) approach. This representation is especially useful for TC when no prior knowledge about the specific features is available [212]; however, the BOW representation does not take any semantic or contextual information into consideration. In addition, the BOW document representation presents the sequence of words independently of how they appear in the document collection, i.e., the order in which the words appear is not recorded.

A common weighting approach for assigning attribute values to each feature in the VSM is to associate the features with their frequency in the document or corpus. This is accomplished by using weighting functions, such as the *Term Frequency Inverse Document Frequency* (TF-IDF):

$$x_i = TF_{i,d}.IDF_i$$

where $TF_{i,d}$ is the term frequency and represents the number of times term i occurs in document d. $IDF_i = log(n/n_i)$ is the inverse document frequency, where N represents the total number of documents in a corpus, and n_i is the number of documents that contain term i [213]. The TF-IDF weights take into account that (i) the most frequent terms that occur in a document are more representative of the content of the document, and (ii) if a term occurs in many of the documents in the corpus, the less discriminating it is [205].

Obtaining the features for the VSM representation commonly follows specific feature extraction techniques, such as tokenization[1], chunking[2], parsing[3], etc. Each TC task is different, and there are no "best" features for each task. It is the researcher's task to identify which features can best represent a specific domain. Regardless of the choice of features, it is always good practice in TC to pre-process the documents, as a significant number of features can be irrelevant [216]. Preprocessing the documents to reduce the number of features and remove irrelevant

[1]Tokenization is the task of dividing a text into its constituent units called tokens [214].

[2]Chunking is the task of dividing text into syntactically related non-overlapping groups of words [215].

[3]Parsing is the task of deriving the syntactic structure of a sentence [214].

features is a beneficial step in improving the performance of TC classifiers. Pre-processing has an additional benefit for classifiers, as it tends to reduce overfitting, which is when a classifier is so specific to the patterns in the training data that it does not perform well with new data [216].

Pre-processing also includes making decisions regarding whether to keep or discard features such as punctuation, capitalization, very common words (stop words, e.g., "and," "the," and "a"), and rare words (words appearing only once or twice in the corpus, which can be considered as not discriminating) and resolving spelling mistakes [217]; however, the techniques that should be applied must be considered for each TC research task.

5.3.2 Feature generation

In the previous section, the ways the features of a document can be structured in a vector format with numerical values, which can then be processed by classification algorithms, were described. In this section, selecting and extracting the features are described.

The feature selection and feature extraction processes are preliminary steps in ML, as the resulting feature set captures the relevant information about each input document that a classifier uses to predict a class for that particular document. It is important to select the type of features that are informative and discriminative enough to result in good classifier performance, as no standard or best feature set exists. Thus, it is the task of each researcher to investigate and select suitable features and to discard those that are irrelevant; however, researchers must be careful because discarding too many features might result in the algorithm lacking the information that it needs to meaningfully learn from the documents.

Though no standard feature set exists, there are commonly used features in TC, which include lexical, syntactic, structural, and content-specific features [218, 219]. Table 5.1 lists and explains the features.

Once the features have been identified, they are extracted from a corpus. Word or character level features, such as word counts, may require simple NLP tools, but for higher-level features, such as emotions or other semantic information, additional resources might be required, such as lexicons or ontologies.

Identifying emotions as features

The automatic analysis of emotions in text has generated considerable interest. Understanding what people are saying and feeling allows for improved decision making in various fields, such as recommendation and customer feedback systems, the analysis of product reviews, understanding online communication, health and counseling, education, and security and early warning systems. In addition, emotions present in text can be used to compare written materials, including comparing emotional profiles of genres of literature, such as thriller and horror novels [226], or comparing emotional profiles of males and females [227].

The most common method used to identify emotions in text begin with identifying patterns, keywords, or phrases that have been associated (i.e., annotated) with the emotions that they express. The lists of emotion-bearing words and phrases used in this process are created either manually [228, 229] or semi-automatically using seed words [230–232]. These lists (or *lexicons*) play an important role in capturing the emotional information present in text. The manner in which emotional

Table 5.1: Description of features.

Type of feature	Description
Lexical features	Lexical features are those that can be captured directly from the text [220]. They include features such as n-grams, e.g., unigrams, bigrams, and collocations. Unigrams are single tokens or single stemmed tokens (e.g., "internal," "attribution," and "turmoil"). Bigrams correspond to two elements in a sequence and thus unlike the unigram feature set, bigrams retain some word order. In addition, bigrams are able to capture certain lexical distinctions such as the difference in meaning of the word "internal" in the following phrases, "internal attribution" and "internal turmoil" [221]. Trigrams on the other hand correspond to three elements in sequence (e.g., "human internal turmoil"). Collocations represent a set of tokens that have a dependency relationship that occurs significantly more frequently than by chance (e.g., emotional baggage) [222]. Bigrams, trigrams, or collocations are often included in feature sets as a way to keep the dependencies and relative positions of the tokens [223]. The lexical features capture lexical variations in a document at both character- and word-level (e.g., average word length and total number of characters) [218].
Syntactic features	Syntactic features indicate patterns used to form sentences and include features such as; Parts-of-Speech (noun, verbs and adjectives), parse structures, frequency of punctuation, and function words. They further include features such as html markup and authors' habits of organizing a document (e.g., paragraph length and use of signature) [219,224].
Content-specific features	Content-specific features are those that represent specific topics and are comprised of specific keywords and phrases. For example, content-specific features on a discussion on violence may include the words "kill," "death," and "gun." These features can be manually or automatically selected from the documents [219]. In addition, it is possible to supplement content-specific features with semantic information such as emotional information.
Semantic features	Semantic features reflect the semantic meaning of words [225].

information is associated with emotion-bearing words and phrases is dependent on the choice of the psychological emotional model (categorical, dimensional, and appraisal), which were discussed in detail in Chapter 2.

Manually annotating words, phrases, and sentences with emotional information is a difficult process. It requires more than one expert to manually assign emotional annotations based on an agreed *annotation scheme*. As emotions are highly subjective,

more than one annotator is required to achieve reliable results. Hence, in Munezero et al. [233], the ways *crowdsourcing* could be used to collect emotion annotations for phrases and sentences were explored.

Fortunately, publicly available lexicons exist that can be readily used for the identification of emotions in text. Table 5.2 lists and describes some of the most commonly used lexical resources.

Table 5.2: Lexicons annotated with emotional information.

Name of resource	**No. of words**	**Model of emotion**	**Description**
General Enquirer [228]	8,641	Categorical	Includes categories such as anger, fury, distress, happy, positive, and negative.
SentiWordNet [234]	147,278	Categorical	Consists of positive, negative, and neutral categories.
Affective Norms for English Words (ANEW) [235]	1,034	Dimensional	Three dimensions of pleasure, arousal, and dominance.
WordNetAffect [236]	> 4,787	Categorical	Includes categories such as emotion, cognitive state, behavior, attitude, feeling, etc.
NRC emotion lexicon [237]	14,182	Categorical	Includes Plutchik's [100] eight emotion categories, positive, and negative categories
Linguistic Inquiry and Word Count (LIWC) [6]	4,500	Categorical	Includes several social, psychological categories, e.g. sadness, negative emotions, and positive emotions.
Whissell's Dictionary of affect in Language [238]	8,742	Dimensional	Three dimensions of activation, evaluation, and imagery

Table 5.2 shows that the representation of emotions is in accordance with two of the emotional models discussed in Chapter 2 (i.e., categorical and dimensional). In particular, Table 5.2 clearly shows that there are more resources with categorical representations than dimensional, indicating a limitation of resources in literature, especially as Whissell's affect dictionary is not publicly available. In addition, the dimensional representations do not include annotations for the intensity dimension of an emotion, and according to Frijda et al. [109], "not paying attention to emotion intensity may lead to misinterpretations of empirical findings."

The above mentioned lexical resources have been extensively utilized in identifying emotions present in text. In particular, the NRC emotion lexicon was utilized in two papers that this researcher co-authored [239, 240]. Unfortunately, a comprehensive comparison of the resources to identify which is the best or most useful

resource is difficult to perform. This is because many of these resources use different lists of words and are developed based on different models of emotion. In his analysis of lexicons, Potts [241] observed that some lexicons disagree with each other in that they assign opposing valence values to a specific word. Hence, it is the task of the researcher to evaluate, select, or develop a lexicon that is most suitable for the specific task.

Moreover, there are differences in the extent of the semantic information provided. For example, although it contains a large number of annotated words, SentiWordNet only provides valence annotations for the words included. Moreover, in comparison with the list of lexicons available for detecting valence in text (the valence detection process is broadly defined as *Sentiment Analysis* (SA) [242]), few are available for emotions [237,243].

In addition, although the categorical resources, such as WordNet-Affect [236], General Inquirer (GI) [244], and NRC word-emotion association lexicon [237], have proven to be useful in some tasks, they are limited in their capability of expressing the plethora of emotions in texts. As discussed in Section 2.4.2, it is unlikely that all the emotion-bearing words fit neatly into discrete sets of basic emotions. As Posner et al. [245] asserted, even "individuals do not experience, or recognize, emotions as isolated, discrete entities" but "rather recognize emotions as ambiguous and overlapping experiences." Moreover, there is a lack of consensus regarding what constitutes a basic emotion, which has led to several categorical models with varying sizes and labels for the set of basic emotions.

From Table 5.2, ANEW and Whissell's are the only resources that are based on a multidimensional model of emotions; however, the ANEW resource, for instance contains a relatively small number of words (1,034), which is arguably insufficient to be used in a practical emotion detection system [246,247]. Based on the performance of existing word list-based tools, such as SentiStrength [248], for acceptable results in emotion detection, a word list that contains at least 3000 words is needed. Admittedly, a highly accurate real-world working system might require a higher number. Whissell's resource has a larger number of words, but unfortunately, it does not account for the intensity dimension.

In Chapter 2, a suitable dimensional representation for identifying emotions in text was identified. After comparing the chosen representation model (that is, three dimensions: unpleasantness-pleasantness, low-high arousal, and intensity) to the available resources, none of the resources discussed meet the requirements. Intensity information, which was identified in Chapter 2 as an important aspect to the core affect, is not included in any of the available lexicons. Intensity was found to more accurately model people's perceptions of how they feel, such as little, somewhat, or very angry [103], rather than feeling in control or being controlled, as captured by the ANEW resource. Moreover, if intensity information is missing, a faithful representation of the text's emotional experience is not achieved.

To address the shortage of multidimensional resources as well as the lack of intensity annotations, a comprehensive and fine-grained resource was developed to efficiently identify emotions in text. The development of the resource, which was based on the core-affect dimensional model discussed in Chapter 2 with the addition of intensity, is described in Chapter 6.

5.3.3 Classification

To perform TC, a classification algorithm must be chosen. The algorithms learn from patterns in a collection of documents and use the patterns to assign a document to one or more categories. Several classification algorithms are available to researchers, and the choice of the algorithm also has an impact on the classification results. It was not the purpose of this research to provide an exhaustive evaluation of the existing algorithms, as they have been extensively covered elsewhere (see [249–251]). IEEE International Conference on Data Mining (ICDM)[4] identified 10 algorithms that are among the most effective algorithms for data mining. The algorithms used for classification include decision trees, SVM, k-nearest neighbor, naive Bayes, and CART (see [252]) for a review of the algorithms).

5.3.4 Evaluation

Once a classifier has been constructed, the next step in the supervised ML process is to evaluate how well it performs in classifying new documents. The evaluation of the performance of a classification model is based on the number of test instances correctly and incorrectly predicted by the classifier [253]. The performance is commonly measured along four parameters: accuracy, precision, recall, and F-measure. *Accuracy* measures the "correctness" of the classifier, or how often a document is classified correctly. *Precision* measures the "exactness" of the classifier and indicates the proportion of classified documents that have been labeled correctly. *Recall*, on the other hand, measures the "completeness" of the classifier and indicates the proportion of "correct" documents that have been classified correctly. *F-measure* is the weighted harmonic mean of precision and recall. To achieve a high F-measure, both precision and recall must be high, which in performance terms is an ideal condition for classifiers.

The four evaluation measures are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2TP}{2TP + FN + FP}$$

where TP, FP, FN, and TN are explained in Table 5.3:

The classification of new document instances can be carried out in two ways: (i) the test and train approach and (ii) the *k*-fold cross validation approach [205]. In the train and test approach, the whole corpus is split into two sets, a training set and a test set, with the training set usually being larger in size. A classifier is first developed on the training set and is then applied to the test set to measure the effectiveness of the classifier. It is important that the data in the test set do not participate in any way in the training of the classifier. In addition, the test data in this approach must also be represented with the same feature set representation as the training data.

[4] http://www.cs.uvm.edu/ icdm/algorithms/CandidateList.shtml

Table 5.3: Classification evaluation measures (Ikonomakis et al. [209]).

True Positive (TP)	Documents that have been classified correctly as relating to a class.
False Positive (FP)	Documents that have been related to a class incorrectly.
False Negative (FN)	Documents that have not been marked as related to a class but should be.
True Negative (TN)	Documents that should not be marked as being in a particular class and are not.

In the second evaluation approach, the k-fold cross evaluation, the whole corpus is first split into k equal and disjoint subsets. During each fold, one of the subsets is "held-out" and forms part of the test set, while the rest of the $k-1$ subsets are combined to form a training set. This procedure is repeated k times, such that each subset is used for testing exactly once [253]. Thus, with each fold, a classifier is trained and tested by applying the train and test approach. The final effectiveness is measured by averaging the effectiveness of the k classifiers [205]. This method ensures that the maximum amount of data for the training is used and also avoids testing the same data that the model is trained with. The k-fold approach is useful in cases in which a corpus is relatively small.

With the supervised ML approach, care must be taken not to fit the algorithm to the training data too perfectly, as it might memorize the noise and all peculiarities of the training data too well, a phenomenon called *overfitting* [254]. When this occurs, the training algorithm does not generalize well on new data, and it will most likely result in poor predictive accuracy for the new data. Overfitting is one of the key issues in ML. Since the interest of this research was capturing the generic underlying relationships in data and not overfitting them to once-off noise and occasional errors, a method used for preventing overfitting was adopted: the cross validation approach [255].

5.4 RELATED WORK

In the last few decades, there has been an increase in automated approaches with the goal of detecting and analyzing online content to achieve a safe Internet devoid of hate speech, racism and xenophobia [256]. These works include projects such as the Dark Web Project by the National Science Foundation [257] and the Princip Project [258]. Notable work in this area also includes the project entitled "Intelligent Information System Supporting Observation, Searching and Detection for Security of Citizens in Urban Environment" (INDECT) [259]. The project aims to detect terrorist threats and recognize criminal behavior and violence from multi-media content.

To the best of the author's knowledge, there are no previous detection ML models that detect ASB in text. Thus, an overview of the research conducted in the context of the automatic detection of other harmful behaviors such as cyberbullying, harassment, flaming, and terrorist behaviors, which could be considered subcategories of ASB, is provided. The discussion serves to highlight the techniques and approaches that have been used as well as to identify research gaps.

Table 5.4 summarizes the related work. The studies were analyzed along a set of facets to structure the summary:

- Technique: which technique was used to detect and analyze behavior content?
- Features: which features did these studies use?
- Evaluation results: what are the best evaluation results reported in these studies (if stated)?
- Modalities: which modalities of text did they consider (e.g., blogs, forums, YouTube, etc.)?
- Genre: which genre was the focus?

Table 5.4: Related studies (sorted according to publication date).

Ref.	**Genre**	**Modality**	**Technique**	**Features**	**Evaluation results**
Spertus [260]	Flames	Forums	Multi-label TC	Syntactic and semantic	64% true positive rate for the 'flame' labeled messages with C4.5 decision tree classifier
Greevy and Smeaton [261]	Racism	Web pages	Binary TC	BOW and bi-grams	87.33% and 84.77% accuracy with BOW and bi-grams features respectively, with SVM classifier
Last et al. [262]	Terrorism	Web documents	binary and multi-lingual TC	Graph-based features	98.5% accuracy with C4.5 decision tree classifier
Abbasi and Chen [263]	Extremism	Extremist forums	Affect analysis	Affect lexicon	Qualitative evaluation results

Abbasi et al. [218]	Hate and violence	Hate / extremist forums	Binary TC	Syntactic and stylistic attributes	92.84% and 93.84% accuracy on US and Middle Eastern Forums respectively, with both syntactic and stylistic features and with SVM classifier
Chen [264]	Violence	Jihadist web forums	TC, affect and sentiment analysis	Character, word and root n-grams, and collocations	92% accuracy for affect intensities and 88% accuracy for sentiment polarities with SVR ensemble
Yin et al. [265]	Harassment	Online chat and discussion posts	Binary TC	Content (TF-IDF), sentiment and contextual	44.2% F-measure with selected features of all feature types and libSVM classifier in chat posts
Fu et al. [224], Huang et al. [266]	Extremism	YouTube video comments, descriptions, titles, categories, tags, etc.	Binary TC	Lexical, syntactic, content-specific (e.g., video tags and categories)	Over 80% accuracy with selected features of all feature types and SVM classifier
Razavi et al. [267]	Flames	Log files and newsgroup messages	Multi-level, binary TC	Insulting and abusing language dictionary	96.78% correctly classified instances and 100% precision on the flame class.

Dinakar et al. [268]	Cyber-bullying	YouTube video comments	Multi-class and binary TC	General (e.g., TF-IDF and Ortony lexicon) and topic specific features.	66.7% accuracy with SVM multi-class classifier, and 80.20% accuracy with rule-based JRip classifier on the sexuality class
Bogdanova et al. [269]	Cyber-pedophilia	Chats	Binary TC	High-level (positive and negative words) and low level (character bi-grams and tri-grams, word unigrams, bigrams, and trigrams) features	94% and 92% accuracy with high level features and naive Bayes classifier on two datasets respectively
Del Bosque and Garza [270]	Aggression	Tweets	Linear regression	Swear words, ANEW, 2nd-person singular, SentiWordNet, and number of words	2-attribute linear regression with a 4.8 Mean Squared Error

Table 5.4 summarizes the state-of-the-art approaches of automated solutions used to detect harmful behavior in text. The studies indicated an increasing need to develop automated tools that help understand and detect negative behaviors. The studies were summarized along five facets, and the following observations were made:

- Technique: Table 5.4 shows that most studies approached the behavior detection problem as a binary TC task. In addition, many of the studies applied the ML approach and achieved high accuracy rates, with five studies obtaining over 90% accuracy. Based on the accuracy rates, it was found that these high rates were achieved with support vector machines (SVM).
- Features: Table 5.4 illustrates that which feature representations were most effective for harmful behavior detection is unclear. It seems that the results were dependent on the genre and modality used. The studies that included emotions (e.g., [263,264,269]) used either manually generated lexicons or semantic orientation and utilized the categorical model for representing emotions.
- Modalities: Many of the studies used generic, publicly available collections, such as YouTube, MySpace, and Twitter. A disadvantage of these types of collections is that much of the data retrieved had few instances of the case

being researched (e.g., [271,272]). It would require considerable effort to filter and obtain a representative set on which one can do an in-depth study of linguistic and emotion features pertaining to ASB. Table 5.4 also shows that no corpus of harmful behavior has yet become publicly available.

- Genre: Table 5.4 shows that related works have covered genres such as flames, racism, extremism, terrorism, and cyberbullying, all negative behaviors that can cause harm. Furthermore, many of the studies primarily focused generally on one genre; however, as the table shows, no research has yet addressed ASB, which is characterized by covert and overt hostility and intentional aggression toward others, as explained in Chapter 4.

From Table 5.4, it can be observed that few studies utilized emotional information, in Abbasi and Chen [263] for instance, the affect analysis was restricted to violence and hate-related terms. In addition, only one study analyzed emotions at the dimensional level [270].

The research objective to develop automatic solutions for detecting harmful behavior in text is similar to the studies outlined in Table 5.4; however, as discussed, limitations and challenges exist, which were addressed in this dissertation research.

5.5 CONCLUSION

In this chapter, the computational approaches used to address the problem formulation of ASB detection as a TC problem were described. Based on the approaches and techniques reviewed, the supervised ML approach was selected for the ASB detection, as the approach is flexible, robust, and effective.

The review of the related work has shown that various feature sets and classification algorithms were used. Thus, in this work, experiments with several combinations of algorithms and feature sets were performed to identify those that were best suited to the research problem; however, the options were narrowed to the features and algorithms that have been widely used and have been shown to produce good results across various tasks and datasets.

Moreover, based on the review of techniques and resources used to identify emotions in text, it has been shown that the current resources are limited and that few fine-grained dimensional resources for analyzing emotions in text exist. In addition, the related works have shown that no publicly available corpus of ASB exists that can be used to study and analyze ASB for a more comprehensive understanding of the behavior.

Part II of the dissertation describes the experiments performed as well as the research that was conducted to address all the limitations highlighted.

Part II – Computational Implementations

6 TOWARDS THE DETECTION OF ANTISOCIAL BEHAVIOR

Every action that you take is stimulated by an emotion of some kind, either positive or negative. - Brian Tracy

In Chapter 5, it was concluded that there is no known set of features and techniques that are most effective for ASB detection. Thus, this chapter focuses on testing and comparing several features and techniques for ASB detection. Figure 6.1 presents the experimental framework. The framework involves three major phases: (i) data collection and preprocessing, (ii) feature extraction, and (iii) classification, evaluation, and interpretation. Each phase is discussed in detail in the next subsections.

6.1 PHASE 1: DATA COLLECTION AND PREPROCESSING

Based on the reviewed literature, it was identified that automatic detection of ASB in text is a novel research avenue. The lack of research in this area has largely been due to difficulties of access to textual data that contains ASB. Information pertaining to criminal, violent, and threatening behaviors is often difficult to obtain due to legal and privacy reasons; however, such data is needed to study and identify linguistic, stylistic, and emotional features pertaining to ASB. In addition, these features are useful in developing ML algorithms capable of predicting future incidences of ASB.

Thus, due to the lack of domain-relevant data, this section first describes the collection of ASB relevant data. In addition, the section presents the development of a lexicon for detecting emotions in text. The lexicon is based on the three dimensional model selected in Chapter 2. Both the collected ASB data and the lexicon serve as input to the ASB detection framework illustrated in Figure 6.1.

6.1.1 Corpora collection

Since a supervised binary ML classification approach was selected (see Chapter 5), both positive and negative examples of ASB were collected. From henceforth, the positive examples of ASB will be referred to as ASBT (ASB Texts) and the negative examples as non-ASBT. For the non-ASBT, a popular SA corpus (movie reviews [211]), an emotion annotated corpus (ISEAR [273]), and factual Wikipedia text extracts (Wikipedia[1]) were collected. Following from Greevy and Smeaton's [261] advice, since in reality the amount of generated ASB content is less than non-ASBT, an imbalanced corpus was utilized in developing the detection models. The following subsections describe each corpus.

ASB corpus

The author and Dr Kakkonen searched online for UGC that could conclusively be identified as being ASB. This process was guided by the research on ASB, which

[1] `https://www.wikipedia.org/`

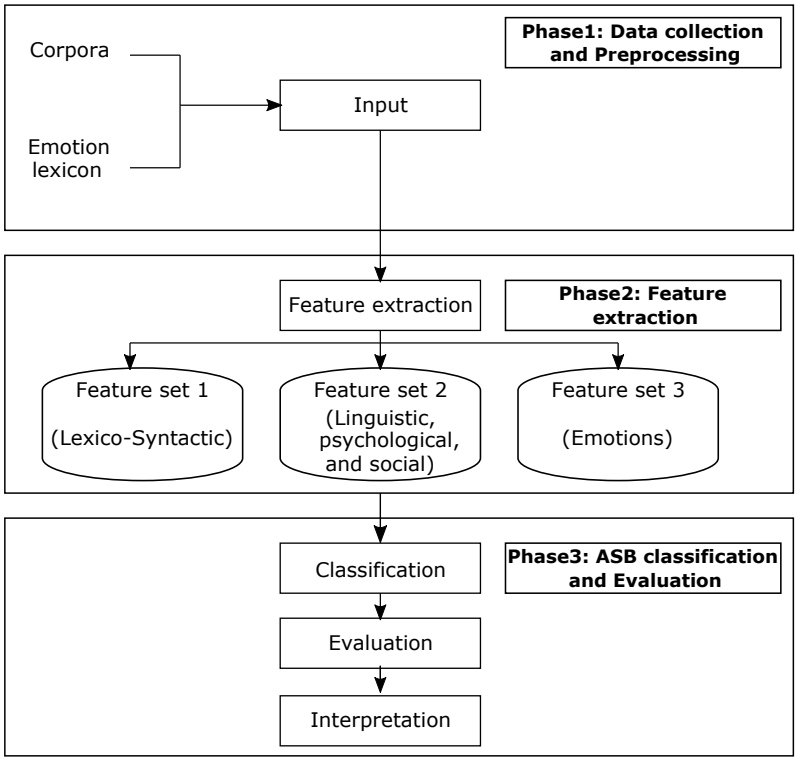

Figure 6.1: ASB detection framework.

revealed that aggression, violence, hostility, and lack of empathy were among the factors most associated with ASB (see Chapter 4). The UGC included blog posts, news articles, and song lyrics, and ranged from topics such as serial killing, terrorism, extremism, violence, and suicide. In total, 150 documents all in English, were identified as ASBT. In addition, the documents were written in a way that communicated the author's thoughts and emotions, and not those of other people.

The collected ASB documents allowed for the investigation of linguistic and emotional features of ASB and the identification of those features that are most beneficial for its prediction[2].

International Survey on Emotion Antecedents and Reactions (ISEAR)
The ISEAR corpus is a collection of student reports on situations in which the student felt any of the following seven emotions: joy, fear, anger, sadness, disgust, shame, and guilt [273]. The responses include descriptions of how a situation was appraised and the reactions to the situation.

[2]The current work-in progress version of the corpus is available to the research community.

Table 6.1: Corpora description with source, number of files, and average file sizes.

Corpus	Source	Documents	Avg. File Size (characters)
ASB	online	150	527
ISEAR	Scherer and Wallbott [273]	258	95
MovieReviews	Pang et al. [211]	179	3173
Wikipedia	wikipedia.org	205	470
Total		**792**	**4265**

Movie reviews
The research made use of the movie reviews corpus by Pang et al. [211]. The corpus is annotated with polarity data (negative vs. positive) and is often used in SA and opinion mining[3].

Wikipedia text extracts
A Wikipedia corpus was created through a similar approach as that used for the collection of the ASB corpus. Factors such as those found to be characteristic of ASB: killing, violence, aggression, terror, and frustration were used to search and extract Wikipedia articles. The aim of including these texts was to observe how well the developed detection models could distinguish between ASB texts and informative texts containing similar factors.

Table 6.1 summarizes the collected corpora. All corpora were preprocessed in order to remove irrelevant features (such as URL links, email addresses, and html tags).

6.1.2 Emotion lexicon

One of the research objectives was to investigate the presence and role of emotions in ASB detection. As discussed in Chapter 5, a common approach to analyzing emotions present in text is to make use of a lexicon of words annotated with emotional information. Thus, based on the analysis of emotions conducted in Chapter 2 and the limited coverage of existing lexicons (see Chapter 5), a comprehensive and fine-grained lexicon for identifying emotions in text was developed.

The development of the lexicon was based on the analysis of emotions conducted in Chapter 2, where a dimensional model of emotions with three dimensions: pleasure-displeasure, arousal-sleep, and intensity, was selected for the analysis of emotions in text. The dimensional model was an adaptation of the two-dimensional circumplex model by Russell [65]. In the *circumplex model*, emotions were viewed as a linear combination of two dimensions rather than being judged as belonging or not belonging to a specific basic emotion category. This allows for a "fuzzy" characterization of emotions. In addition, the circular nature of the circumplex model provides explicit information about the similarity of emotions. That is, emotions that are close together are very similar, while those that are opposite

[3]Opinion mining is the field dedicated to the research and detection of opinions in text [242].

are "maximally" different from each other [46]. A third dimension, intensity, was added to the circumplex model's two dimensions and this new model was called the *Circumplex-based Model for EmotioN AnalysiS* (CENSe).

In this section, the development of the CENSe-based lexicon by the researcher in cooperation with Dr Suero Montero and Dr Kakkonen is described.

Annotation tool

In order to create a CENSe-based annotated lexicon, a purposeful tool for the annotation was developed. The researcher was responsible for developing the annotation tool's functionality, Dr Kakkonen was responsible for developing the dictionary function in the tool and the analysis tool that is also described in this section, and Dr Suero Montero assisted in improving the code. The annotation tool presented a unified platform in which the three annotators could easily annotate words. In addition, the tool made it easy for the annotators to transverse and edit annotations.

Figure 6.2 presents a screenshot of the annotation tool. It illustrates how a word was annotated for valence, arousal, and intensity. The valence and arousal values were captured in the angle [0°, 360°], which describes the emotion *quality* of the word. In addition, the annotation tool included a sample of emotion qualities as guidelines (denoted with dashed lines). The sample emotion qualities were obtained from Ross's [274] study which involved 36 subjects classifying 28 emotion-bearing words by using the circumplex model. For example, the word *happy* was placed at 352°whereas *sad* was at 172°; however, unlike in the CENSe model, intensity was not considered in Ross's study.

Intensity, which captured the strength of the emotion, was annotated, as shown in Figure 6.2 by the distance from the neutral point to the placement of the word in the circle (following Reisenzein [103]). Intensity values ranged from the neutral 0.0 to the maximum intensity 1.0.

As an example, Figure 6.2 shows that the word *stimulate* was annotated with an emotional quality (angle) of 282°and 0.55 intensity, indicating that the word is pleasant and has a high level of arousal. An example of an unpleasant word with a high intensity value was blackmail, which was annotated with an emotion quality of 215°and with an intensity value of 0.70. In addition, the following words were annotated as being pleasant and with high arousal, but with varying degrees of intensity: studious (293°, 0.43), passionate (282°, 0.66), and magnificent (7°, 0.70). Hence, subtle nuances in emotional information were captured. With 360 degrees of freedom, the annotation tool provided an open canvas to accommodate complex emotional states.

Annotation process

The annotation was performed by the author and Dr Kakkonen and Dr Suero Montero. The annotation process began by using a popular emotion analysis resource, GI [244], as the seed list. The GI word list can be considered as the predecessor for all the resources that contain information relevant to identifying emotions in text. It consisted of 1,580 positive and 1,921 negative words.

The annotation process was divided into several phases. First, two test annotation rounds were conducted in which the aim was to calibrate the annotators' work in order to achieve as high level of agreement as possible and to determine the threshold values for acceptable levels of agreement on angles and intensities. After each test round, the annotators compared the results and discussed the reasons for disagreements. The discussion was necessary because of the complexity of the an-

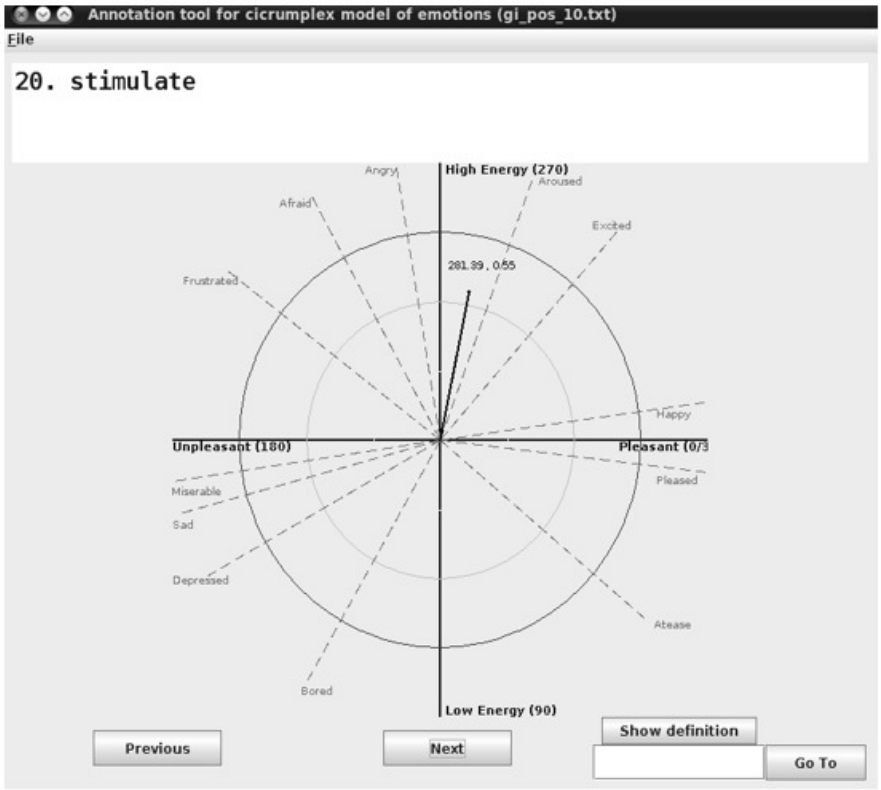

Figure 6.2: Annotation tool for CENSe (showing the word stimulate annotated with 281.99°and 0.55 intensity).

notation task: due to its implicit subjectivity, it is a difficult endeavor to achieve a high level of agreement on emotion annotation [275]. For the same reason the annotators chose not to have strict written annotation guidelines, but rather agree on some common annotation rules and to conduct a review meeting after each dataset had been annotated.

After the two test rounds, consisting of about 150 test words (both positive and negative), it was decided that the annotators had an acceptable rate of agreement for the actual annotation process to begin. The actual annotation process was also performed in rounds. After each annotation round, an analyzer tool was used to verify the agreement and produce lists of words that the annotators agreed on and those that they did not agree on. The analyzer tool and the results of these annotation rounds are reported on next. Furthermore, each annotation round was followed by a review meeting in which consensus was achieved for the list of words that the annotators did not agree on. Those words for which the annotators could not find an acceptable level of agreement were disregarded from the final annotated list.

Measuring inter-annotator agreement and reliability

In order to measure the agreement between the annotators and to evaluate the reliability of the annotations, a purpose-built Java tool, CENSe Analyzer, was implemented by Dr Kakkonen. The tool compared the annotation tool output files

it received as input and produced as output a pair-wise comparison between the annotators along with aggregate accuracy values for all the annotators. While the exact values are parameterized in the analyzer application, they were defined for this research as follows:

- At least two annotations had to be within 30°from each other for a word to be accepted as well-annotated.
- At least two annotations had to be within 0.15 from each other on the intensity scale for a word to be accepted as well-annotated.
- If either of the above requirements was not met, then the word was added to a list of words to be discussed in an annotation review meeting.

Under the above parameters, the CENSe Analyzer then produced a list of accepted words with the corresponding angle and intensity values, averaged over the closest two annotations (i.e., the consensus values). For those words where the annotators did not reach an agreement, either on the angle or the intensity or both, the analyzer outputted them as a list. This list served as the starting point for the discussions during the review meetings. Finally, the analyzer outputted a list of words to be ignored if they had too low an intensity to contribute to the list of emotion-bearing words. The intensity threshold was set to 0.25, i.e., if a word received an average intensity score that was less than 0.25, it was disregarded.

In addition to the annotator agreement rates, Krippendorff's alpha [276] was calculated to assess the reliability of the annotations. Since the late 1960s, Krippendorff's alpha has been known to be a useful statistical tool to measure the agreement among the observers of a phenomenon. Krippendorff's alpha in its general form is given by the following formula, where D_o is a measure of the observed disagreement and D_e is the expected disagreement when the values assigned to a phenomenon are given randomly by the observers:

$$\alpha = 1 - \frac{D_o}{D_e}$$

when $D_o = 0$, then $\alpha = 1$, which indicates that there is no disagreement in the data; hence the data is perfectly reliable. If $D_o = D_e$, then $\alpha = 0$, which implies little to no reliability in the data. Hence, reliability close to $\alpha = 1$ is desired.

Table 6.2 summarizes the results of all 24 annotation rounds. The "Annot." column within the table gives the total number of annotated words in each dataset. The column "Disreg." refers to the number of words that the analyzer automatically skipped due to low intensity (the average intensity was less than 0.25). These words were removed from the corpus. Examples of such words included amenable (0.1) and amnesty (0.13).

As shown in Table 6.2, a total of 3,380 words were annotated, out of which 3,121 (3,380-259) were either accepted directly to the final dataset or were to be discussed in an annotation review meeting. The consensus agreement between the two closest matching annotations was on average 87.8% for the angles and 95.4% for the intensities. The average pair-wise agreement scores between the annotators were 56.0%, 56.4%, and 60.8% for the angles and 69.2%, 74.0%, and 73.5% for the intensities. Excluding the Pos1 dataset, the pair-wise agreement rates for the angles ranged from 45.4% to 72.2% while the average agreement on intensities ranged between 42.3% and 90.3%. All these figures indicate that the data can indeed be reliably annotated by using the proposed CENSe model.

Table 6.2: Pair-wise agreement rates between annotators and the consensus agreement and Krippendorff's alpha for the annotators. A1, A2 and A3 refer to the three annotators respectively. "Pos" refers to a dataset of positive/pleasant words. "Neg" datasets consisted of negative/unpleasant words. The three first columns under the heading "Agreement" give the pair-wise agreement rates between the annotators. The two columns under "Consensus" give the consensus agreement rate and alpha scores. In all the cells, the first figure refers to angles and the second to intensities.

Dataset	No. of words		Agreement				
	Annot.	Disreg.	A1 vs.A2	A1 vs.A3	A2 vs.A3	Consensus	
						%	alpha
Pos1	144	2	29.1,25.0	34.5,13.5	44.6,59.6	79.1,77.0	0.77,0.62
Neg1	150	1	62.4,47.0	54.4,42.3	57.0,69.1	90.6,91.3	0.85,0.80
Pos2	142	1	53.2,63.8	56.7,58.2	45.4,67.4	85.8,90.8	0.82,0.74
Neg2	150	9	60.3,52.5	64.5,56.7	51.1,70.9	86.5,91.5	0.79,0.67
Pos3	140	5	63.0,68.9	50.4,62.2	65.9,71.9	91.9,96.3	0.86,0.81
Neg3	150	6	50.7,85.4	64.6,84.7	68.8,84.7	90.3,98.6	0.81,0.75
Pos4	146	2	70.1,79.2	61.8,89.6	72.2,81.9	95.1,98.6	0.91,0.82
Neg4	150	7	61.5,65.0	72.0,88.1	63.6,65.0	90.1,97.9	0.82,0.58
Pos5	146	10	58.8,75.7	58.2,75.7	64.0,75.7	86.0,97.8	0.73,0.73
Neg5	150	14	59.6,65.4	62.5,80.1	59.6,71.3	88.2,93.4	0.79,0.60
Pos6	149	9	49.3,62.1	47.4,76.4	52.1,72.9	85.0,95.0	0.61,0.75
Neg6	150	21	63.8,66.2	52.3,83.1	69.2,70.0	91.5,94.6	0.86,0.74
Pos7	148	21	53.5,64.6	47.2,73.2	60.6,78.0	88.2,95.3	0.78,0.78
Neg7	150	6	56.3,73.6	56.9,90.3	56.3,77.1	90.3,97.9	0.82,0.67
Pos8	135	10	56.8,65.6	52.0,65.6	64.0,64.8	82.4,93.6	0.62,0.56
Neg8	150	17	53.4,83.4	54.9,86.5	67.7,78.2	89.5,99.2	0.85,0.86
Pos9	150	11	54.0,71.2	61.2,85.6	65.5,77.7	89.2,99.3	0.81,0.73
Neg9	150	14	54.4,77.9	55.1,77.2	66.9,80.1	91.1,96.3	0.84,0.79
Pos10	143	21	50.0,69.7	60.7,78.7	66.4,82.0	85.2,96.7	0.80,0.70
Neg10	150	20	60.8,75.4	50.8,81.5	49.2,73.8	83.4,96.2	0.75,0.71
Pos11	50	5	48.9,82.2	57.8,82.2	60.0,75.6	80.0,97.8	0.70,0.79
Neg11	150	22	61.7,81.3	49.2,71.9	59.4,71.9	83.6,96.9	0.83,0.75
Neg12	150	17	63.9,85.0	70.7,90.2	70.7,77.4	94.7,99.2	0.83,0.76
Neg13	87	8	48.1,74.7	58.2,83.5	59.5,67.1	89.8,97.5	0.84,0.81
TOTAL	**3380**	**259**	**56.0,69.2**	**56.4,74.0**	**60.8,73.5**	**87.8,95.4**	**0.80,0.73**

Furthermore, column "alpha" in Table 6.2 gives Krippendorff's alpha scores for the two closest matching annotations for the angles and intensities. Taking into account the subjectivity of the data, the alpha scores of 0.80 and 0.73 for the angles and intensities respectively, also indicate a reliably annotated dataset.

From Table 6.2, it can be observed that the first dataset, Pos1, had the lowest agreement values. This was mostly due to the lack of initial consensus over how to annotate some particularly challenging words, for example, with the word *acknowledgement*, the annotators had about a 70°difference in the angles. Other examples of challenging words to annotate included *accession*, *allow*, and *alive*, to name a few. Within Pos1, annotator 1 in particular had a low level agreement with the other two annotators. However, as can be observed in Table 6.2, the trend in the agreement rate increased as the annotation work progressed. This was due to the synchronization

of the annotations based on agreed guidelines during the annotation review meetings. Hence, as of the Pos3 dataset, angle and intensity consensus agreement rates were 91.9% and 96.3% respectively. The highest agreement rates on angles were observed on the Pos4 set (95.1%). The highest intensity agreement was obtained on Pos9 (99.3%).

Final dataset

All the words in which no pair of annotators agreed within the set threshold values (30°for angles, 0.15 for intensities) were discussed by the three annotators. Following the results of each annotation review meeting, the consensus agreement rate was 100%. A total of 509 words were reviewed (Table 6.3). Thus, in the final dataset, there were 2,974 words, i.e., 147 of the 3,121 words from the initially annotated set were removed during the review meetings. These words were removed from the dataset either because the annotators could not reach an agreement or the word in question was deemed to be ambiguous (i.e., a contextual word). The words belonging to the latter category will be investigated in more detail in future work.

Table 6.3: Final dataset.

Dataset	No. of words			
	All	**Reviewed**	**Removed**	**Accepted**
Pos1	142	66	0	142
Neg1	149	28	4	145
Pos2	141	29	5	136
Neg2	141	26	4	137
Pos3	135	15	2	133
Neg3	144	15	11	133
Pos4	144	10	1	143
Neg4	143	20	7	136
Pos5	136	22	5	131
Neg5	136	24	12	124
Pos6	140	25	12	128
Neg6	129	16	3	126
Pos7	127	27	13	114
Neg7	144	16	2	142
Pos8	125	29	10	115
Neg8	133	14	6	127
Pos9	139	16	5	134
Neg9	136	16	7	129
Pos10	122	22	9	113
Neg10	130	22	10	120
Pos11	45	10	5	40
Neg11	128	22	4	124
Neg12	133	9	5	128
Neg13	79	10	5	74
Total	**3121**	**509**	**147**	**2974**

Examples of the contextual words that were removed during the review meetings included *contrary*, *cost*, *costliness*, *costly*, *craze*, *craziness*, *crazy*, *dark*, and *darken*, which were removed due to their ambiguity. Another example, was the word *covert*,

which was totally excluded from analysis since the annotators were not able to reach an agreement on how to annotate it. Such words were referred to as *genuine disagreements*.

The final set of words and their annotations were written in an *EmotionML* (Emotion Markup Language) document that provided a standard interface between the annotation tool and other components in the CENSe-based tagging system (discussed in Section 7.4.1). EmotionML is an XML based language that was introduced by the World Wide Web Consortium as a working draft standard for representing emotions in text [277]. It is used as a "plug-in" for three different areas: manual annotation of emotion, automatic recognition of emotion, and generation of emotion.

EmotionML defines a set of vocabularies for representing emotion-related states [278]. This set, however, does not support the CENSe representation of emotions, thus the research defined a CENSe-based EmotionML vocabulary, whereby an emotion element in the EmotionML document was represented as follows:

```
<EmotionWord>
   <String>stimulate</String>
   <Intensity>0.495</Intensity>
   <Angle>317.2795</Angle>
</EmotionWord>
```

Summary

This section described the creation of an emotion lexicon based on the CENSe model. The final resource consisted of 2,974 words which allowed for a "fuzzy" characterization of emotions as they are considered as a linear combination of these dimensions rather than belonging to a specific basic emotion category. Although the valence, arousal, and intensity representation of emotions was not a novel approach, it had not been applied to annotating a large resource of emotion-bearing words as was created in this research. The model is particularly beneficial when dealing with real-world data that include a wide range of emotional states.

6.2 PHASE 2: FEATURE EXTRACTION

Features selected to represent documents play a crucial role in the performance of ML detection models; therefore, the feature extraction phase described in this section, sought to identify those features that efficiently characterize ASB in text. As mentioned in Chapter 5, no standard feature set exists, and as ASB is a new area of investigation in NLP, a variety of features were investigated and their significance for the detection of ASB was analyzed.

The research selected features that reflected word usage, writing styles, and emotions. In selecting the features, the research did not attempt to infer or interpret the meaning of what was written but aimed at developing detection models based on explicit features or near surface features. These features included word-level, sentence structure, and semantic features such as emotions. The selected features were grouped into three groups such as lexico-syntactic; linguistic, psychological, and social-based; and semantic features. Table 6.4 summarizes the features.

All features listed in Table 6.4 are beneficial in capturing writing styles and the word usage of the collected corpora. The lexico-syntactic features (POS and word n-grams) determine how an author constructs a message [242]. The linguistic, psycho-

Table 6.4: Features and a sample of examples.

Features	Examples
Lexico-Syntactic	Frequency of POS tags (e.g., NP and VB) Word n-grams (e.g., "fight" and "kiss")
Linguistic, psychological, and social states	Total words, % words per text, word length, pronouns, verb tenses, social words, biological process words, etc.
Emotion-based	Emotion dimensions (e.g., valence, arousal, and intensity)

logical, and social state features provide richer information regarding word usage than the n-gram features [279]. In addition, the semantic features reveal the emotions present in text. The semantic information will be obtained from the CENSe resource created in Section 6.1.2. The benefits of these features are evaluated for ASB detection in the next chapter.

6.3 PHASE 3: ANTISOCIAL BEHAVIOR CLASSIFICATION AND EVALUATION

As the research adopted a binary supervised ML classification approach to detect whether a given text is ASB or not, a series of experiments to empirically evaluate efficient combinations of classification algorithms and features sets for the detecting of ASB, were designed. The classification algorithms included three frequently used classification algorithms in TC, NB, SVMs, and C4.5 decision trees.

Naive Bayes
The naive Bayes (NB) algorithm is a probabilistic algorithm where the input is assumed to be independent. The NB algorithm estimates the probability of a class given the data to be proportional to the probability of the class multiplied by the probability of the data given the class [206]. In other words, the NB classifier assigns a given document d the class $c* = argmax_c P(c|d)$ [211].

Support vector machine
Support Vector Machines (SVM) are based on the concept of decision planes that define decision boundaries. A decision boundary is a hyperplane that separates a set of objects in one class from another [280]. In particular, a SVM classifier aims to find a hyperplane that is represented by a vector that maximally separates the document vectors in one class from those in another [211].

C4.5 decision tree
C4.5 is an algorithm that generates decision tree classifiers, a structure that is either a leaf indicating a class or a decision node that specifies some test to be carried out on a single attribute value, with one branch and subtree for each possible outcome of a test. A decision tree can be used to classify a document by starting at the root of the tree and moving through it until a leaf is encountered. At each non-leaf decision node, the instance's outcome of a test at the node is determined and the attention is shifted to the root of the subtree corresponding to this outcome. When this process

finally (and inevitably) leads to a leaf, the class of the document is predicted to be that recorded at the leaf [281].

In addition, the evaluation of the classification algorithms was performed using the *k*-fold cross validation approach that was discussed in Section 5.3.4. The approach was found to be beneficial for an imbalanced ASBT and non-ASBT corpus. The experiments with the described three classification algorithms and feature sets are presented next in Chapter 7.

7 LEVERAGING FEATURES FOR ANTISOCIAL BEHAVIOR DETECTION

People are their emotions. To understand who a person is, it is necessary to understand emotion. – Denzin (1984)

7.1 EXPERIMENT DESIGN

This chapter presents experiments to identify efficient features for the detection of of ASB in text. For each experiment, a binary classifier was developed with the three classification algorithms explained in Section 6.3 to detect whether a given message was ASB or not. Each classifier was trained on the datasets described in Section 6.1.1. The *Waikato Environment for Knowledge Analysis* (WEKA) software suite for ML [282] and its implementation of the three classification algorithms discussed in Section 6.3 was used for the development of the detection models. For the probabilistic model, WEKA's NB classifier was selected, while the sequential minimal optimization (SMO), WEKA's implementation of SVM, and J48, an implementation of the C4.5 decision tree, were also selected. These are explained in Table 7.1.

To begin the experiments, the datasets were first preprocessed by tokenizing each document, removing stop words, stemming, and removing rare words with a frequency of less than three. After preprocessing, an ARFF file (Attribute Relation File Format) was generated, whereby each data entry[1] had two fields; a text field consisting of the document's textual content and a binary class label (i.e., a 1 = ASB or 0 = non-ASB).

In addition, for all experiments, the ten-fold cross validation approach, whereby samples of the corpora are reserved for testing while the rest are used for training a model was applied. The performance of the classifiers was assessed and compared in terms of accuracy, precision, recall, and F-measure. More specifically, the balanced F-measure, which gives equal weights to recall and precision was adopted. The performance of all the three classifiers was further compared to baseline values that were obtained from a primitive classifier that basically predicted the majority class. Such a classifier gave an idea of the minimal performance that any classification algorithm should be able to obtain [285]. In addition, it is a beneficial baseline for the experiments since there was a skewed distribution of ASBT to non-ASBT. To develop the baseline classifier, WEKA's implementation of the ZeroR classification algorithm was utilized. All experiments are described in Sections 7.2 to 7.4.

7.2 LEXICO-SYNTACTIC FEATURES

The first experiment involved the use of the following lexical and syntactic features: word unigrams, bigrams, and POS-bigrams.

[1] A data entry in all the experiments was the whole document, thus document classification was being performed.

Table 7.1: Description of used algorithms.

Algorithm	Description
NB	The NB classifier implemented in WEKA is based on the Bayes' theorem with independence assumptions between predictors as described in Section 6.3. That is, the NB classifier assigns the most likely class to a given example described by its feature vector and assumes that the features are independent given the class. Furthermore, it assumes that no hidden or latent attributes influence the classification [283].
SMO	SMO represents the "sequential minimal optimization" algorithm which is an improved training algorithm for SVMs. Normally, training an SVM requires the solution of a very large quadratic programming (QP) optimization problem. SMO breaks this QP problem into a series of smallest possible QP problems, which are solved quickly and analytically, generally improving its scaling and computation time significantly [284].
J48	J48 is an implementation of Quinlan's C4.5 algorithm for generating pruned or unpruned C4.5 decision trees. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used [281].

7.2.1 Unigrams

The initial experiment involved representing each document with the individual words (unigrams) occurring in it. That is, a BOW approach was taken and the unigram features were represented in the VSM format that was described in Section 5.3.1. The BOW approach considers all words as independent features and makes an implicit assumption that the order of the words in a document does not matter [206]. The unigram feature set generally works well for text classification, as the collection of words appearing in the document (in any order) is usually sufficient to differentiate between semantic concepts [206]; however, it does not take into consideration any semantic and contextual information.

Three supervised binary ML algorithms using different combinations of the ASBT and non-ASBT datasets were trained. Combinations such as the ASB corpus together with the ISEAR corpus (ASB + ISEAR), with movie reviews (ASB + MovieReviews), with the Wikipedia corpus (ASB + Wiki), and finally with all corpora combined (All) were part of the experiments. Table 7.2 displays the average of the ten-fold cross validation results on the corpora for each algorithm.

Table 7.2: Results for the ASB classification using the unigram feature set (%).

Corpora	Classifier	Accuracy	Precision	Recall	F-measure
ASB + ISEAR	NB	91.67	93.28	83.33	88.03
	SMO	**94.12**	**100.00**	**84.00**	**91.30**
	J48	86.52	94.39	67.33	78.60
	Baseline	63.24	-	0.00	0
ASB + MovieReviews	NB	96.96	99.30	94.00	96.58
	SMO	99.39	98.68	**100.00**	99.34
	J48	**99.70**	**99.34**	**100.00**	**99.67**
	Baseline	54.41	-	0.00	0.00
ASB + Wikipedia	**NB**	**97.19**	**98.61**	**94.67**	**96.60**
	SMO	94.67	94.56	92.67	93.60
	J48	87.92	87.41	83.33	85.32
	Baseline	57.87	-	0.00	0.00
All	NB	64.65	32.62	**81.33**	46.56
	SMO	**94.57**	**94.21**	76.00	**84.13**
	J48	91.54	86.73	65.33	74.52
	Baseline	81.06	-	0.00	0

Based on the results shown in Table 7.2, the ML algorithms NB, SMO, and J48 achieved high accuracies, with SMO performing the best in two of the corpora (ASB + ISEAR and All). With the 'All' corpora, SMO achieved the highest accuracy with about 95%. In terms of precision and recall, all the classifiers outperformed the baseline.

As identifying an efficient ASB detection model is one of the research objective, performance of the three ML algorithms for the 'All' corpora was examined in terms of their TP, FN, FP, and TN results (presented in Table 7.3). The comparative table not only allows for the description of classification accuracy but the characterization of errors as well [286].

Table 7.3: Comparative table of TP, FN, FP, and TN for the 'All' corpora.

Corpora	Classifier	True Pos.	False Neg.	False Pos.	True Neg.
All	NB	122	28	252	390
	SMO	114	36	7	635
	J48	98	52	15	627
	Baseline	0	150	0	642

From the table, it can be seen that even though the SMO classifier had the higher accuracy levels (95%), the NB classifier captured more of the ASB instances (122 vs. 114). The NB classifier, however, performed poorly in detecting the non-ASB instances, hence the low accuracy rate observed in Table 7.2, where as the SMO classifier was able to capture majority of the non-ASBT.

Results from the first experiment have illustrated that the developed classifiers can successfully distinguish ASBT from non-ASBT, with the SMO and J48 classifiers performing the best for the 'All' corpora.

7.2.2 Bigrams

The next experiment investigated the impact that sequences of word pairs (i.e., bigrams) had on ASB classification. Table 7.4 shows the average of the ten-fold cross validation results on the corpora for each algorithm using the bigram feature set.

Table 7.4: Results for the ASB classification using the bigram feature set (%).

Corpora	Classifier	Accuracy	Precision	Recall	F-measure
ASB + ISEAR	NB	70.59	**100.00**	20.00	33.33
	SMO	**75.98**	**100.00**	**34.67**	**51.49**
	J48	63.24	-	0.00	0.00
ASB + MovieReviews	**NB**	**93.62**	**91.08**	95.33	**93.16**
	SMO	86.89	77.72	**100.00**	87.46
	J48	54.27	-	0.00	0.00
ASB + Wikipedia	NB	69.66	**95.65**	29.33	44.90
	SMO	**70.79**	92.59	**33.33**	**49.02**
	J48	57.87	-	0.00	0.00
All	NB	83.84	**95.83**	15.33	26.44
	SMO	**84.97**	79.25	**28.00**	**41.38**
	J48	81.06	-	0.00	0.00

In comparing the bigram experiment results with those of unigrams, Table 7.4 illustrated that bigrams were not as efficient as the unigrams in distinguishing between ASBT and non-ASBT. It was found that the SMO classifier performed better on three corpora (ASB+ISEAR, ASB+Wiki, and All) while the NB classifier performed better on the ASB+MovieReviews. In particular, the J48 classifier performed poorly in terms of precision and recall, showing the same performance results as the baseline values (accuracy = 81.06%, precision = n/a, and recall = 0%). Although the accuracy levels, for instance for the 'All' corpora was 81%, a closer look at the TP, FP, FN, and TN values revealed that the J48 classifier was unable to capture the ASBT instances and the accuracy results were based on its ability to detect non-ASBT (see Table 7.5). As a consequence, the recall and F-measure values for the J48 classifier were 0, and since true positive and negative values were both 0, precision was not applicable.

Table 7.5: Comparative table of TP, FN, FP, and TN for 'All' corpora.

Corpora	Classifier	True Pos.	False Neg.	False Pos.	True Neg.
All	NB	23	127	1	641
	SMO	42	108	11	631
	J48	0	150	0	642

Further comparison of the experimentation results with bigrams and unigrams, revealed that the SMO and NB classifiers achieved low recall values in three of the corpora, with the exception of the ASB+MovieReviews corpora where the recall results were higher. The reason behind the poor performance could have been a result of the relatively small ASB corpus which could have led to sparseness of data when bigrams were used, thus affecting the accuracies of the classifiers. In addition, when comparing the classification features for 'All' corpora, it was observed that 458 bigram features were utilized while in the first experiment, 3,880 unigram features

were utilized. From Table 7.5, it was observed that the decision tree algorithm (J48) with the bigram feature set resulted in poor classification performance, which could be because decision tree algorithms require sufficient sample training data to build a decision tree [287]. Overall, the second experiment with bigram feature set, revealed that the classifiers had a tendency to classify documents into the majority class which was the non-ASBT class.

7.2.3 POS-bigrams

POS-bigrams as features can be used to capture stylistic information such as distinction between "the answer, which . . ." and "which is the answer" [221]. Table 7.6 shows the average of the ten-fold cross validation results of the corpora combinations for each algorithm using the POS-bigrams feature set.

Table 7.6: Results for the ASB classification using the POS-bigram feature set (%).

Corpora	Classifier	Accuracy	Precision	Recall	F-measure
ASB + ISEAR	NB	89.46	89.05	81.33	85.02
	SMO	**91.67**	**92.65**	**84.00**	**88.11**
	J48	84.56	85.37	70.00	76.92
ASB + MovieReviews	NB	97.87	97.99	97.33	97.66
	SMO	**99.39**	**98.68**	**100.00**	**99.34**
	J48	97.56	97.33	97.33	97.33
ASB + Wikipedia	NB	89.61	88.97	86.00	87.46
	SMO	**91.29**	**92.20**	**86.67**	**89.35**
	J48	83.99	83.45	77.33	80.28
All	NB	48.86	24.65	**82.67**	37.98
	SMO	**90.78**	**79.39**	69.33	**74.02**
	J48	87.50	72.57	54.67	62.36

POS-bigrams as shown in Table 7.6 lead to high accuracy rates, especially when compared to the rates with the bigram feature set. This could be because the number of POS-bigram features utilized for the classification were higher in number than the bigram features (753 vs. 458 in the 'All' corpora). More specifically, the SMO classifier performed the best with all the four corpora. Notably, the NB classifier performed poorly on the 'All' corpora in terms of precision (24.65%) and accuracy (48.86%) levels, with accuracy levels lower than the baseline (81.06%); however, it achieved higher recall rates (82.67% vs 0%). In looking at the TP, FP, FN, and TN results with the POS-bigrams features on the 'All' corpora, the NB classifier especially misclassified the non-ASBT as ASBT (FP=379 and TN=263). This could also be due to the imbalanced datasets that might affect the decision boundary of the NB classifier.

7.2.4 Unigram, bigram, and POS-bigram combinations

The contribution of each lexico-syntactic feature set to the classification accuracy rates was investigated on the 'All' corpora with the SMO classifier, since the SMO classifier achieved the best classification performance in all three experiments on the 'All' corpora. The resulting accuracy rates are displayed in Figure 7.1.

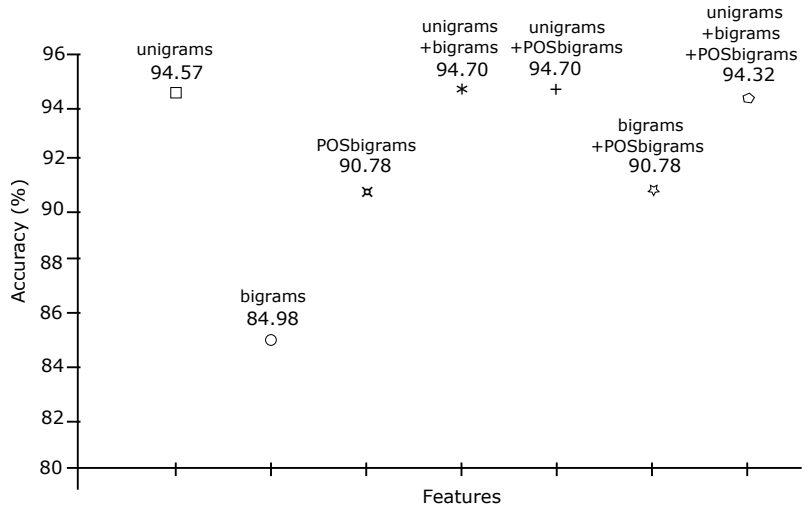

Figure 7.1: Comparison of accuracy rates achieved with the lexico-syntactic features.

From Figure 7.1, it can be observed that combining unigrams with either bigrams or POS-bigrams resulted in the highest accuracy rate of about 95%. Bigrams as identified earlier, resulted in the lowest accuracy. In addition, looking at the top three best accuracy results (see Table 7.7), it was possible to identify that combining POS-bigrams with unigrams increased the number of ASBT instances accurately detected as ASBT; however, more of the non-ASBT instances were also misclassified (see Table 7.7).

Table 7.7: Comparative table of TP, FN, FP, and TN of the top three best accuracy rates in Figure 7.1.

Features	**Accuracy (%)**	**True Pos.**	**False Neg.**	**False Pos.**	**True Neg.**
Unigrams	94.57	114	36	7	635
Uni + Bigrams	94.70	116	34	8	634
Uni + POSbigrams	94.70	120	30	12	630

From the results, it is apparent that unigrams seem to be efficient features for ASB classification. To analyze the unigrams more deeply, a feature selection filter - *information gain* - which is often used in ML to identify how well each single feature classifies a given dataset [288], was applied. The filter was applied to the best performing classifier, SMO with the unigram feature set. It was identified that the unigrams contributing most to the classification included words such as: nigger, fu*k, shit, kill, will, all, white, like, etc., which indicate that the ASBT consist more of profanity and insult related language (e.g., nigger and chink).

7.3 LINGUISTIC, PSYCHOLOGICAL, AND SOCIAL FEATURES

Writing styles are influenced by an extensive array of factors such as text genre [289], thus, this section focuses on investigating linguistic features that are stylistic in nature and appear in writing. In contrast to content words (e.g., nouns, regular verbs, adjectives, and adverbs), stylistic features (also referred to as particles or functional words) (e.g., pronouns, prepositions, articles, and conjunctions) describe "how" people talk or write. Stylistic features were found by Pennebaker [290] in their development of the well-known text analysis tool - the Linguistic Inquiry and Word Count (LIWC) - to be far more interesting than content-based features since linguistic styles provide far richer psychological information. Research has also shown that the presence and frequency of stylistic features in writings can be indicative of certain psychological states (see [289]) and as Chung and Pennebaker [152] explained, can reveal behavioral intent. Thus, it was hypothesized that stylistic features might be useful in revealing harmful behavior intent such as ASB. In order to analyze and extract the linguistic and stylistic features in the ASBT and non-ASBT datasets, the LIWC tool was applied.

7.3.1 The LIWC tool

The LIWC tool is based on a word count strategy which takes on the assumption that the "words people use convey psychological information over and above their literal meaning and are independent of their semantic context" [291]. The tool calculates the degree various categories of words across a wide array of text genres are used [6]. LIWC works by searching and counting both content and style words within any given text and comparing it to its dictionary of almost 4,500 words and word stems and assigning them to linguistically, socially, and psychologically meaningful categories. There are in total 74 categories (e.g., articles, pronouns, cognitive, emotions, and social which have been validated by independent judges) [6]. For example, the word 'cried' falls into five categories: sadness, negative emotion, overall affect, verb, and past tense verb [279].

The linguistic categories of LIWC have been shown to be reliable markers for a number of psychologically meaningful constructs [6, 291]. For example, the categories have been shown to be indicative of psychological disorders such as depression [292–294]. Furthermore, first person singular pronouns were linked to negative emotions [152, 295], and were found useful in the analysis of coherence, disclosure, and online participation among females and males in an Internet cancer support group [296].

LIWC takes as input a corpora and analyzes at word-by-word basis comparing each word in a given file to words and word stems in the LIWC internal dictionary [297]. LIWC then calculates for each input file the number of words that match each of the 74 LIWC categories. This is expressed in percentages of the total words in the text [297]. It is these 74 categories that were taken as classification features for the ASB classification task.

This section investigates which and whether the LIWC features form good features for ASB detection. The use of LIWC features for harmful behavior detection is to the author's knowledge novel work in the research field.

7.3.2 LIWC feature set

To analyze the linguistic style of the datasets, two sets of experiments on the 'All' corpora using LIWClite7 were conducted. The first experiment tested how well the LIWC features performed by as features for detecting ASB. The second experiment combined the LIWC features with the unigram feature set. Table 7.8 summarizes the average of the ten-fold cross validation results on the 'All' corpora for each classification algorithm.

Table 7.8: Results for the ASB classification using the LIWC features (%).

Features	Classifier	Accuracy	Precision	Recall	F-measure
LIWC	NB	94.07	81.21	**89.33**	85.08
	SMO	**95.20**	**93.08**	80.67	**86.43**
	J48	94.44	86.30	84.00	85.14
Unigram + LIWC	NB	93.81	79.89	**90.00**	84.64
	SMO	95.58	**92.59**	83.33	87.72
	J48	**95.83**	92.09	85.33	**88.58**

From Table 7.8, it can be observed that the SMO classifier performed the best when only the LIWC features were used (95.20%) and the J48 classifier was the best performing classifier when both the LIWC and unigram features were combined (95.83%). With the LIWC features, all three classifiers performed better than the baseline values (accuracy = 81.06% , precision = n/a , and recall = 0%). The classification results of the SMO and J48 classifiers offered substantial support that ASBT exhibit common stylistic features and that these features are distinguishable from non-ASBT. Especially, it was found that the LIWC features alone performed better than the unigram features (94.57%) in Section 7.2.1. More specifically, by analyzing the TP, FP, FN, and TN values in Table 7.9, it was observed that the LIWC features did indeed improve the detection of ASBT in terms of true positives. With the LIWC features, the SMO classifier was able to detect seven more ASBT instances than with the unigram features alone, and fourteen more when the LIWC features were combined with unigrams.

Table 7.9: Comparative table of TP, FN, FP, and TN of the top three best accuracies.

Features	**Classifier** (%)	**True Pos.**	**False Neg.**	**False Pos.**	**True Neg.**
Unigrams	94.57	114	36	7	635
LIWC	95.20	121	29	9	633
LIWC + Unigrams	95.83	128	22	11	631

In addition to the improved accuracy, another advantage of the LIWC features over the unigrams was that the size of the feature set was much smaller (74 vs. 3880) which lead to the SMO classifier performing much faster on the LIWC feature set than on the unigram feature set.

In addition, to identify the LIWC feature categories that were the best discriminants for ASB classification, the feature selection filter - information gain was applied. The filter was applied to the best performing classifier, SMO. Table 7.10 lists in no particular order the 15 LIWC categories that were the top in discriminating ASBT from non-ASBT as well as examples of words belonging to each category.

Table 7.10: Top contributing 15 LIWC features to ASB classification (Examples taken from Pennebaker et al. [297].

No.	LIWC categories	Examples
Linguistic processes		
1	Swear words	Damn, piss, and f*ck
2	Word Count	
3	Question marks	?
4	Words > 6 letters	Underprivileged and eliminated,
5	Total pronouns	I, them, and itself
6	Personal pronouns	I, them, and her
7	1st-person singular	I, me, and mine
8	2nd-person singular	You, your, and thou
9	Present tense verbs	Is, does, and hear
10	Future tense verbs	Will and gonna
Psychological processes		
11	Negative emotion	Hurt, ugly, and nasty
12	Anger	Hate, kill, and annoyed
13	Biological processes	Eat, blood, and pain
14	Body	Cheek, hands, and spit
15	Sexual	Horny and incest

The top 15 categories were grouped in Table 7.10 under the process names given by Tausczik and Pennebaker [6]. From Table 7.10, it can be observed that the top 15 categories were under the linguistic and psychological processes, with 10 out of 15 belonging to linguistic processes. In addition, Table 7.11 presents the differences between the top 15 LIWC categories in the ASB, Wikipedia, ISEAR, and MovieReviews corpora.

It can be observed from Table 7.11 that the ASB corpus particularly differed from the three non-ASBT datasets (ISEAR, MovieReviews, and Wikipedia) in the following categories: swear words, 2nd-person singular, anger, negative emotion, future and present tense verbs, and words related to the body and sexuality. Furthermore, the ASB corpus contained more swear words than the non-ASBT datasets. Swearing was described by Montagu [298] as a "culturally acquired way of expressing anger". Swear words were also identified by Tausczik and Pennebaker [6] as being correlated with the use of informal language and aggression. Aggression and anger were among the main ASB factors identified in Chapter 4. In Chapter 4, negative emotions were also identified as being a factor of ASB and this can be observed in Table 7.11. As Table 7.11 shows, the ASB corpus contained more negative emotions, particularly anger, than the three non-ASBT datasets.

In addition, it was observed that ASBT had higher presence of 2nd-person singular pronouns. As Tausczik and Pennebaker [6] explained, personal pronouns provide us with information about the subject of attention and this is in line with the research definition of ASB, which specifies that the focus of harm is another person or community.

Moreover, looking at the tense of the verbs used mostly in ASBT, it was found that the temporal organization of the communication in ASBT was more in the present tense as well as the future tense in comparison to the non-ASBT. The ISEAR corpus in particular scored highest on the past tense category while the Wikipedia

Table 7.11: Mean scores for the top 15 LIWC categories in the four datasets. Grand Means is the unweighted means of the four datasets; Mean SDs refer to the unweighted mean of the standard deviations across the four datasets.

LIWC categories	ASB	ISEAR	Movie-Reviews	Wikipedia	Grand Means	Mean SDs
Linguistic processes						
Swear words	**3.27**	0.02	0.10	0.35	0.94	1.56
Word Count	17880	5824	**120323**	17660	40422	53564
Question marks	**0.38**	0.00	0.00	0.00	0.10	0.19
Words > 6 letters	15.30	16.60	20.31	**32.30**	21.13	7.74
Total pronouns	16.45	**19.47**	9.80	4.30	12.51	6.80
Personal pronouns	11.53	**14.70**	4.81	0.86	7.98	6.29
1st-person singular	3.85	**11.11**	0.90	0.06	3.98	5.02
2nd-person singular	**3.66**	0.19	0.42	0.06	1.08	1.72
Present tense verbs	**9.02**	2.80	7.28	4.78	5.97	2.74
Future tense verbs	**1.49**	0.24	0.61	0.45	0.70	0.55
Psychological processes						
Negative emotion	**5.68**	3.21	2.42	4.08	3.85	1.40
Anger	**4.19**	1.27	0.89	2.34	2.17	1.48
Biological processes	**4.09**	1.32	1.31	2.73	2.36	1.33
Body	**1.62**	0.50	0.48	0.37	0.74	0.59
Sexual	**1.87**	0.24	0.30	0.70	0.78	0.76

and MovieReviews were more in the present tense. With the ISEAR corpus, this finding was expected as the corpus consisted of documents where subjects were asked to remember and report on past events and feelings. The use of high past tense in communicating events was also identified in Hancock et al.'s [299] work on the language of psychopaths, where they identified that subjects used mostly the past tense when writing about harm they had committed. Thus, by identifying in this work that ASBT contained more future and present tense, might be indicative of leakage, which is referred to by Brynielsson et al. [92] as the communication to a third party of an intent to do harm to another. In their study on detecting lone wolf terrorists and violent extremism in online environments, Brynielsson et al. [92] and Johansson et al. [300], respectively found that this leakage is likely to contain intent signaling future tense verbs such as "...will..." or "going to" followed by an expression of a violent action, e.g., beat or destroy. Examples of such occurrences in the ASB corpus were also found, for example: "I will kill every motherfuc*king

one of you"; "This will be deadliest school shooting in American..."; "I will write my manifesto in her spilled blood"; "I will get revenge on the students and teachers"; "I'm going to make sure they die"; or "I am going to chop off your arm."

In addition, it was found that ASBT contained a higher reference to body and sexual related words than non-ASBT. Upon a closer look at the ASBT, it was identified that these references were mostly related to insults. This is in line with Jay's [301] finding that body and sexual terms were related to the semantics of insults and anger. That is, angry or aggressive persons will use high taboo or dirty words to offend, and these terms stem from body parts, body sexual processes, and ethnic terms, among others. Hence, it was observed that the ASBT contained body related insults that used words such as "dick," "cunt," "asshole," or sexual related terms such as "get screwed," and "jerk off." The presence of insults in ASBT as an expression of anger confirms the research findings in Chapter 4 that anger was a characteristic of ASB.

7.4 EMOTION FEATURES

This section explores the role that emotions as features play in distinguishing ASBT from non-ASBT. The research focused on identifying explicit expression of emotions despite what the writer might privately believe. This is because a user might be insecure or be deceitful, or express feigned emotions. The research adopted Irvine's [150] perspective that, "depending on the nature of the acts involved, it does not always matter whether the speaker (or actor) really has the feeling attributed to him. Sometimes what is much more important is simply the public expression of some feeling - so that the public social order can continue, despite what anyone may privately believe." For example, a person may express condolences over a death without actually feeling sympathy. "Such insincere acts are not void, but merely infelicitous; they still "count", whether the feeling is present or not" [150]. Thus, it is beyond the scope of this research to delve into whether the emotions expressed in text are the writer's actual emotions or to identify the extent to which the emotions are sincere.

With this in mind, using the CENSe resource, emotions present in the 'All' corpora were identified.

7.4.1 CENSe-based tagging system description

Based on the CENSe dimensional model, Dr Tuomo Kakkonen and a Master degree student at the University of Eastern Finland named Ehsan Khakifirooz, developed a system that was able to annotate emotion in text and output an annotated file in XML format. The annotations were based on the EmotionML vocabulary structure described in Section 6.2. The annotation system was created using the General Architecture Text Engineering (GATE[2]) platform. The system has a user interface as shown in Figure 7.2. Data is entered by clicking the 'Open File' button and then specifying the file path of a document. Once inputted, the document's content is visible in display 1 (marked in Figure 7.2). Results of the analysis are then visible in display area 1 and 2 (shown in Figure 7.3).

[2]https://gate.ac.uk/

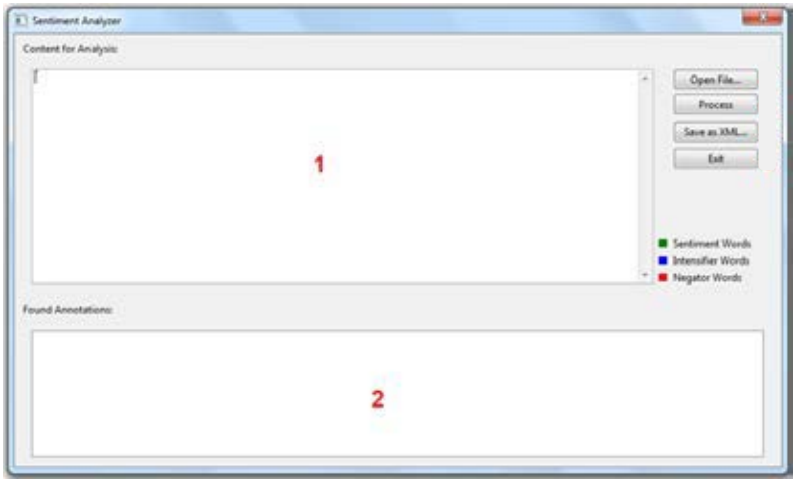

Figure 7.2: CENSe-based tagging system user interface.

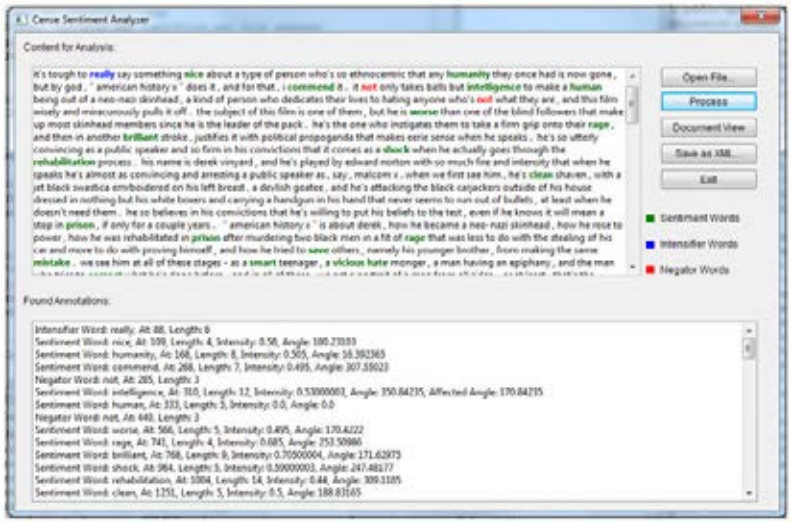

Figure 7.3: CENSe-based tagging system with results visible.

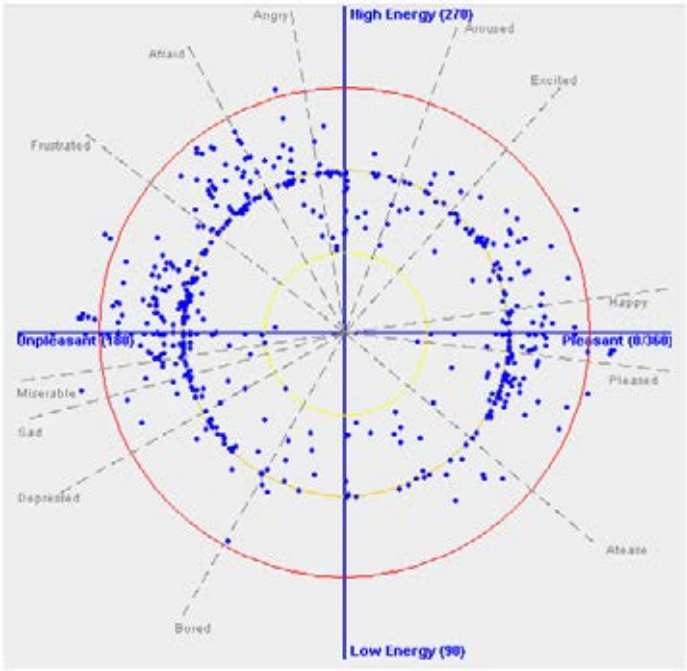

Figure 7.4: Distribution of emotions in the ASB corpus along the dimensional emotional space.

7.4.2 ASB detection with emotion features

Detection models developed with three emotion feature sets (i.e., emotion words; emotion quality and intensity; and emotion words, quality and intensity) are presented in this section along with their evaluation. Figure 7.4 shows the distribution of the identified emotions in ASBT and Figure 7.5 illustrates their distribution in percentages.

From Figure 7.5, it can be observed that ASBT contained more unpleasant than pleasant emotion-based words (76.07% vs. 23.93%). This finding is in line with the research expectations from Chapter 4's analysis of ASB, that is, ASB was highly associated with negative emotions. Moreover, of the unpleasant terms, 81.81% of them were identified to be unpleasant with high arousal and of those, 80.18% were with high intensity.

Table 7.12 presents all the feature sets used to develop the NB, SMO, and J48 classifiers, and the ten-fold cross validation results on the 'All' corpora.

From Table 7.12, it can be observed that with the emotion quality and intensity feature set, the J48 classifier performed the best with an accuracy level of about 89%; however, with the same feature set, recall and precision values were rather low especially for the NB classifier.

By combining the emotion quality and intensity feature set with the emotion words, the accuracy levels for all the three classifiers was observed to increase from 81.94% (NB), 86.11% (SMO), and 89.39% (J48) to 90.91%, 91.41%, and 91.16% respectively. Emotion words by themselves proved to be the best features, with the NB and SMO classifiers achieving accuracy levels of 92.42% and 92.80% respectively. In

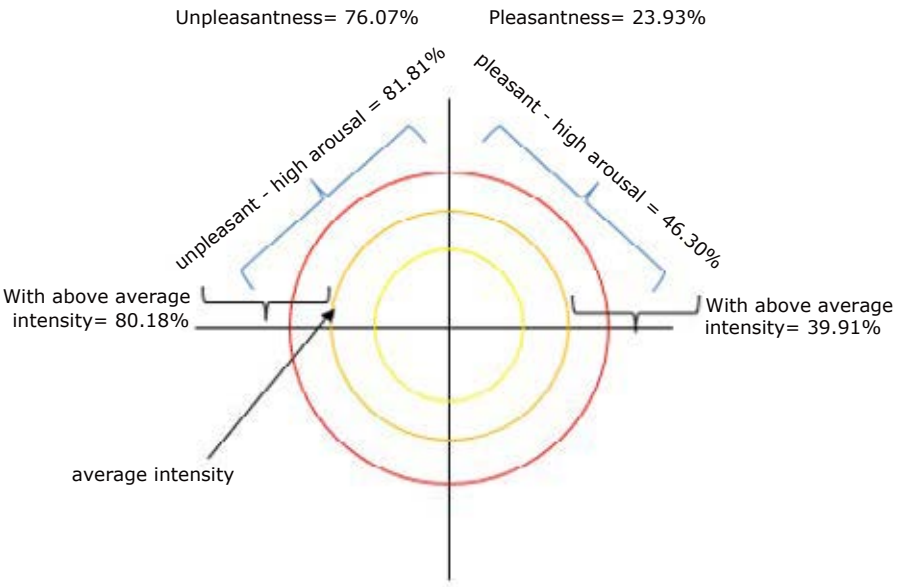

Figure 7.5: Distribution of emotions in the ASB corpus.

Table 7.12: Results for the ASB classification using the emotion features (%).

Features	Classifier	Accuracy	Precision	Recall	F-measure
Intensity + Quality	NB	81.94	52.59	47.33	49.83
	SMO	86.11	64.49	59.33	61.81
	J48	**89.39**	**78.45**	**60.67**	**68.42**
Emotion words	NB	92.42	82.61	76.00	79.17
	SMO	**92.80**	**83.45**	**77.33**	**80.28**
	J48	89.52	77.24	63.33	69.60
Emotion words + Intensity + Quality	NB	90.91	77.86	72.67	75.17
	SMO	**91.41**	79.71	**73.33**	**76.39**
	J48	91.16	**90.00**	60.00	72.00

addition, based on the results achieved with the three feature sets, it can be observed that all three classifiers performed better than the baseline performance (accuracy = 81.06%, precision = n/a, and recall = 0%). Table 7.13 compares the classification performance of the best classifier with each of the three feature sets presented in Table 7.12.

Based on the three emotion features, the SMO classifier with the emotion words feature set was able to better accurately classify ASBT as ASBT; however, the J48 classifier with the emotion quality and intensity feature set misclassified a higher number of ASBT as non-ASBT. Based on the results achieved with the three feature sets, it can be observed that the emotion quality and intensity feature set were not efficient features for ASB detection. In addition, using Information Gain as a feature selection method, possible reasons for the low accuracy, precision, and recall results were investigated. In looking at the emotion quality and intensity features that contributed most to the classification, it was found that up to 60% of the contributing

Table 7.13: Comparative table of TP, FN, FP, and TN for the top performing classifier with each of the three feature sets presented in Table 7.12.

Features	Classifier	True Pos.	False Neg.	False Pos.	True Neg.
Intensity + Quality	J48	91	59	25	617
Emotion words	SMO	116	34	23	619
Emotion words + Intensity + Quality	SMO	110	40	28	614

features belonged to the pleasant emotion quality, most likely due to the larger non-ASBT datasets, especially since earlier findings showed that the distribution of emotion features in the ASB corpus was 76% unpleasant versus 24% pleasant (see Figure 7.5). This gave some insight into why the classifiers performed poorly in classifying ASBT as ASB. When the emotion quality and intensity features were combined with emotion words, the accuracy of the classifiers was greatly improved. The emotion words proved to be better performing features. Using the Information Gain feature selection approach, it was identified that the top contributing words included swear words (e.g., fu*k, motherfu*ker, and shit), insulting words (e.g., ass, nigger, bitch, stupid, and scum), violence-related words (e.g., destroy, kill, violence, and burn), and negative emotion-bearing words (e.g., hate and angry), which were identified to be factors of ASB in Chapter 4.

In addition, experiments with the three emotion feature sets combined with unigrams and top 15 LIWC categories (identified in Section 7.3.3) were performed. Table 7.14 shows the ten-fold cross validation results for the NB, SMO and J48 classifiers on the 'All' corpora.

From the results in Table 7.14, it was identified that the SMO classifier achieved the best performance in several combinations of feature sets. Figure 7.6 visually illustrates the best accuracies with each feature set.

From Figure 7.6, it can be observed that not only do all the accuracy results outperform the baseline values, the feature sets combining emotion quality and intensity with unigrams and the top 15 LIWC categories resulted in the highest accuracy (96%). To consider the types of errors that the above classifiers made, the TP, FP, FN, and TN values for the best performing classifiers in Table 7.14 are presented in Table 7.15.

From Table 7.15, it can be observed that the SMO classifier with the emotion quality, intensity, unigrams and top 15 LIWC categories performed the best in correctly classifying ASBT as ASBT and non-ASBT as non-ASBT.

Table 7.14: Results for the ASB classification with a combination of emotion, unigrams, and 15 top LIWC categories features (%) (top 15 = the top 15 LIWC categories).

Features	Classifier	Accuracy	Precision	Recall	F-measure
Intensity + Quality					
+ unigrams	NB	92.93	**97.96**	64.00	77.42
	SMO	**95.83**	95.35	**82.00**	**88.17**
	J48	91.41	83.61	68.00	75.00
+ top 15	NB	91.79	81.48	73.33	77.19
	SMO	91.04	77.24	74.67	75.93
	J48	**93.43**	**84.03**	**80.67**	**82.31**
+ unigrams, top 15	NB	93.18	87.50	74.67	80.58
	SMO	**96.46**	**96.21**	**84.67**	**90.07**
	J48	94.95	90.44	82.00	86.01
Emotion words					
+ unigrams	NB	91.79	**97.75**	58.00	72.80
	SMO	**94.44**	94.17	**75.33**	**83.70**
	J48	91.41	85.35	66.00	74.47
+ top 15	NB	93.94	85.92	81.33	83.56
	SMO	93.94	84.46	**83.33**	83.89
	J48	**94.32**	**90.08**	78.67	**83.99**
+ unigrams, top 15	NB	93.31	89.43	73.33	80.59
	SMO	**95.96**	**96.83**	**81.33**	**88.41**
	J48	94.19	87.68	80.67	84.03
Emotion words + Intensity + Quality					
+ unigrams	NB	92.55	**97.90**	62.00	75.92
	SMO	**95.71**	94.62	**82.00**	**87.86**
	J48	91.29	87.85	62.67	73.15
+ top 15	NB	93.18	**84.29**	78.67	81.38
	SMO	91.79	80.14	75.33	77.66
	J48	**93.94**	83.55	**84.67**	**84.11**
+ unigrams, top 15	NB	92.30	77.64	**83.33**	80.39
	SMO	**95.46**	**94.53**	80.67	**87.05**
	J48	94.57	89.05	81.33	85.02

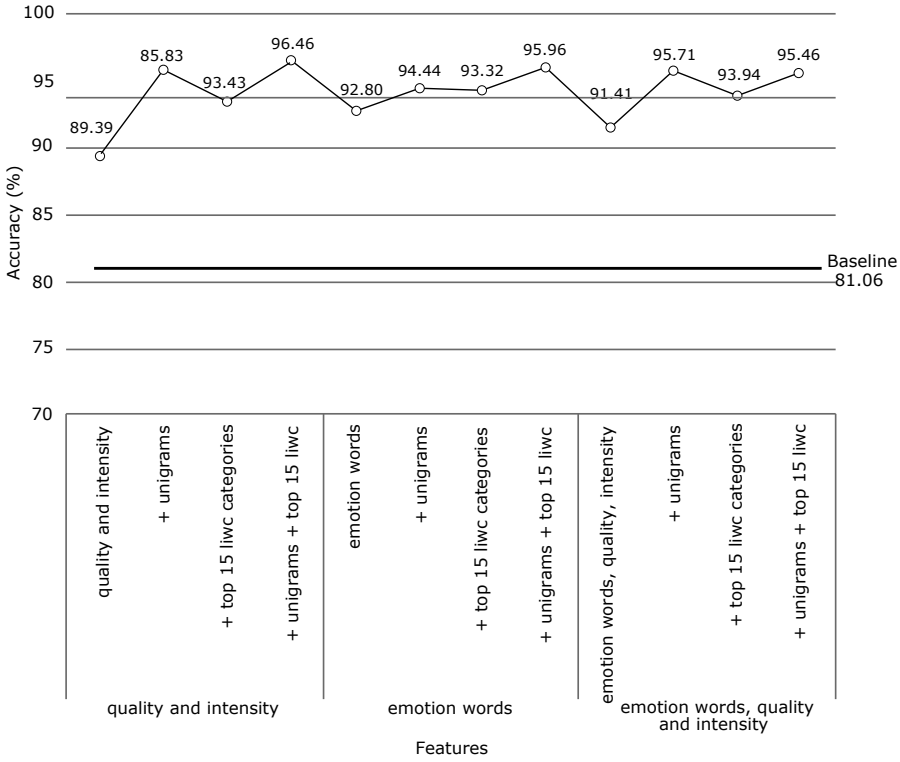

Figure 7.6: Effect of combining emotion features with unigrams and top 15 LIWC features for ASB detection (Baseline classifier performance also shown).

Table 7.15: Comparative table of TP, FN, FP, and TN for the top performing classifier for each of the three feature sets presented in Table 7.14 (top 15 = the top LIWC categories).

Features	Classifier	True Pos.	False Neg.	False Pos.	True Neg.
Intensity + Quality					
+unigrams	SMO	123	27	6	636
+ top 15	J48	121	29	23	619
+ unigrams, top 15	SMO	127	23	5	637
Emotion words					
+unigrams	SMO	113	37	7	635
+ top 15	J48	118	32	13	629
+ unigrams, top 15	SMO	122	28	4	638
Emotion words + Intensity + Quality					
+unigrams	SMO	123	27	7	635
+ top 15	J48	127	23	25	617
+ unigrams, top 15	SMO	123	27	7	635

8 INTERPRETATION AND DISCUSSION

Emotions cut to the core of people. Within and through emotion, people come to define the surface and essential, or core, meanings of who they are. – Denzin (2007)

Feature selection always plays a key role in constructing classifiers. In the previous chapter, an experimental framework for ASB detection was presented in which three sets of features were selected: lexico-syntactic; linguistic, psychological, and social; and emotion-based. The selected features allowed for examining the way ASB is expressed in written language. By leveraging the selected features and their combinations, various classification models were developed and evaluated to detect ASBT and to determine which features best distinguish ASBT from non-ASBT. In this chapter, the results of these experiments in terms of their implications for ASB detection are presented. In addition, the applicability of the results for practitioners and researchers as well as the possible limitations of the research are discussed.

8.1 FEATURE SETS PERFORMANCE

The proposed framework included three different feature sets for the detection of ASB. Figure 8.1 summarizes the classification results by showing the best accuracy rates along with the F-measure values among each of the three experiments, including their combinations. Figure 8.2 illustrates the true positive values for the same feature sets.

Based on Figure 8.1 and Figure 8.2, all framework feature sets performed well in terms of how accurately they could distinguish ASBT from non-ASBT, with accuracy results above 90% in all datasets. In particular, combining emotion quality, intensity, unigrams, and the top 15 LIWC categories resulted in the highest accuracy of over 96% with an F-measure of 90% and a true positive score of 127. The full LIWC features were second with accuracy results of about 93%, an F-measure of 86%, and a true positive score of 121. After specifically examining the true positive values, it was found that the classifier that used only emotion words achieved a higher true positive value than the classifier that used only unigrams, though the accuracy levels were higher for the unigram than when only emotion word features were used. This indicates that the classifiers using emotion words could detect more instances of ASBT than when using unigrams, while at the same time, more mistakes were made in correctly classifying non-ASBT as non-ASBT. This could be due to the finding that swearing, insults, and violence-related words contributed most to the classification, and thus their presence in, for instance, the Wikipedia collection, resulted in some misclassification. This is the limitation of the keyword-based approaches in detecting emotions, as the context of the words was not taken into consideration.

Moreover, though the unigram-based models achieved good results, it was found that the accuracy levels were increased by adding semantic information, such as the emotion quality and intensity values. Initially, it was hypothesized that emotions were good classification features due to the nature and connection identified between emotions and ASB (see Chapter 4). To test this hypothesis, a fine-grained

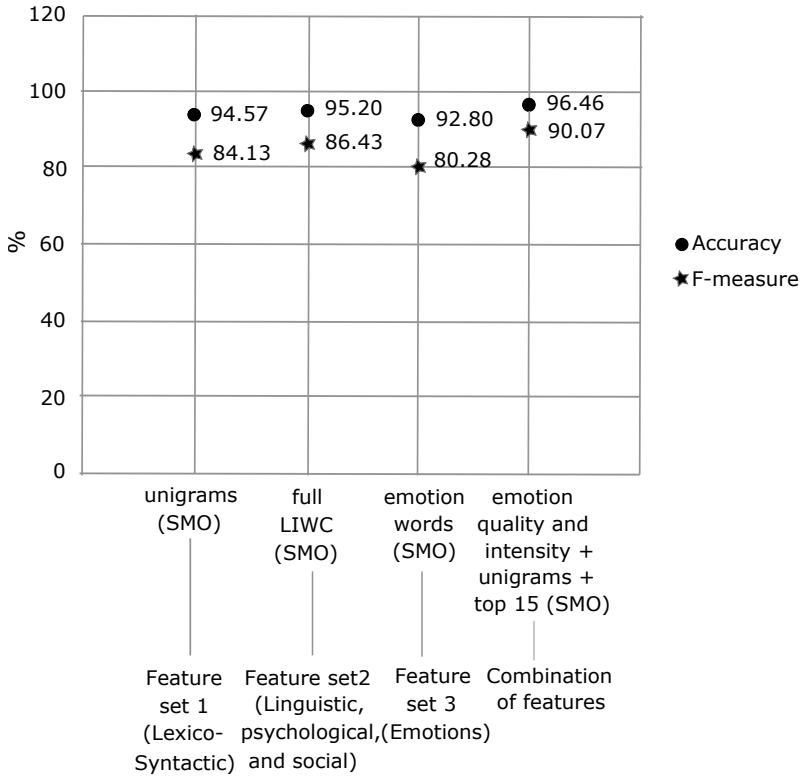

Figure 8.1: Comparison of the best classifiers in terms of accuracy and F-measure results among each of the three feature sets and their combinations.

resource was developed to identify emotions in text. For the experiments, the features provided by the CENSe resource were used: emotion words, intensity, and emotion quality values. First, the experiments showed that it is possible to use the resource to observe and analyze the distributions of emotions in the dimensions selected, i.e., pleasantness-unpleasantness, high-low arousal, and intensity (see Figure 7.4 and 7.5). Second, it was found that the resource provides good features for ASB detection. In addition, the experiments revealed that the best emotion features were the emotion words alone with accuracy levels of about 93%, and a combination of emotion words, intensity, and emotion quality achieved an accuracy rate of 91%.

Based on the results, each of the feature sets - lexico-syntactic; linguistic, psychological, and social features; and emotion-based - performed relatively well in detecting ASB independently, and thus it logically followed that combining the feature sets resulted in improved accuracy results of up to 96%.

8.2 CLASSIFIER PERFORMANCE

As can be observed in Figure 8.1 and Figure 8.2, it was also found that the best accuracy results for each feature set were achieved with the SMO classifier, indicating that SVM classifiers perform better with the targeted dataset. SVM classifiers were also found by Joachims [302] to be-well suited for text classification tasks, as they

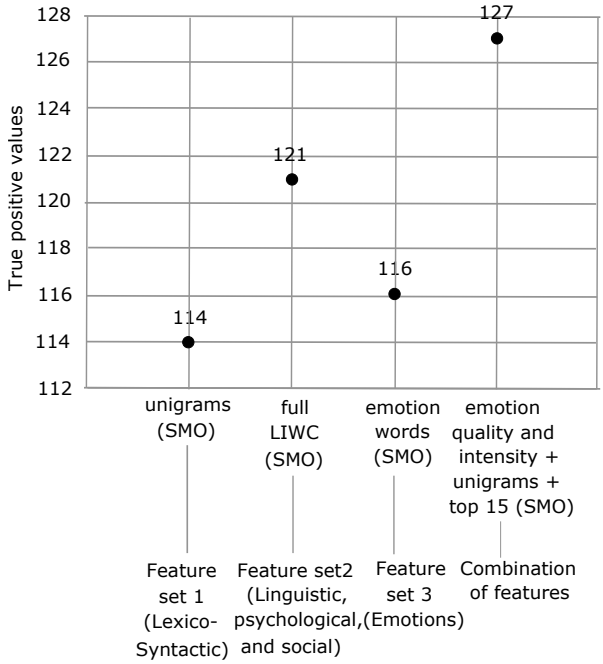

Figure 8.2: Comparison of the best classifiers in terms of number of true positives in the ASB dataset among each of the three feature sets and their combinations.

are better able to manage the properties of text, such as high dimensional features, vectors that are dense (i.e., where most of the features are relevant), and sparse instance vectors.

Based on the performance of the classifiers in all the experiments conducted, it was observed that all three classifiers achieved a considerably better performance than the baseline, particularly in terms of precision and recall. In addition, it was observed that an imbalanced corpora had an impact on the performance of some of the classifiers. As Rennie et al. [303] also observed, imbalanced data, where one class has more examples, can cause the decision boundary weights to be biased, which might cause the classifier to unintentionally prefer one class over the other, as was the case with the trivial baseline classifier. This was particularly noticeable with the experiments with bigrams using the J48 classifier as well as the experiments with the emotion quality and intensity features with all three classifiers, where the recall and precision values were rather low. Because the ASB corpus was the minority class in the dataset, the corpus was poorly represented by an excessively reduced number of examples, which was insufficient for the classifiers to learn from the examples for more accurate classification.

Regardless of this impact, it is important to highlight that in this work, models that can manage real-world situations were developed, where the majority of UGC was non-ASB; however, future work should involve expanding the ASB corpus and training the classifiers on the expanded corpus again.

Moreover, when measuring classifier performance, almost no single performance measure can capture every desirable property; however, for the detection of harmful behaviors, a balance between precision and recall is desired over accuracy results.

If Facebook and Twitter are used as example users of the developed ASB detection models, both are platforms that have a large amount of data, and both aim to combat hate speech and harmful behaviors. Using the classifiers, the two companies would be interested in detecting all hate speech or harmful behavior instances (high recall), and due to their interests in fairness and users' rights, they would want to reprimand or shut down the accounts of people who were truly expressing hate speech or harmful behavior (high precision). Thus, the SMO classifier, which had over a 90% F-measure with unigrams, emotion quality, and intensity, and LIWC features would be the most suitable classifier for this scenario as well as several related scenarios.

To extend the analysis of UGC for ASB, future research will focus on transforming the detection task into a ranking problem, where the output of a ranking algorithm would then be a graded relevance scale that determines the likelihood of UGC being ASB. This type of ranking would allow practitioners to focus on UGC that is ranked highest.

Furthermore, to ensure the reproducibility of the reported methods and to assess the actual effects of the feature sets, no parameter was modified from the classifier's default settings in WEKA. It is acknowledged that better results might have been possible by adjusting the parameters of each classifier; however, given the staggeringly large number of potential combinations of parameter values for each algorithm, the aim was to avoid a situation in which the tuned parameters would result in overfitting the data, especially considering the small size of the current ASB corpus. Nevertheless, additional research on the influence of tuning the classifiers' parameters for a larger ASB corpus is advisable.

As a preliminary test to determine whether parameter tuning would have had a significant influence on actual classifier performance, WEKA's automated process for identifying optimal parameters for a classifier was used. The CVParameterSelection meta classifier[1] was used, which automatically optimizes an arbitrary number of parameters of a base classifier. The base classifier used for testing was the best performing SMO classifier with unigrams, emotion quality, intensity, and LIWC features that had a 96.46% accuracy and a 90.07% F-measure. After running the classification with parameter tuning, the SMO classifier achieved a 96.34% accuracy and an 89.68% F-measure. This indicates a slightly lower performance than the classifier without parameter tuning. Nevertheless, further testing should be conducted to determine whether the classifiers can be optimized by tuning the parameters.

Regardless, the objectives of the experiments were achieved, which were to show the robustness of the methods in detecting ASB in text and to identify the ASB indicators discussed next.

8.3 ANTISOCIAL BEHAVIOR FEATURES

The aim of this dissertation was to investigate the manner in which ASB is represented or can be present in written language. Based on the results, it has been concluded that the way in which people write and the words used can efficiently be used to distinguish between ASBT and non-ASBT. Many of the factors that lead to ASB were detected in text, which is in line with the expectations from Chapter 4's analysis.

[1] http://weka.sourceforge.net/doc.stable/weka/classifiers/meta/CVParameterSelection.html

Firstly, aggression was identified in Section 4.2 as one of the most prominent aspects associated with ASB. Aggression was defined as a behavior carried out with the proximate intent to cause harm to another individual. In the experiments, based on the most distinguishing unigrams for ASBT, words were identified that would be considered harm-related, e.g., "beat," "kill," and "hit." In addition, future intent signaling verbs, such as "will" that occurred before the harm-related words were detected, which indicated that the ASB characteristic, i.e., intent to harm, can be present in written form. This is an important finding because it shows that detecting ASBT might allow for the opportunity to intervene before acts of harm take place. The finding also contributes to the objective of the study to develop an early detection system for ASB. Furthermore, in ASBT, there was a higher use of 1st-person and 2nd-person singular pronouns, which is indicative of a personal expression of harm towards another individual.

Moreover, based on the findings presented in Chapter 4, anger and frustration are negative emotions that are considered prominent instigators of aggressive behavior. The experiments discussed in Sections 7.3 and 7.4 showed that this assertion is correct. Using the use of the LIWC categories, it was concluded that in comparison with non-ASBT, ASBT included the use of more swear words and insults, which are expressions of anger and aggression [6,298]. ASBT also scored higher than non-ASBT in the negative emotions and anger LIWC categories. Moreover, based on the CENSe resource, it was also observed that ASBT contained a high number of unpleasant emotions, with about 81% being unpleasant and with high arousal. High arousal emotions were also identified to be associated with ASB and aggression [16]. In addition, it was observed that pleasant emotions were present in the ASB text as well. As discussed in Chapter 4, there are instances in which love, desire, and even excitement can be instigators of aggressive acts.

Thus, based on the analysis and comparison of ASBT and non-ASBT, the presence of distinguishing features of ASB that are in line with what constitutes ASB was identified (see Chapter 4).

8.4 LIMITATIONS OF THE RESEARCH

The research aimed at leveraging word usage and emotion-based features to detect ASB and obtain insights on how ASB is expressed in written language. Despite the evidence provided throughout the work that ASB texts have distinguishing features allowing for its accurate detection in text, it must be borne in mind that the classifiers described are proof-of-concepts and need to be tested in the real-world. The results of the research are just the beginning, with continued research on the relationships between natural language and public health concerns and with a multidisciplinary effort in building models to assess the probability of harmful behavior, much progress can be made.

In this section, limitations of the research are discussed along three main areas: ASB corpus, CENSe as an emotion detection resource, and text as a communication medium.

ASB corpus

It is acknowledged that the size of the ASBT corpus was relatively small in size. Ideally, ML techniques should be applied to corpora containing thousands of documents; however, the corpus consisted of texts that could confidently be identified

as ASB and this was needed in developing effective detection models. As future work, the developed models could be used to semi-automatically expand the ASB corpus by identifying potential ASBT. In addition, only three non-ASBT datasets for building the models were used, thus, it cannot be guaranteed that the results in the empirical studies can be generalized to other corpora or achieve similar results; however, the datasets used to build the models constitute texts that are commonly found online, i.e., formal, informal, and are UGC. This indicates that the results are relevant to real-world use cases.

CENSe as an emotion detection resource

The CENSe resource proved to be a useful resource in terms of accurately distinguishing between ASBT and non-ASBT. In particular, the emotion words by themselves achieved good accuracy rates, with the second best classification results being achieved by combining the emotion words with their intensity and emotion quality values. As with many lexicon-based resources, CENSe has two limitations; first, to expand the resource using an annotator agreement approach is expensive in terms of time and money. Using the 360 degrees of freedom for the annotation did increase the cognitive load of the annotators, so that only a small amount of annotations were possible in a day. It is understandable hence why previous research have focused on a smaller number of categories or range of emotions; however, having such a fine-grained representation reveals a greater range in differences in emotions, and allows for comparisons to be performed easily.

Second, the emotion words in the lexicon are still limited in number. For improved emotion analysis, the resource will need to include more annotated words. In addition, the effect that punctuation and capitalization might have had on the annotation values, specifically on intensity would need to be investigated. This aspect was however beyond the scope of the research. The effects of capitalization and punctuation on intensity is a possible future research avenue. This is an example of an improvement that could be implemented for the CENSe-based tagging system.

Overall, the resource and the tagging system need to be extended to allow for the adjustment of the angle and intensity values depending on context. In future, whether it is worth to incorporate other NLP techniques such as the use of normative databases such as commonsense databases in order to obtain the context will be investigated. In addition, during the annotation, a question arose on whether there should be a difference in how emotions for different verb tenses should be annotated. For example, 'like' and 'liked' might communicate different emotion qualities or intensity levels. Such questions and uncertainty give light to the limitations still remaining in the research of accurately detecting emotions in written language.

Text as a communication medium

The research focused only on what is written, it did not go into the analysis of factors that may affect and or influence an individual's choice of behavior. These factors might include individual characteristics, personality traits, environment, and cultural background, which were not explored in the research. Analyzing such factors might be beneficial for understanding the author of the ASBT, but as observed during the experiments, when analyzing text, the above information may not necessarily be readily available in text. Thus, confirming that the use of text only can only give a probabilistic prediction and that no absolute predictions can be made.

Moreover, the research focused on what was expressed by the authors of the texts; however, text can also be analyzed from the perspective of the reader, including analyzing the emotions evoked from reading a piece of text. In future, the ability to discriminate between these two perspectives will be investigated. Specifically in the area of harmful behavior detection, it might be useful to know whether a reader feels threatened or harmed by reading certain ASB texts.

8.5 APPLICATIONS AND IMPLICATIONS OF THE RESEARCH

The work presented here translates to practical implications for ASB detection and is a step towards ASB prevention in society. The study and its results have great potential for various applications and in addition pose a number of implications for the use of NLP methods for harmful behavior detection. The overarching aim of the study was to provide a new way to identify ASB to a number of stakeholders such as researchers, policy makers, educators, and even parents.

In particular, the main applications of the developed ASB detection models and the CENSe resource in envisioned to belonging to two groups: e-counseling/e-health and security, which are discussed in detail in the next subsections.

8.5.1 Security

In line with the research motivation, the research results are beneficial for real world security deployment. The developed ASB detection models can be augmented into other systems. For instance, the models can be augmented into social networks or school networks so as to monitor and predict incidents of ASB and or act as early warning systems. More specifically:

- Law enforcement officers and forum administrators can use the resulting ASB detection models to identify conflicts and problematic users on the web. Often the amount of text produced in these mediums is too much to read through manually.
- For schools particularly, the models can be narrowed to specifically detect harmful behaviors particular to schools, e.g., school shootings and bullying.
- Moreover, the models could be used on social networking sites like Facebook and Twitter to automatically prevent harmful messages from being posted.

8.5.2 e-Counseling

The analysis of emotions and ASB has several clinical implications. Possible avenues include e-counseling and self-reflection systems. In e-counseling, psychologists, for instance can make use of CENSe-based tagging system to analyze patients' diaries, letters, etc. The system can assist in identifying sudden and persistent emotions, which in turn can help in diagnosing a patient's progress. In particular, the automatic detection of emotions and understanding of text can further help physicians identify those patients that might for instance be suffering from depression.

Not only practitioners, but any individual can make use of text analysis methods for self-reflection purposes. As Pennebaker and Graybeal [304] identified, when people talk or write about emotional and personal issues, they are able to achieve greater understanding of themselves. This is because language is a medium by which people come to alter or inspect their self-perceptions.

9 CONCLUSION AND FUTURE WORK

Language is rather important. It is the basis of most human communication and is the filter through which we understand and learn about ourselves and others - whether as researchers, clinicians, or human beings. – Pennebaker (2002)

This work has addressed a growing concern for society, i.e., ASB, and has produced methods and approaches to help pertinent authorities address this concern. The emphasis of the research was the way word usage, writing styles, and emotions can reveal features of ASB in text. By analyzing written materials pertaining to ASB, the principle feasibility of automatically detecting ASB in texts has been demonstrated. In this final chapter, the research findings will be discussed and summarized in response to the research questions posed at the beginning of this dissertation. In addition, the research contributions will be reviewed, and directions for future research will be discussed.

RQ1: How can emotions in text be identified and automatically analyzed effectively?

This question was answered by extensively analyzing the current literature on emotions and language (see Chapters 2, 3, 4, and 5). Emotions were identified as a complex phenomenon to define and thus difficult to measure. Currently, there is still a lack of a standard definition of emotions or standard approach to identify emotions in text. Therefore, it was determined that the best way to define emotions is to use a description of emotions. Hence, in Chapter 2, a working description of emotions was outlined as an integration of multiple components that are agreed upon by researchers. The description states that emotions are composed of the following components: appraisal, subjective feeling, physiological arousal, expressive behavior, and action tendencies.

Based on the emotion components, the ways this description can be modeled and captured in text was illustrated (Chapters 2 and 3). It was determined that the most beneficial way to capture the emotion components in text was to represent emotions along a three-dimensional model: unpleasant-pleasant, high-low arousal, and intensity. The representation follows the theory that emotions are better represented in a dimensional model. This is substantially different from the categorical representation most commonly followed in the analysis of emotion.

In addition, after reviewing the different existing approaches used for automatically analyzing emotions in text, it was established that the use of lexicons is still a beneficial approach for an emotion detection task. Based on these findings, a resource was developed for detecting emotions in text along the selected dimensional representation of emotions. The resource proved to be efficient in the detection of ASB in text (see Section 7.4).

Nevertheless, based on the literature analysis of emotions, determining which approach best represents and automatically identifies emotions in text is still debatable and still dependent on the research purpose. Unfortunately, many of the

current approaches are still limited in their capabilities of capturing emotions in written language.

RQ2: Which features are most beneficial for the detection of ASB in text?

Selecting features is an important step in developing automated methods for ASB detection. An ASB detection framework was presented in which the use of lexico-syntactic; linguistic, psychological, and sociological; and emotion-based features and their combinations were outlined to develop ASB detection models. In Chapters 7 and 8, experiments with the feature sets were described, and the results were discussed. In Chapter 8, it was shown that each of the feature sets performed relatively well on their own as ASB detection models. In particular, a combination of unigrams, LIWC, and emotion features was most effective in detecting ASB with an accuracy of up to 96%. Further evaluation of the detection model revealed that the features had a high precision, indicating that the texts classified as ASB were actually ASB. The developed automated models provide the opportunity for integrating them with analysis tools, allowing for fast and reliable warnings and interventions.

RQ3: How can research on the language and emotion features of ASB further improve the understanding of ASB?

A primary aim of this study was to develop automated methods for detecting ASB and to explore and analyze the ways ASB is expressed in written language.

The results of RQ1 and RQ2 greatly informed RQ3. In Chapter 7, substantial evidence that ASBT exhibits common features was presented. For instance, it was observed that ASBT use a higher amount of certain linguistic features:

1. ASBT have a high presence of swear words and profanity, which were specifically identified by Tausczik and Pennebaker [6] and Montagu [298] as being correlated with the use of informal language and expressions of aggression and anger. In addition, a high presence of insults that make use of body and sexual references was identified, which are also expressions of anger. This supports the findings discussed in Chapter 4 that aggression and anger are the primary characteristics of ASB.

2. In Chapter 4, it was established that aggression and negative emotions, such as anger, frustration, and shame, in addition to both high and low arousal may cause an individual to engage in aggressive behaviors leading to ASB. Using the CENSe resource, it was observed that ASBT consist of both high and low unpleasant emotion words, with the majority being unpleasant and high arousal words (see Figure 7.4 and Figure 7.5 in Section 7.4).

3. A third characteristic of ASB identified in this research was that ASBT include more future tense verbs, which indicates that the texts refer to the future and are goal-oriented [6]. Thus, it is possible to intervene before acts of violence or harm take place, allowing for early intervention.

4. Lastly, it was discerned that ASBT use more 2nd-person singular pronouns, which indicates that the attention or focus of the aggression or negative emotions is towards a person other than the author. Further analysis in this area will be conducted, as identifying the other people referred to by the pronouns might help identify the target(s) of harm.

Overall, this research has contributed to obtaining a deeper understanding of the manner in which ASB is expressed in written language. The results will aid in informing prevention and intervention approaches.

9.1 THESIS CONTRIBUTION

With the increasing amount of text generated and available online, computational methods have become a necessity, as it is impossible to manually monitor such an enormous amount of content. This dissertation has presented a solution to automatically detect ASB in UGC. It has been shown that it is possible to automatically detect harmful behaviors in written language. A key contribution has been made in the area of NLP and public health safety by extensively reviewing the existing research and providing a solution for the detection of ASB. To the author's knowledge, this is the first exploration of the detection of ASB and its characteristics in text. The results obtained in this research revealed that computational approaches represent a viable solution for the task of ASB detection. The results achieved the overall objective of the research, which has resulted in the following major outcomes:

- One of the difficulties of examining ASB communication and actual behaviors is that researchers typically do not have access to a collection of documents pertaining to ASB. For this reason, an ASBT corpus containing 150 documents that were reliably judged as being ASB was created. The resultant ASB corpus will be available to researchers and practitioners.

- As Egan [14] stated, ASB perpetrators, such as school shooters, have often given sufficient warning signs in detailed texts that described dramatic and violent outbursts. Hence, one of the more significant contributions of this research is the development of automated ASB detection models that can be used to detect ASB, hopefully before acts of violence occur. The research has produced detection models for ASB that achieve over a 90% accuracy rate for the collected datasets and that have also shown high precision and recall rates. The models can be integrated with other systems to alert and warn pertinent authorities, allowing for fast and reliable warnings and interventions.

- The ability to identify emotions pertaining to ASB was an important aspect of this research because as mentioned in Chapter 4, emotions do play a role in behavior, whether direct or indirect. In particular, it was found that when emotion does play a more direct role in priming behavior, it results in less than optimal behaviors, such as ASB. Hence, to capture and visualize the emotions in a psychologically meaningful way, a novel psychological-inspired resource was formulated that represents emotions in text along a three-dimensional space, called CENSe. Moreover, the resource is represented in a format (i.e., emotionML) that allows it to be easily integrated with other systems. In addition, the CENSe features proved to be efficient for the detection of ASB in text, achieving results comparable and even higher than those achieved by related research. In the future, the resource will be expanded and tested using additional UGC.

- Finally, this study has provided a deeper understanding of the content in ASB texts. It has been shown that ASBT contain a higher amount of swear words, profanity, insults, and violence-related words. ASBT contain both negative

and positive emotions, though they contain more negative emotions, including anger. In addition, the high use of 2nd-person singular pronouns indicates that the target of ASB is someone other than the author. Further analysis could reveal the identity of targets. An improved understanding of the characteristics of ASB will also lead to improved intervention measures.

Overall, this study has demonstrated the potential of using NLP techniques in ASB detection, thus providing a foundation for identifying solutions for public health safety. This work is an initial attempt to identify solutions, and with continued research on the relationships between natural language use and societal concerns and a multidisciplinary effort in developing models to assess the probability of harmful behavior, considerable progress can be made.

9.2 DIRECTIONS FOR FUTURE WORK

The research has made substantial contributions to providing automated solutions for detecting potential harmful behaviors in the society. Furthermore, it has raised several interesting theoretical and practical questions in the area of emotion detection and the application of NLP technologies for the detection of harmful behaviors; however, due to the scope of the research, a few of these questions were not addressed, which make them good avenues for future studies. In this section, the avenues are presented as recommendations and opportunities for future research. The avenues can be grouped into three broad categories: A common framework for emotion detection in text, ASB prevention, and ethical and privacy concerns of this kind of work.

9.2.1 A common framework for detecting emotions in text

As a result of the analysis of the different emotion models and representations, a representation of emotions that was suited for the research purpose was selected; however, it would be more beneficial for computational purposes if there was a unification of theories, which would lead to a common framework for analyzing emotions in text. A common framework would allow for the advancement of the NLP field and, in addition, allow comparisons to be made among different researcher or practitioner techniques for detecting emotions.

The common framework would additionally take into account the representation of other emotion-related terms (e.g., affect, sentiments, moods, etc.) that can be expressed in text. These emotion-related terms have their own definitions and some of them overlap, which can create confusion in the NLP community. During research on the different terms, the author found that rather than aim to have a model to represent each of the terms, it might be more convenient to have one common practical framework, based on a sound theoretical ground, which would allow other alternatives to be left out (see [305]).

One approach to resolve this, would be to start by fleshing out attributes that define certain dimensions relevant to making distinctions between all the emotion-related terms. For example, the terms could be represented and differentiated on dimensions such as quality, stability, intensity, level of abstractness, controllable/uncontrollable, triggered from within/outside, idiosyncracy vs. socially or culturally shared, real/feigned, explicitness, etc. Such a dimensional framework would then allow for the placement of a word or phrase on a specific spot in the

multi-dimensional plane. By incorporating all these dimensions, a common formalization of emotions and its related terms could be developed for the computing community.

A common framework that would capture all the emotion-related terms would prevent researchers from getting stuck in the semantics of what each term means and how they are differentiable. Furthermore, the common framework would make research in emotion more comparable which is currently difficult to perform since there exist several models and representations of emotions.

9.2.2 Antisocial behavior prediction

Predicting behavior is a task that is permeated with uncertainty. This is because often the same situation and emotion can induce a variety of different behaviors in people, depending on variables that are not always identifiable from text (e.g., goals, preferences, personality, social, and cultural factors).

The research focused on the detection of ASB based on the content of texts; however, this is an initial step. In order to truly have an effective early warning system of ASB, other criteria need to be included. The criteria that can be considered are for instance; modality of language or being able to distinguish between events or situations which have been asserted to have happened, are happening, or will happen, and those which might, could, should, ought to, or possibly have occurred or will occur [21]. Making such distinctions might lead to different intervention plans for ASB. Approaches for intervening for events that are happening, have happened, or will happen, would be different.

Another criterion is calculating and identifying the severity score of each ASBT, i.e., identifying whether a behavior is imminent or non-imminent, or the threshold at which an ASB detection system should warn the concerned authorities. All the above are very good questions and the answers would be an improved step in creating effective early warning ASB systems.

In addition, ASB detection has direct application in forensic linguistics and ASB prevention. Hence, a logical next step for future research is to include the presence of other factors such as location information (as shown possible in [306]), for instance, location of the author or location of where the potential harm might occur. This would allow for directing alert notifications to specific authorities or institutions. The significance of other factors such as personality (some people are more likely to react aggressively than others), environmental traits, history of similar behaviors, etc., to developing detection models and to better inform responses to ASB will also be investigated. The inclusion of the above information presents interesting opportunities for further research on profiling ASB perpetrators and tailoring ASB prevention strategies. Collecting that data will however not be an easy task, and it might not be readily available. Moreover, including historical information and events requires constant monitoring, which brings up the next proposed area for future research.

9.2.3 Real-world application opportunities and concerns: Privacy and ethical considerations

The detection models built in this research are planned to be incorporated into real-world systems, which will act as early warning systems for organizations and institutions such as schools and law enforcement agencies. Although the current

work did not infringe on ethical or privacy rights, care still needs to be taken in the future that this does not take place. This research respected the privacy of the authors' whose texts were used and it did not collect nor utilize any personal information.

Unfortunately, there has been concern over the access and analysis of personal texts by practitioners and researchers, with concerns that it might just be a step away from 'Big Brother.' Ethical and privacy issues do arouse real concerns that have an impact on the broad areas of NLP, ASB, and emotion detection. It is true, a large amount of data is open, for instance, the Twitter data is mostly public. However, it might not be the wish of a Twitter user to have their data analyzed for purposes that they have not given their consent to. This was observed when it was reported in June 2014[1] that Facebook was using status updates to manipulate the users' moods and observe how that manipulation translated to their status updates. Even though with Facebook one does agree to their data policy when using their application, people however are often not aware of what is being done with their data and it is hard to draw the line of whether it is alright or if it is infringing on privacy rights. If data is public, should researchers still ask for consent to use the data for research purposes, if yes, to whom should the consent be acquired from?

With so much available data, it is the task of researchers to make sure that the data acquired is only used for good. But how do we measure what is good and what might be considered invasive? It might be argued that if data and technology are used to prevent crimes or improve quality of life, then it should be allowed – but it should be done without abusing individuals' personal rights.

To take advice from Ray Kurzweil (in [307]), perhaps the answer to these ethical and privacy concerns would be to have a set of standards that are established through a whole social discussion between technologists and society.

[1] http://www.forbes.com/sites/kashmirhill/2014/06/28/facebook-manipulated-689003-users-emotions-for-science/

BIBLIOGRAPHY

[1] J. Cleland, "Racism, football fans, and online message boards: How social media has added a new dimension to racist discourse in English football," *Journal of Sport and Social Issues* **38**, 415–431 (2014).

[2] R.-L. Punamäki, K. Tirri, P. Nokelainen, and M. Marttunen, "Koulusurmat. yhteiskunnalliset ja psykologiset taustat ja ehkäisy," (2011), Helsinki: Suomalaisen Tiedeakatemian Kannanottoja.

[3] R. Card and R. Ward, *The Crime and Disorder Act 1998* (Jordan Publishing, Bristol, 1998).

[4] B. J. Bushman and C. A. Anderson, "Is it time to pull the plug on hostile versus instrumental aggression dichotomy?," *Psychological Review* **108**, 273–279 (2001).

[5] D. Biber, *University Language: A Corpus-based Study of Spoken and Written Registers*, Vol. 23, (John Benjamins Publishing, Amsterdam, 2006).

[6] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology* **29**, 24–54 (2010).

[7] R. Borum, D. G. Cornell, W. Modzeleski, and S. R. Jimerson, "What can be done about school shootings? A review of the evidence," *Educational Researcher* **39**, 27–37 (2010).

[8] P. J. Piotrowska, C. B. Stride, and R. Rowe, "Social gradients in child and adolescent antisocial behavior: a systematic review protocol," *Systematic Reviews* **1**, 38 (2012).

[9] B. W. Dalton, "Antisocial and prosocial behavior," *Noncognitive Skills in the Classroom: New Perspectives on Educational Research* 145–168 (2010).

[10] R. Armitage, "Tackling anti-social behaviour: what really works," in *Nacro Briefing Note* (2002), Nacro Crime and Social Policy Section.

[11] R. F. Baumeister and J. Lobbestael, "Emotions and antisocial behavior," *Journal of Forensic Psychiatry & Psychology* **22**, 635–649 (2011).

[12] R. Worth, *Children, Violence, and Murder* (Chelsea House Publishers, Philadelphia, 2001).

[13] B. Dedman, "Deadly lessons: School shooters tell why," (10/15/2000), Chicago Sun-Times, `https://archive.org/details/ERIC_ED448359` (visited on 2017-03-10).

[14] T. Egan, "Where rampages begin: A special report; From adolescent angst to shooting up schools," (06/14/1998), New York Times, http://www.nytimes.com/1998/06/14/us/where-rampages-begin-special-report-adolescent-angst-shooting-up-schools.html?pagewanted=1 (visited on 2017-03-10).

[15] T. Kiilakoski and A. Oksanen, "Cultural and peer influences on homicidal violence: A Finnish perspective," *New Directions for Student Leadership* **2011**, 31–42 (2011).

[16] C. A. Anderson and R. L. Huesmann, "Human aggression: A social-cognitive view," in *Handbook of Social Psychology*, M. Hogg and J. Cooper, eds. (Sage Publications, London, 2003), pp. 296–323.

[17] C. Hanrahan, "Antisocial behavior," (2006), The Gale Encyclopedia of Children's Health: Infancy through Adolescence, Encyclopedia.com, http://www.encyclopedia.com/doc/1G2-3447200056.html (visited on 2017-03-11).

[18] J. Walsh, "Rangaistavan vihapuheen levittäminen Internetissä: Rangaistavan vihapuheen määrittäminen ja rikosoikeudellisen vastuun kohdentuminen erilaisiin Internetissä toimiviin toimijoihin," (2012), Työryhmän raportti, Helsinki.

[19] B. Comrie, *Language Universals and Linguistic Typology: Syntax and Morphology* (University of Chicago press, Chicago, 1989).

[20] I. M. Marks, *Fears, Phobias, and Rituals: Panic, Anxiety, and their Disorders* (Oxford University Press, New York, NY, 1987).

[21] L. Polanyi and A. Zaenen, "Contextual valence shifters," in *Computing Attitude and Affect in Text: Theory and Applications*, J. Shanahan, Q. Y, and W. J, eds. (Springer Netherlands, Dordrecht, 2006), pp. 1–10.

[22] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing* **1**, 18–37 (2010).

[23] K. T. Strongman, *The Psychology of Emotion* (Wiley Press, New York, NY, 1978).

[24] A. Ben-Ze'ev, *The Subtlety of Emotions* (MIT Press, Cambridge, Massachusetts, 2000).

[25] R. S. Lazarus, *Emotion and Adaptation* (Oxford University Press, New York, 1991).

[26] J. Hillman, *Emotion: A Comprehensive Phenomenology of Theories and their Meaning for Therapy* (Northwestern University Press, Evanston, Illinois, 1960).

[27] J. Drever, *A Dictionary of Psychology* (Penguim Books, Inc., Oxford, 1952).

[28] C. A. Smith and R. S. Lazarus, "Emotion and adaptation," in *Handbook of Personality: Theory and Research*, L. A. Pervin, ed. (Guilford Press, New York, NY, 1990), pp. 609–637.

[29] C. E. Izard, *The Psychology of Emotions* (Plenum Press, New York, NY, 1991).

[30] M. Gendron, "Defining emotion: A brief history," *Emotion Review* **2**, 371–372 (2010).

[31] K. R. Scherer, "Psychological models of emotion," *The Neuropsychology of Emotion* **137**, 137–162 (2000).

[32] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information* **44**, 695–729 (2005).

[33] R. J. Dolan, "Emotion, cognition, and behavior," *Science* **298**, 1191–1194 (2002).

[34] W. James, *The Principles of psychology* (Dover, New York, NY, 1890).

[35] R. West and L. H. Turner, *Understanding Interpersonal Communication: Making Choices in Changing Times* (Cengage Learning, Boston, MA, 2010).

[36] L. S. Greenberg and J. D. Safran, "Emotion in psychotherapy," *American Psychologist* **44**, 19–29 (1989).

[37] P. R. Kleinginna and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and Emotion* **5**, 345–379 (1981).

[38] M. W. Eysenck and M. T. Keane, *Cognitive Psychology: A Student's Handbook* (Taylor & Francis, Hove, 2000).

[39] N. H. Frijda, *The Emotions* (Cambridge University Press, New York, NY, 1986).

[40] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," in *Appraisal Processes in Emotion: Theory, Methods, Research*, K. R. Scherer, A. Schorr, and T. Johnstone, eds. (Oxford University Press, New York, NY, 2001), pp. 92–120.

[41] N. H. Frijda, S. Markam, K. Sato, and R. Wiers, "Emotions and emotion words," in *Everyday Conceptions of Emotion: An Introduction to the Psychology, Anthropology and Linguistics of Emotion*, J. A. Russell, J. Fernández-Dolls, A. S. Manstead, and J. C. Wellenkamp, eds. (Springer Netherlands, Dordrecht, 1995), pp. 121–143.

[42] N. H. Frijda, "The laws of emotion," *American Psychologist* **43**, 349–358 (1988).

[43] E. B. Roesch, J. Fontaine, and K. R. Scherer, "The world of emotion is two-dimensional–or is it," (2006), Presentation at the HUMAINE Summer school, Genoa, Italy.

[44] C. Darwin, *The Expression of the Emotions in Man and Animals*, Vol. 526, (University of Chicago press, Chicago, US, 1965).

[45] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions from Facial Cues* (Prentice Hall, Englewood Cliffs, NJ, 1975).

[46] M. Schröder, *Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis* (Saarbrücken: Institut für Phonetik, Universität des Saarlandes, 2004).

[47] R. R. Cornelius, "Theoretical approaches to emotion," in *Proceedings of the ISCA Workshop on Speech and Emotion* (ISCA, 2000), pp. 3–10.

[48] W. James, "What is an emotion?," *Mind* **9**, 188–205 (1884).

[49] C. G. Lange, "The mechanism of the emotions," in *The Emotions*, D. Dunlap, ed. (Williams & Wilkins, Baltimore, MD, USA, 1885), pp. 33–92.

[50] J. Walsh, "Theories of emotion," (12/15/2013), Khan Academy, `https://www.khanacademy.org/video/theories-of-emotion` (visited on 2017-03-12).

[51] A. R. Damasio, *Descartes' Error: Emotion, Rationality and the Human Brain* (GP Putnam'Sons, New York, NY, 1994).

[52] C. Antoine, P. Antoine, P. Guermonprez, and B. Frigard, "Awareness of deficits and anosognosia in Alzheimer's disease," *L'Encephale* **30**, 570–577 (2003).

[53] W. B. Cannon, "The James-Lange theory of emotions: A critical examination and an alternative theory," *The American Journal of Psychology* **39**, 106–124 (1927).

[54] W. Cannon, *Bodily Changes in Pain, Hunger, Fear and Rage: An Account of Recent Researches Into the Function of Emotional Excitement* (Appleton-Century, New York, NY, 1929).

[55] P. Bard, "On emotional expression after decortication with some remarks on certain theoretical views: Part I," *Psychological Review* **41**, 309–329 (1934).

[56] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state," *Psychological Review* **69**, 379–399 (1962).

[57] M. B. Arnold, *Emotion and Personality* (Columbia University Press, New York, NY, 1960).

[58] K. Oatley and P. N. Johnson-Laird, "Towards a cognitive theory of emotions," *Cognition and Emotion* **1**, 29–50 (1987).

[59] K. R. Scherer, "On the nature and function of emotion: A component process approach," in *Approaches to Emotion*, Vol. 2293, K. R. Scherer and P. Ekman, eds. (Erlbaum, Hillsdale, NJ, 1984), pp. 293–317.

[60] P. C. Ellsworth and K. R. Scherer, "Appraisal processes in emotion," in *Handbook of Affective Sciences*, K. Scherer and H. Goldsmith, eds. (Erlbaum, Mahwah, NJ, 2003), pp. 572–595.

[61] T. Dalgleish and M. Power, *Handbook of Cognition and Emotion* (Wiley Online Library, Chichester, UK, 2000).

[62] J. Averill, "A constructivist view of emotion," in *In Emotion: Theory, Research and Experience*, R. Plutchik and H. Kellerman, eds. (Academic Press, New York: NY, 1980), pp. 305–339.

[63] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face: Guidelines for Research and an Integration of Findings* (Pergamon Press Inc., Elmsford, NY, 1972).

[64] R. R. Cornelius, *The Science of Emotion: Research and Tradition in the Psychology of Emotions* (Prentice-Hall, Inc., Englewood Cliffs, NJ, 1996).

[65] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology* **39**, 1161–1178 (1980).

[66] R. Cowie, H. Gunes, G. McKeown, L. Vaclau-Schneider, J. Armstrong, and E. Douglas-Cowie, "The emotional and communicative significance of head nods and shakes in a naturalistic database," in *Proc. of LREC Int. Workshop on Emotion* (2010), pp. 42–46.

[67] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (ISCA, 2000), pp. 19–24.

[68] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine* **18**, 32–80 (2001).

[69] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition* **17**, 484–495 (2008).

[70] G. Chanel, K. Ansari-Asl, and T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm," in *IEEE International Conference on Systems, Man and Cybernetics, 2007. ISIC.* (IEEE, 2007), pp. 2662–2667.

[71] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *Affective Dialogue Systems. ADS 2004. LNCS*, Vol. 3068, E. André, L. Dybkjær, W. Minker, and P. Heisterkamp, eds. Berlin, Heidelberg, 2004), pp. 36–48.

[72] T. Pun, T. I. Alecu, G. Chanel, J. Kronegg, and S. Voloshynovskiy, "Brain-computer interaction research at the Computer Vision and Multimedia Laboratory, University of Geneva," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **14**, 210–213 (2006).

[73] J. Reilly and L. Seibert, "Language and emotion," in *Handbook of Affective Sciences*, R. Davidson, K. Scherer, and H. Goldsmith, eds. (Oxford University Press, New York, NY, 2003), pp. 535–559.

[74] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE* **91**, 1370–1390 (2003).

[75] N. Sebe, I. Cohen, T. S. Huang, et al., "Multimodal emotion recognition," *Handbook of Pattern Recognition and Computer Vision* **4**, 387–419 (2005).

[76] P. Ekman, "Facial expression and emotion," *American Psychologist* **48**, 384–392 (1993).

[77] P. Ekman, "Expression and the nature of emotion," in *Approaches to Emotion*, Vol. 3, K. Scherer and P. Ekman, eds. (Lawrence Erlbaum, Hillsdale, NJ, 1984), pp. 19–344.

[78] D. Kaplan, "The meaning of ouch and oops: Explorations in the theory of Meaning as Use," (1999), Draft 3, ms., UCLA.

[79] J. E. LeDoux, "Emotion, memory and the brain," *Scientific American* **270**, 50–57 (1994).

[80] K. R. Scherer, "Appraisal theory," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, eds. (John Wiley & Sons Ltd., New York, NY, 1999), pp. 637–663.

[81] C. A. Smith, B. David, and L. D. Kirby, "Emotion-eliciting appraisals of social situations," in *Affect in Social Thinking and Behavior*, J. Forgas, ed. (Psychology Press, New York, NY, 2006), pp. 85–101.

[82] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and Emotion* **23**, 209–237 (2009).

[83] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions* **1**, 68–99 (2010).

[84] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication* **40**, 5–32 (2003).

[85] C. M. Whissell, "The dictionary of affect in language," in *Emotion: Theory, Research, and Experience*, R. Plutchik and H. Kellerman, eds. (Academic Press, New York, NY, 1989), pp. 113–131.

[86] J. R. Averill, *A Semantic Atlas of Emotional Concepts* (American Psycholog. Ass., Journal Suppl. Abstract Service, Washington, D.C, 1975).

[87] W. G. Parrott, *Emotions in Social Psychology: Essential Readings* (Psychology Press, Philadelphia, 2001).

[88] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: further exploration of a prototype approach," *Journal of Personality and Social Psychology* **52**, 1061–1086 (1987).

[89] B. Fehr and J. A. Russell, "Concept of emotion viewed from a prototype perspective," *Journal of Experimental Psychology: General* **113**, 464–486 (1984).

[90] C. Storm and T. Storm, "A taxonomic study of the vocabulary of emotions," *Journal of Personality and Social Psychology* **53**, 805–816 (1987).

[91] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz, "What a neural net needs to know about emotion words," *Computational Intelligence and Applications* **404**, 5311–5316 (1999).

[92] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson, "Harvesting and analysis of weak signals for detecting lone wolf terrorists," *Security Informatics* **2**, 2–11 (2013).

[93] H. Scholsberg, "A scale for the judgment of facial expressions," *Journal of Experimental Psychology* **29**, 497–510 (1941).

[94] M. Wöllmer, F. Eyben, S. Reiter, B. W. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, et al., "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech* (2008), pp. 597–600.

[95] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication* **49**, 787–800 (2007).

[96] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant," *Journal of Personality and Social Psychology* **76**, 805–819 (1999).

[97] H. Spencer, *The Principles of Psychology*, Vol. 1, (Appleton, New York, NY, 1890).

[98] W. Wundt, *Outlines of Psychology* (Wilhelm Engelmann, Leipzig, Germany, 1907).

[99] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science* **18**, 1050–1057 (2007).

[100] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Emotion Theory, Research, and Experience: Theories of Emotion*, Vol. 1, R. Plutchik and H. Kellerman, eds. (Academic Press, New York, NY, 1980), pp. 3–33.

[101] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality* **11**, 273–294 (1977).

[102] R. Reisenzein and T. Hofmann, "Discriminating emotions from appraisal-relevant situational information: Baseline data for structural models of cognitive appraisals," *Cognition & Emotion* **7**, 271–293 (1993).

[103] R. Reisenzein, "Pleasure-arousal theory and the intensity of emotions," *Journal of Personality and Social Psychology* **67**, 525–539 (1994).

[104] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology* (The MIT Press, Cambridge, MA, 1974).

[105] J. A. Russell, "Evidence of convergent validity on the dimensions of affect," *Journal of Personality and Social Psychology* **36**, 1152–1168 (1978).

[106] H. R. Markus and S. Kitayama, "Culture and the self: Implications for cognition, emotion, and motivation," *Psychological Review* **98**, 224–253 (1991).

[107] T. Church, M. S. Katigbak, J. A. S. Reyes, and S. M. Jensen, "Language and organisation of Filipino emotion concepts: Comparing emotion concepts and dimensions across cultures," *Cognition & Emotion* **12**, 63–92 (1998).

[108] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), pp. 4009–4012.

[109] N. Fridja, A. Ortony, J. Sonnemans, and G. Clore, "The complexity of intensity: Issues concerning the structure of emotion intensity," in *Review of Personality and Social Psychology*, Vol. 13, S. Margaret, ed. (SAGE, 1992), pp. 60–89.

[110] J. J. Gross and L. F. Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion Review* **3**, 8–16 (2011).

[111] K. Scherer, "Profiles of emotion-antecedent appraisal: Testing theoretical predictions across cultures," *Cognition & Emotion* **11**, 113–150 (1997).

[112] I. J. Roseman, "Cognitive determinants of emotion: A structural theory," *Review of Personality & Social Psychology* **5**, 11–36 (1984).

[113] I. J. Roseman, "Appraisal determinants of discrete emotions," *Cognition & Emotion* **5**, 161–200 (1991).

[114] B. Weiner, S. Graham, and C. Chandler, "Pity, anger, and guilt: An attributional analysis," *Personality and Social Psychology Bulletin* **8**, 226–232 (1982).

[115] B. Weiner, "An attributional theory of achievement motivation and emotion," *Psychological Review* **92**, 548–573 (1985).

[116] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions* (Cambridge university press, Cambridge, UK, 1988).

[117] N. H. Frijda and B. Mesquita, "The analysis of emotions," in *What Develops in Emotional Development?*, M. Mascolo and S. Griffin, eds. (Plenum Press, New York, NY, 1998), pp. 273–295.

[118] K. R. Scherer, "Emotion as a process: Function, origin and regulation," *Social Science Information/Sur les Sciences Sociales* **21**, 555–570 (1982).

[119] A. Balahur, J. M. Hermida, A. Montoyo, and R. Muñoz, "EmotiNet: a knowledge base for emotion detection in text built on the appraisal theories," in *Natural Language Processing and Information Systems. NLDB 2011. LNCS*, Vol. 6716, R. Muñoz, A. Montoyo, and E. Métais, eds. (Springer, Berlin, Heidelberg, 2011), pp. 27–39.

[120] D. Derks, A. H. Fischer, and A. E. Bos, "The role of emotion in computer-mediated communication: A review," *Computers in Human Behavior* **24**, 766–785 (2008).

[121] S. Planalp, *Communicating Emotion: Social, Moral, and Cultural Processes* (Cambridge University Press, Cambridge, UK, 1999).

[122] J. M. Wilce, *Language and Emotion* (Cambridge University Press, Cambridge, UK, 2009).

[123] S. Schachter, "Cognition and peripheralist-centralist controversies in motivation and emotion," in *Handbook of Psychobiology*, M. Gazzaniga and C. Blakemore, eds. (Academic Press, New York, NY, 1975), pp. 529–564.

[124] P. N. Johnson-Laird and K. Oatley, "The language of emotions: An analysis of a semantic field," *Cognition and Emotion* **3**, 81–123 (1989).

[125] R. Reisenzein, "On Oatley and Johnson-Laird's theory of emotion and hierarchical structures in the affective lexicon," *Cognition & Emotion* **9**, 383–416 (1995).

[126] A. Wierzbicka, "Emotion, language, and cultural scripts," in *Emotion and Culture: Empirical Studies of Mutual Influence*, S. Kitayama and H. R. Markus, eds. (American Psychological Association, Washington, DC, 1994), pp. 133–196.

[127] Z. Kövecses, *Metaphor and Emotion: Language, Culture, and Body in Human Feeling* (Cambridge University Press, Cambridge, 2000).

[128] R. Paul, *Language Disorders from Infancy through Adolescence: Assessment & Intervention* (Elsevier Health Sciences, St. Louis, Missouri, 2007).

[129] T. Rus, "A unified language processing methodology," *Theoretical Computer Science* **281**, 499–536 (2002).

[130] A. Stavrianou, *Modeling and mining of web discussions*, PhD thesis (Doctoral Dissertation, Lyon: Universite De Lyon, 2011).

[131] R. Huddleston, *Introduction to the Grammar of English* (Cambridge University Press, Cambridge, UK, 1984).

[132] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A Comprehensive Grammar of the English Language* (Longman, London and New York, 1985).

[133] L. Hickey, "Stylistics, pragmatics and pragmastylistics," *Revue Belge de Philologie et D'histoire* **71**, 573–586 (1993).

[134] N. Ishihara and A. D. Cohen, *Teaching and Learning Pragmatics: Where Language and Culture Meet* (Routledge, New York, NY, 2010).

[135] M. Bates, "Models of natural language understanding," *Proceedings of the National Academy of Sciences of the United States of America* **92**, 9977–9982 (1995).

[136] M. G. Frank, A. Maroulis, and D. J. Griffin, "The voice," in *Nonverbal Communication: Science and Applications*, D. Matsumoto, M. Frank, and H. Hwang, eds. (SAGE Publications, Inc., Thousand Oaks, California, 2013), pp. 53–74.

[137] R. Parkins, "Gender and emotional expressiveness: An analysis of prosodic features in emotional expression," *Pragmatics and Intercultural Communication* **5**, 46–54 (2012).

[138] V. Lee and H. Wagner, "The effect of social presence on the facial and verbal expression of emotion and the interrelationships among emotion components," *Journal of Nonverbal Behavior* **26**, 3–25 (2002).

[139] J. B. Walther and K. P. D'Addario, "The impacts of emoticons on message interpretation in computer-mediated communication," *Social Science Computer Review* **19**, 324–347 (2001).

[140] J. Siegel, V. Dubrovsky, S. Kiesler, and T. W. McGuire, "Group processes in computer-mediated communication," *Organizational Behavior and Human Decision Processes* **37**, 157–187 (1986).

[141] H. Yang, A. Willis, A. De Roeck, and B. Nuseibeh, "A hybrid model for automatic emotion recognition in suicide notes," *Biomedical Informatics Insights* **5**, 17–30 (2012).

[142] J. Tao, "Context based emotion detection from text input," in *Interspeech* (2004), pp. 1–4.

[143] A. Stavrianou, P. Andritsos, and N. Nicoloyannis, "Overview and semantic issues of text mining," *ACM Sigmod Record* **36**, 23–34 (2007).

[144] A. Verma, *Techniques for Human Emotion Recognition in Text Documents* (Dept. of Computer Science and Engineering. IIT Kanpur, 2009-10).

[145] A. Osherenko, *Opinion Mining and Lexical Affect Sensing* (Universität Augsburg, Augsburg, Germany, 2010).

[146] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo, "A survey on the role of negation in sentiment analysis," in *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP '10)* (Association for Computational Linguistics, Stroudsburg, PA, 2010), pp. 60–68.

[147] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation* **39**, 165–210 (2005).

[148] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 conference on Empirical methods in Natural Language Processing (EMNLP '03)* (Association for Computational Linguistics, Stroudsburg, PA, 2003), pp. 105–112.

[149] M. Ostendorf, E. Shriberg, and A. Stolcke, "Human language technology: Opportunities and challenges," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'05)* (IEEE, 2005), pp. 949–952.

[150] J. T. Irvine, "Language and affect: Some cross-cultural issues," in *Contemporary Perceptions of Language: Interdisciplinary Dimensions*, H. Byrnes, ed. (Georgetown University Press, Washington, D.C., 1982), pp. 31–47.

[151] P. Ekman, "Facial signs: Facts, fantasies, and possibilities," in *Sight, Sound, and Sense*, T. Sebeok, ed. (Indiana University Press, Bloomington, Indiana, 1978), pp. 124–156.

[152] C. K. Chung and J. W. Pennebaker, "Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language," *Journal of Research in Personality* **42**, 96–132 (2008).

[153] J. Nixon, S. Blandy, C. Hunter, A. Jones, and K. Reeve, *Tackling Anti-Social Behaviour in Mixed Tenure Areas* (Office of the Deputy Prime Minister, London, 2003).

[154] A. Bandura, *Aggression: A Social Learning Analysis* (Prentice-Hall, Oxford, England, 1973).

[155] E. Andreou and P. Metallidou, "The relationship of academic and social cognition to behaviour in bullying situations among Greek primary school children," *Educational Psychology* **24**, 27–41 (2004).

[156] M. E. O'Toole, *The School Shooter a Threat Assessment Perspective* (FBI Academy, Quantico, Virginia, 1999).

[157] T. E. Moffitt, "Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy," *Psychological Review* **100**, 674–701 (1993).

[158] D. Clarke, *Pro-social and Anti-social Behaviour* (Taylor & Francis, Abingdon, UK, 2003).

[159] E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist, "Integrating models of personality and emotions into lifelike characters," in *Affective Interactions: Towards a New Generation of Computer Interfaces*, A. Paiva, ed. (Springer, Berlin, Heidelberg, 2000), pp. 150–165.

[160] C. E. Izard and B. P. Ackerman, "Motivational, organizational, and regulatory functions of discrete emotions," in *Handbook of Emotions (2nd Edition)*, M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, eds. (Guilford Press, New York, NY, 2000), pp. 253–264.

[161] G. F. Loewenstein, E. U. Weber, C. K. Hsee, and N. Welch, "Risk as feelings," *Psychological Bulletin* **127**, 267–286 (2001).

[162] R. F. Baumeister, E. Masicampo, and K. D. Vohs, "Do conscious thoughts cause behavior?," *Annual Review of Psychology* **62**, 331–361 (2011).

[163] R. F. Baumeister, K. D. Vohs, C. Nathan DeWall, and L. Zhang, "How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation," *Personality and Social Psychology Review* **11**, 167–203 (2007).

[164] J. C. Bishop, *Natural Agency: An Essay on the Causal Theory of Action* (Cambridge University Press, 1989).

[165] D. Davidson, *Essays on Actions and Events: Philosophical Essays Volume 1* (Oxford University Press, Clarendon, Oxford, 2001).

[166] A. I. Goldman, *Theory of Human Action* (Princeton University Press, Princeton, New Jersey, 1970).

[167] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological Review* **110**, 145–172 (2003).

[168] R. B. Zajonc, "Feeling and thinking: Preferences need no inferences," *American Psychologist* **35**, 151–175 (1980).

[169] J. Zhu and P. Thagard, "Emotion and action," *Philosophical Psychology* **15**, 19–36 (2002).

[170] K. P. Leith and R. F. Baumeister, "Why do bad moods increase self-defeating behavior? Emotion, risk tasking, and self-regulation," *Journal of Personality and Social Psychology* **71**, 1250–1267 (1996).

[171] C. A. Anderson and D. C. Anderson, "Ambient temperature and violent crime: Tests of the linear and curvilinear hypotheses," *Journal of Personality and Social Psychology* **46**, 91–97 (1984).

[172] G. K. Manucia, D. J. Baumann, and R. B. Cialdini, "Mood influences on helping: Direct effects or side effects?," *Journal of Personality and Social Psychology* **46**, 357–364 (1984).

[173] R. B. Cialdini, B. L. Darby, and J. E. Vincent, "Transgression and altruism: A case for hedonism," *Journal of Experimental Social Psychology* **9**, 502–516 (1973).

[174] R. Baron and D. Byrne, "Interpersonal attraction: Getting acquainted, becoming friends," in *Social Psychology: Understanding Human Interaction* (Allyn and Bacon, Boston, 1994), pp. 262–303.

[175] R. F. Baumeister, C. N. DeWall, K. D. Vohs, and J. L. Alquist, "Does emotion cause behavior (apart from making people do stupid, destructive things)," in *Then a Miracle Occurs: Focusing on Behavior in Social Psychological Theory and Research*, C. Agnew, D. Carlston, W. Graziano, and J. Kelly, eds. (Oxford University Press, New York, NY, 2010), pp. 119–136.

[176] D. M. Tice, E. Bratslavsky, and R. F. Baumeister, "Emotional distress regulation takes precedence over impulse control: If you feel bad, do it!," *Journal of Personality and Social Psychology* **80**, 53–67 (2001).

[177] J. E. LeDoux, *The Emotional Brain: The Mysterious Underpinnings of Emotional Life* (Simon & Schuster, New York, 1996).

[178] R. F. Baumeister, *Evil: Inside Human Violence and Cruelty* (W. H. Freeman and Company, New York, NY, 1997).

[179] W. Wood, J. M. Quinn, and D. A. Kashy, "Habits in everyday life: thought, emotion, and action," *Journal of Personality and Social Psychology* **83**, 1281–1297 (2002).

[180] O. H. Green, *The Emotions: A Philosophical Theory (vol.53)* (Kluwer academic publishers, Dordrecht, The Netherlands, 1992).

[181] D. Zillmann, "Excitation transfer in communication-mediated aggressive behavior," *Journal of Experimental Social Psychology* **7**, 419–434 (1971).

[182] L. Berkowitz, "Pain and aggression: Some findings and implications," *Motivation and Emotion* **17**, 277–293 (1993).

[183] A. D. Berkowitz, "Applications of social norms theory to other health and social justice issues," in *The Social Norms Approach to Preventing School and College Age Substance Abuse: A Handbook for Educators, Counselors, and Clinicians*, W. H. Perkins, ed. 2003), pp. 259–279.

[184] K. M. Lagerspetz, K. Björkqvist, and T. Peltonen, "Is indirect aggression typical of females? Gender differences in aggressiveness in 11-to 12-year-old children," *Aggressive Behavior* **14**, 403–414 (1988).

[185] K. Björkqvist, K. M. Lagerspetz, and A. Kaukiainen, "Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression," *Aggressive Behavior* **18**, 117–127 (1992).

[186] N. R. Crick and J. K. Grotpeter, "Relational aggression, gender, and social-psychological adjustment," *Child Development* **66**, 710–722 (1995).

[187] K. M. Lagerspetz and K. Björkqvist, "Indirect aggression in boys and girls," in *Aggressive Behavior: Current Perspectives*, L. R. Huesmann, ed. (Springer US, Boston, MA, 1994), pp. 131–150.

[188] R. Loeber and D. Hay, "Key issues in the development of aggression and violence from childhood to early adulthood," *Annual Review of Psychology* **48**, 371–410 (1997).

[189] L. Berkowitz, "Aversive conditions as stimuli to aggression," *Advances in Experimental Social Psychology* **15**, 249–288 (1982).

[190] L. Berkowitz, "Frustration-aggression hypothesis: examination and reformulation," *Psychological Bulletin* **106**, 59–73 (1989).

[191] J. M. Gottman, *What Predicts Divorce?: The Relationship Between Marital Processes and Marital Outcomes* (Lawrence Erlbaum Associates, Inc., New York, NY, 1994).

[192] J. De Rivera, "The structure of emotional relationships," *Review of Personality & Social Psychology* **5**, 116–145 (1984).

[193] N. H. Frijda, P. Kuipers, and E. Ter Schure, "Relations among emotion, appraisal, and emotional action readiness," *Journal of Personality and Social Psychology* **57**, 212–228 (1989).

[194] I. J. Roseman, C. Wiest, and T. S. Swartz, "Phenomenology, behaviors, and goals differentiate discrete emotions," *Journal of Personality and Social Psychology* **67**, 206 (1994).

[195] J. R. Averill, *Anger and Aggression: An Essay on Emotion* (Springer-Verlag, New York: NY, 1982).

[196] J. R. Averill, "Studies on anger and aggression: Implications for theories of emotion," *American Psychologist* **38**, 1145–1160 (1983).

[197] D. H. O'Hair, D. R. Bernard, and R. R. Roper, "Communication-based research related to threats and ensuing behavior," in *Threatening Communications and Behavior: Perspectives on the Pursuit of Public Figures*, C. Chauvin, ed. (National Academies Press, Washington, DC, 2011), pp. 33–74.

[198] E. A. Locke and G. P. Latham, "Building a practically useful theory of goal setting and task motivation: A 35-year odyssey," *American Psychologist* **57**, 705–717 (2002).

[199] J. P. Tangney, "Situational detenninants of shame and guilt in young adulthood," *Personality and Social Psychology Bulletin* **18**, 199–206 (1992).

[200] D. R. Olson, *The World on Paper: The Conceptual and Cognitive Implications of Reading and Writing* (Cambridge University Press, Cambridge, 1994).

[201] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open Mind Common Sense: Knowledge acquisition from the general public," in *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE. OTM 2002*, Vol. 2519, R. Meersman and Z. Tari, eds. (Springer, Berlin, Heidelberg, 2002), pp. 1223–1237.

[202] H. Liu and P. Singh, "Commonsense reasoning in and over natural language," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. KES 2004. LNCS*, Vol. 3215, N. M.G, H. R.J, and J. L.C, eds. (Springer, Berlin, Heidelberg, 2004), pp. 293–306.

[203] A. Cruse, *Meaning in Language: An Introduction to Semantics and Pragmatics* (Oxford University Press, New York, NY, 2004).

[204] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing* (Cambridge, MA: MIT Press, 1999).

[205] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)* **34**, 1–47 (2002).

[206] G. Miner, J. Elder IV, T. Hill, B. Nisbet, D. Dursun, and A. Fast, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (Academic Press, Waltham, MA, 2012).

[207] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics* **3**, 143–157 (2009).

[208] L. Zhang, S. Ferrari, and P. Enjalbert, "Opinion analysis: the effect of negation on polarity and intensity," in *KONVENS workshop PATHOS - 1st Workshop on Practice and Theory of Opinion Mining and Sentiment Analysis* (2012), pp. 282–290.

[209] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS Transactions on Computers* **4**, 966–974 (2005).

[210] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM* **18**, 613–620 (1975).

[211] 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 (EMNLP '02)* (Association for Computational Linguistics, Stroudsburg, PA, 2002), pp. 79–86.

[212] B. Yu, "An evaluation of text classification methods for literary study," *Literary and Linguistic Computing* **23**, 327–343 (2008).

[213] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *Journal of Documentation* **60**, 503–520 (2004).

[214] E. Loper and S. Bird, "NLTK: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, (ETMTNLP '02)* (Association for Computational Linguistics, Stroudsburg, PA, 2002), pp. 63–70.

[215] E. F. Tjong Kim Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task: Chunking," in *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7 (ConLL '00), Vol. 7* (Association for Computational Linguistics, Stroudsburg, PA, 2000), pp. 127–132.

[216] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *Journal of Advances in Information Technology* **1**, 4–20 (2010).

[217] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis* **21**, 267–297 (2013).

[218] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Transactions on Information Systems (TOIS)* **26**, 12 (2008).

[219] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *International Conference on Privacy, Security, Risk and Trust (PASSAT), and International Conference on Social Computing (SocialCom)* (IEEE, 2012), pp. 71–80.

[220] B. Kessler, G. Numberg, and H. Schütze, "Automatic detection of text genre," in *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics (EACL '97)* (Association for Computational Linguistics, Stroudsburg, PA, 1997), pp. 32–38.

[221] C. Rosé, Y.-C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer, "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning," *International Journal of Computer-supported Collaborative Learning* **3**, 237–271 (2008).

[222] D. Lin, "Extracting collocations from text corpora," in *First Workshop on Computational Terminology* (1998), pp. 57–63.

[223] L. Ciya, A. Shamim, and D. Paul, "Feature preparation in text categorization," *Oracle Text Selected Papers and Presentations* 1–8 (2001).

[224] T. Fu, C.-N. Huang, and H. Chen, "Identification of extremist videos in online video sharing sites," in *IEEE International Conference on Intelligence and Security Informatics, ISI'09* (IEEE, 2009), pp. 179–181.

[225] M. Mishra, V. K. Mishra, and H. Sharma, "Question classification using semantic, syntactic and lexical features," *International Journal of Web & Semantic Technology* **4**, 39–47 (2013).

[226] T. Kakkonen and G. G. Kakkonen, "SentiProfiler: creating comparable visual profiles of sentimental content in texts," in *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop* (2011), pp. 62–69.

[227] C. S. Montero, M. Munezero, and T. Kakkonen, "Investigating the role of emotion-based features in author gender classification of text," in *Computational Linguistics and Intelligent Text Processing. CICLing 2014. LNCS*, Vol. 8404, A. Gelbukh, ed. (Springer, Berlin, Heidelberg, 2014), pp. 98–114.

[228] P. J. Stone, D. C. Dunphy, and M. S. Smith, *The General Inquirer: A Computer Approach to Content Analysis* (MIT press, Cambridge, Massachusetts, USA, 1966).

[229] R. M. Tong, "An operational system for detecting and tracking opinions in on-line discussion," in *Proceedings of SIGIR Workshop on Operational Text Classification* (2001), pp. 1–6.

[230] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL '98)* (ACL, Stroudsburg, PA, 1997), pp. 174–181.

[231] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)* (Association for Computational Linguistics, Stroudsburg, PA, 2002), pp. 417–424.

[232] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems (TOIS)* **21**, 315–346 (2003).

[233] M. Munezero, T. Kakkonen, C. I. Sedano, E. Sutinen, and C. S. Montero, "EmotionExpert: Facebook game for crowdsourcing annotations for emotion detection," in *IEEE International Games Innovation Conference (IGIC)* (IEEE, 2013), pp. 179–186.

[234] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)* (2006), pp. 417–422.

[235] M. M. Bradley and P. J. Lang, *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Technical Report C-1, the Center for Research in Psychophysiology, University of Florida, 1999).

[236] C. Strapparava and A. Valitutti, "WordNet-Affect: An affective extension of WordNet," in *Proceedings of the International Conference on Language Resources and Evaluation*, Vol. 4 (2004), pp. 1083–1086.

[237] S. M. Mohammad and P. D. Turney, "Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10)* (Association for Computational Linguistics, Stroudsburg, PA, 2010), pp. 26–34.

[238] C. Whissell, "Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language," *Psychological Reports* **105**, 509–521 (2009).

[239] M. Munezero, C. S. Montero, M. Mozgovoy, and E. Sutinen, "Exploiting sentiment analysis to track emotions in students' learning diaries," in *Proceedings of the 13th Koli Calling International Conference on Computing Education Research* (ACM, New York, NY, 2013), pp. 145–152.

[240] M. Munezero, C. S. Montero, M. Mozgovoy, and E. Sutinen, "EmoTwitter–A Fine-Grained Visualization System for Identifying Enduring Sentiments in Tweets," in *Computational Linguistics and Intelligent Text Processing. CICLing 2015. LNCS*, Vol. 9042 (Springer, Cham, 2015), pp. 78–91.

[241] C. Potts, "Sentiment symposium tutorial: Lexicons," (08/13/2011), Stanford Linguistics, `http://sentiment.christopherpotts.net/lexicons.html` (visited on 2017-03-11).

[242] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval* **2**, 1–135 (2008).

[243] S. M. Mohammad, "Word association lexicons: Capturing word-emotion, word-sentiment, and word-colour associations," (04/29/2013), `http://www.saifmohammad.com/WebPages/lexicons.html` (visited on 2017-03-12).

[244] E. F. Kelly and P. J. Stone, *Computer Recognition of English Word Senses*, Vol. 13, (North-Holland Linguistic Series, Amsterdam, 1975).

[245] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology* **17**, 715–734 (2005).

[246] K. Mulcrone, "Detecting emotion in text," (2012), Morris CS Senior Seminar Paper - University of Minnesota.

[247] X. Hu and J. S. Downie, "When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis," in *11th International Society for Music Information Retrieval Conference (ISMIR)* (2010), pp. 619–624.

[248] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology* **63**, 163–173 (2012).

[249] R. D. King, C. Feng, and A. Sutherland, "Statlog: comparison of classification algorithms on large real-world problems," *Applied Artificial Intelligence an International Journal* **9**, 289–333 (1995).

[250] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)* (ACM, New York, NY, 2006), pp. 161–168.

[251] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, I. Maglogiannis, K. Karpouzis, M. Wallace, and J. Soldatos, eds. (IOS Press, Amsterdam, 2007), pp. 3–24.

[252] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems* **14**, 1–37 (2008).

[253] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, vol.1* (Pearson Addison Wesley, Boston, 2006).

[254] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM computing surveys (CSUR)* **27**, 326–327 (1995).

[255] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," (2003), National Taiwan University, Department of Computer Science, Taipei, Taiwan.

[256] D. Correa and A. Sureka, "Solutions to detect and analyze online radicalization: a survey," (2013), arXiv preprint arXiv:1301.4916.

[257] J. Qin, Y. Zhou, G. Lai, E. Reid, M. Sageman, and H. Chen, "The dark web portal project: collecting and analyzing the presence of terrorist groups on the web," in *Intelligence and Security Informatics. ISI 2005. LNCS*, Vol. 3495, P. Kantor and et al., eds. Berlin, Heidelberg, 2005), pp. 623–624.

[258] European Parliament, "Princip: Multilingual system for the analysis and detection of racist and revisionist content on the Internet," **6** (2002), The Internet Anti-Fascist.

[259] 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," (06/30/2009), `http://www.indect-project.eu/files/deliverables/public/INDECT_Deliverable_4.1_v20090630a.pdf` (visited on 2017-03-10).

[260] E. Spertus, "Smokey: Automatic recognition of hostile messages," in *Innovative Applications of Artificial Intelligence (IAAI) '97* (1997), pp. 1058–1065.

[261] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in *Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)* (ACM, New York, NY, 2004), pp. 468–469.

[262] M. Last, A. Markov, and A. Kandel, "Multi-lingual detection of terrorist content on the web," in *Intelligence and Security Informatics. LNCS*, Vol. 3917, H. Chen, F. Wang, C. Yang, D. Zeng, M. Chau, and K. Chang, eds. (Springer, Berlin, Heidelberg, 2006), pp. 16–30.

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

[264] H. Chen, "Sentiment and affect analysis of dark web forums: Measuring radicalization on the Internet," in *IEEE International Conference on Intelligence and Security Informatics* (IEEE, 2008), pp. 104–109.

[265] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," in *Proceedings of the Content Analysis in the WEB* (2009), pp. 1–7.

[266] C. Huang, T. Fu, and H. Chen, "Text-based video content classification for online video-sharing sites," *Journal of the American Society for Information Science and Technology* **61**, 891–906 (2010).

[267] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive language detection using multi-level classification," in *Advances in Artificial Intelligence. AI 2010. LNCS*, Vol. 6085, A. Farzindar and V. Kešelj, eds. (Springer, Berlin, Heidelberg, 2010), pp. 16–27.

[268] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Workshop on the Social Mobile Web, International AAAI Conference on Weblogs and Social Media* (2011), pp. 11–17.

[269] D. Bogdanova, P. Rosso, and T. Solorio, "On the impact of sentiment and emotion based features in detecting online sexual predators," in *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (WASSA '12)* (Association for Computational Linguistics, Stroudsburg, PA, 2012), pp. 110–118.

[270] L. P. Del Bosque and S. E. Garza, "Aggressive text detection for cyberbullying," in *Human-Inspired Computing and Its Applications. MICAI 2014. LNCS*, Vol. 8856, A. Gelbukh, F. Espinoza, and G.-H. S.N, eds. (Springer, Cham, 2014), pp. 221–232.

[271] J.-M. Xu, X. Zhu, and A. Bellmore, "Fast learning for sentiment analysis on bullying," in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '12)* (ACM, New York, NY, 2012), pp. 1–6.

[272] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques," in *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13)* (ACM, New York, NY, 2013), pp. 195–204.

[273] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *Journal of Personality and Social Psychology* **66**, 310–328 (1994).

[274] R. T. Ross, "A statistic for circular series," *Journal of Educational Psychology* **29**, 384–389 (1938).

[275] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (2005), pp. 579–586.

[276] K. Krippendorff, *Content Analysis: An Introduction to its Methodology, 3rd edition* (Sage Publications, Inc., California, 2012).

[277] M. Schröder, P. Baggia, F. Burkhardt, J.-C. Martin, C. Pelachaud, C. Peter, B. Schuller, I. Wilson, and E. Zovato, "Elements of an EmotionML 1.0," (11/20/2008), W3C Incubator Group Report, `https://www.w3.org/2005/Incubator/emotion/XGR-emotionml/` (visited on 2017-03-11).

[278] M. Schröder, C. Pelachaud, K. Ashimura, P. Baggia, F. Burkhardt, A. Oltramari, C. Peter, and E. Zovato, "Vocabularies for EmotionML. W3C working draft," (04/07/2011), World Wide Web Consortium, `https://www.w3.org/TR/2011/WD-emotion-voc-20110407/` (visited on 2017-03-11).

[279] J. W. Pennebaker, "Current issues and new directions in Psychology and Health: Listening to what people say–the value of narrative and computational linguistics in health psychology," *Psychology & Health* **22**, 631–635 (2007).

[280] P. Lewicki and T. Hill, "Statistics: methods and applications," *Statsoft* (2006).

[281] J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers, San Mateo, California, 1993).

[282] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter* **11**, 10–18 (2009).

[283] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence (UAI'95)*, B. Philippe and H. Steve, eds. (Morgan Kaufmann Publishers Inc., San Francisco, CA, 1995), pp. 338–345.

[284] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," (04/01/1998), Microsoft Research, `https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/` (visited on 2017-03-10).

[285] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice-Hall, Englewood Cliffs, 1995).

[286] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sensing of Environment* **80**, 185–201 (2002).

[287] C. Li, J. Wang, L. Wang, L. Hu, and P. Gong, "Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery," *Remote Sensing* **6**, 964–983 (2014).

[288] A. Janecek, W. N. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," *FSDM* **4**, 90–105 (2008).

[289] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference," *Journal of Personality and Social Psychology* **77**, 1296–1312 (1999).

[290] J. W. Pennebaker, "What our words can say about us: Toward a broader language psychology," *Psychological Science Agenda* **15**, 8–9 (2002).

[291] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual Review of Psychology* **54**, 547–577 (2003).

[292] N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, "The psychology of word use in depression forums in English and in Spanish: Texting two text analytic approaches," in *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)* (2008), pp. 102–108.

[293] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion* **18**, 1121–1133 (2004).

[294] S. W. Stirman and J. W. Pennebaker, "Word use in the poetry of suicidal and nonsuicidal poets," *Psychosomatic Medicine* **63**, 517–522 (2001).

[295] W. Weintraub, *Verbal Behavior in Everyday Life* (Springer Publishing Co, New York, NY, 1989).

[296] J. E. Owen, E. J. Yarbrough, A. Vaga, and D. C. Tucker, "Investigation of the effects of gender and preparation on quality of communication in Internet support groups," *Computers in Human Behavior* **19**, 259–275 (2003).

[297] J. W. Pennebaker, C. K. Chung, G. A. Ireland, Molly E, and R. J. Booth, "The development and psychometric properties of LIWC2007," (2007), LIWC net.

[298] A. Montagu, *The Anatomy of Swearing* (Mcmillan Publishing Company, New York, NY, 1967).

[299] J. T. Hancock, M. T. Woodworth, and S. Porter, "Hungry like the wolf: A word-pattern analysis of the language of psychopaths," *Legal and Criminological Psychology* **18**, 102–114 (2013).

[300] F. Johansson, L. Kaati, and M. Sahlgren, "Detecting linguistic markers of violent extremism in online environments," in *Countering Violent Extremism and Radicalisation in the Digital Era*, M. Khader, L. Neo, G. Ong, E. Tan, and J. Chin, eds. (IGI Global, Hershey, PA, 2016), pp. 374–390.

[301] T. Jay, *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets* (John Benjamins Publishing, Philadephia, 1992).

[302] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98. LNCS (Lecture Notes in Artificial Intelligence)*, Vol. 1398, C. Nédellec and C. Rouveirol, eds. (Springer, Berlin, Heidelberg, 1998), pp. 137–142.

[303] J. D. Rennie, L. Shih, J. Teevan, D. R. Karger, et al., "Tackling the poor assumptions of naive bayes text classifiers," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Vol. 3 (2003), pp. 616–623.

[304] J. W. Pennebaker and A. Graybeal, "Patterns of natural language use: Disclosure, personality, and social integration," *Current Directions in Psychological Science* **10**, 90–93 (2001).

[305] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Transactions on Affective Computing* **5**, 101–111 (2014).

[306] D. Inkpen, "Text mining in social media for security threats," in *Recent Advances in Computational Intelligence in Defense and Security*, R. Abielmona, R. Falcon, Z.-H. N, and A. H.A, eds. (Springer International Publishing, Cham, 2016), pp. 491–517.

[307] B. Joy and R. Kurzweil, "Future shock: High technology and the human aspect," (12/7/2001), Hoover Institution, `http://www.hoover.org/research/future-shock-high-technology-and-human-prospect` (visited on 2017-03-10).

MYRIAM DOUCE MUNEZERO

The words we use and our writing styles can reveal information about our preferences, thoughts, emotions, and behaviors. Using this information, this work has demonstrated the potential of using natural language processing techniques to develop state-of-the-art solutions to detect antisocial behavior - behavior carried out with the immediate intent to cause harm (for instance violence and terrorism) - in user-generated written content.

UNIVERSITY OF
EASTERN FINLAND

uef.fi

**PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND**
Dissertations in Forestry and Natural Sciences

ISBN 978-952-61-2463-6
ISSN 1798-5668