Identifying Expressions of Emotions and Their Stimuli in Text

by

Diman Ghazi

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the Ph.D. degree in
Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

Emotions are among the most pervasive aspects of human experience. They have long been of interest to social and behavioural sciences. Recently, emotions have attracted the attention of researchers in computer science and particularly in computational linguistics. Computational approaches to emotion analysis have also focused on various emotion modalities, but there is less effort in the direction of automatic recognition of the emotion expressed. Although some past work has addressed detecting emotions, detecting why an emotion arises is ignored.

In this work, we explore the task of classifying texts automatically by the emotions expressed, as well as detecting the reason why a particular emotion is felt.

We believe there is still a large gap between the theoretical research on emotions in psychology and emotion studies in computational linguistics. In our research, we try to fill this gap by considering both theoretical and computational aspects of emotions. Starting with a general explanation of emotion and emotion causes from the psychological and cognitive perspective, we clarify the definition that we base our work on. We explain what is feasible in the scope of text and what is practically doable based on the current NLP techniques and tools.

This work is organized in two parts: first part on *Emotion Expression* and the second part on *Emotion Stimulus*.

In emotion expression detection, we start with shallow methodologies, such as corpus-based and lexical-based, followed by deeper methods considering syntactic and semantic relations in text. First, we demonstrate the usefulness of external knowledge resources, such as polarity and emotion lexicons, in automatic emotion detection. Next, we provide a description of the more advanced features chosen for characterizing emotional content based on the syntactic structure of sentences, as well as the machine learning techniques adopted for emotion classification.

The main novelty of our learning methodology is that it breaks down a problem into hierarchical steps. It starts from a simpler problem to solve, and uses what is learnt to extend the solution to solve harder problems. Here, we are learning emotion of sentences with one emotion word and we are extending the solution to sentences with more than one

emotion word.

Next, we frame the detection of causes of emotions as finding a stimulus frame element as defined for the emotion frame in FrameNet – a lexical database of English based on the theory of meaning called Frame Semantics, which was built by manually annotating examples of how words are used in actual texts. According to FrameNet, an emotion stimulus is the person, event, or state of affairs that evokes the emotional response in the Experiencer. We believe it is the closest definition to emotion cause in order to answer *why* the experiencer feels that emotion.

We create the first ever dataset annotated with both emotion stimulus and emotion class; it can be used for evaluation or training purposes. We applied sequential learning methods to the dataset. We explored syntactic and semantic features in addition to corpus-based features. We built a model which outperforms all our carefully-built baselines. To show the robustness of our model and to study the problem more thoroughly, we apply those models to another dataset (that we used for the first part as well) to go deeper than detecting the emotion expressed and also detect the stimulus span which explains why the emotion was felt.

Although we first address emotion expression and emotion stimulus independently, we believe that an emotion stimulus and the emotion itself are not mutually independent. In the last part, we address the relation of emotion expression and emotion stimulus by building four cases: both emotion expression and emotion stimulus occur at the same time, none of them appear in the text, there is only emotion expression, or only the emotion stimulus exists while there is no explicit mention of the emotion expressed. We found the last case the most challenging, so we study it in more detail.

Finally, we showcase how a clinical psychology application can benefit from our research. We also conclude our work and explain the future directions of this research.

Note: see `http://www.eecs.uottawa.ca/~diana/resources/emotion_stimulus_data/` for all the data built for this thesis and discussed in it.

Acknowledgements

This work has occupied all of my waking thoughts for the last couple of years. Completing it gives me at once satisfaction and bitter sadness as I close these chapters of my life. However, I cannot finish without taking the time to thank the people who have helped me along the way, contributed to my research, and stood by me always.

First I would like to thank my supervisors. To Stan Szpakowicz, whose critical eye has helped me deliver something of quality. Thank you for pushing me to constantly dive deeper and for holding me to your high standards. Your insistence on attention to detail while still being able to see the big picture was a unique asset that kept me grounded and focused on what matters. To Diana Inkpen, for being both an advisor and mentor who is supportive and encouraging, sincerely cares about students professionally and personally. Thank you for always ensuring that my research was relevant to real life by seeking industrial collaborations and social applications.

Thanks to the members of my committee – Cecilia Alm, Ash Asudeh, Nathalie Japkowicz and Peter Turney – for their insightful, wise comments and feedbacks.

To the human annotators that have contributed in creating the dataset that was used in Chapter 5 and for to evaluating my methodology in Chapter 6, thank you for all of your time and energy.

Next I must thank Hamid, whose words of encouragement have motivated me to face this challenge with enthusiasm. Thanks for always having belief in me and confidence in that I can achieve anything.

To my dear friends Laleh and Allyshia, whose care and support have been a treasure to me. Thank you for your nurturing presence during times of high pressure and for your friendship.

Last but certainly not least, I would like to thank my family. Thanks for your persistent love, care and support which always brought me forward in life. Despite being miles away, your presence in the back of my mind while pursuing my PhD was a driving force. I owe you the constant desire that I have to aim high, to achieve big, and to strive for more, for better.

This journey has forever changed me. Rising to the challenge, amidst its ups and downs, its disappointments and rewards, has shaped the person I am today. At the end of this experience, I find myself with new perspectives on solving problems and better equipped to face the more real and less scripted challenges ahead. Thank you all for making this possible.

Contents

List of Tables

List of Figures

Chapter 1

Motivation, overview, contributions

1.1 Research motivation

Language is a powerful tool to communicate and convey information; therefore at the crossroads of computer science, artificial intelligence, and linguistics a discipline has emerged that is concerned with the interactions between computers and human (natural) languages, called Natural Language Processing (NLP). Recently there has been a growing interest in automatic identification and extraction of sentiment, opinions and emotions in text, due to the huge amount of available text data on the Internet, such as blogs, emails and chats which are full of emotions. This wave of activity links to the rapid growth of the social media, namely reviews, discussion fora, blogs, micro-blogs, Twitter and social networks, and the ensuing easy access to a mass of subjective and emotional data recorded in digital form.

Another reason why sentiment analysis is such an active area of NLP research is that its applications have spread to multiple domains, from consumer product reviews, health care and financial services to social events and political elections. As affect discovery, extraction and generation continue to penetrate areas that take advantage of computational linguistics methods, work in areas such as medical decision-making and mental health will benefit from attention to approaches considering affect Alm (2012). Although our research can benefit all of these applications, our main focus and motivation for this research is the latter application, which we will explore more here.

1.1.1 Clinical psychology and mental health

Mental illness is a major health concern worldwide. According to the Canadian Mental Health Association, mental illness indirectly affects all Canadians at some time through a family member, friend or colleague. Twenty percent of Canadians will personally experience a mental illness in their lifetime. It affects people of every age, education, income level and culture (`http://www.cmha.ca/`).

In any given year, one in five people in Canada experiences a mental health problem or illness, with a cost to the economy of well in excess of $50 billion (`http://strategy.mentalhealthcommission.ca/the-facts/`). In the United States, these costs were projected to increase to $203 billion for 2014.

Moreover, research shows that mental illness is a prominent risk factor for suicide; and more than 90% of people who commit suicide (about 4000 people each year only in Canada) have a mental or addictive disorder. Depression is the most common illness among those who die from suicide, with approximately 60% suffering from this condition (Statistics Canada, 2012). Major depression is a clinical condition that, in any given year, affects about 14 million American adults, or about 6.7% of the population 18 or older. Some of the symptoms of depression are as follows: feeling sad, empty, hopeless, or numb; losing interest in things one used to enjoy; experiencing irritability or anxiety; having trouble making decisions (depression can make it hard to think clearly or concentrate, so making a simple choice can seem overwhelming); feeling guilty or worthless; having thoughts of death and suicide. Although the reasons why people experience depression vary, emotions play an important role both as the cause and the symptoms of depression.

Suicide is one of the leading causes of death for people of all ages. In 2009, it ranked ninth in Canada. Among those aged 15 to 34, suicide was the second leading cause of death, preceded only by accidents (unintentional injuries) (Statistics Canada, 2011).

Nonetheless, global provisions and services for identifying, supporting and treating mental illness of this nature have been considered insufficient (Detels, 2009). In fact, there is no reliable laboratory test for diagnosing most forms of mental illness; typically, the diagnosis is mainly based on the patient's self-reported experiences. However, many patients underreport

psychiatric symptoms or lack the self-awareness to report accurately.

Another issue in treating mental health is that only one in three people who experience a mental health problem or illness reports having sought and received services and treatment. It can be due to various reasons such as high expenses for people without adequate insurance or the fact that many people who live in rural areas cannot even obtain access to a clinician qualified to perform a psychological evaluation. If we want health services to keep up with the problems mentioned, as well as with such a growth of mental illnesses, we need to explore alternatives in the way patients access the health care system, so that it improves both its service quality and timeliness. The importance of clinical psychology as a problem space cannot be overstated.

For clinical psychologists, language plays a central role in diagnosis. Indeed, many clinical instruments fundamentally rely on what is, in effect, manual annotation of patients' language. Automating this process and applying language technology in this domain can have an enormous effect.

Access to the World-Wide Web and social networks has motivated many individuals to report their feelings and share their emotional experience through the Web. Those data can reveal psychiatric symptoms that can be helpful for identifying the onset of depression. Also, physicians can gain insights into the trend of their patients' stages of the illness. Last but not least, bringing language technology to bear on these problems could potentially lead to inexpensive screening measures that could be administered by a wider range of healthcare professionals. It is particularly important since the majority of individuals who present with symptoms of mental health problems do so in the office of a primary care physician's. Given the burden on primary care physicians to diagnose mental health disorders in little time, the American Academy of Family Physicians has recognized the need for diagnostic tools for physicians that are "suited to the realities of their practice" (`http://acl2014.org/acl2014/W14-32/pdf/W14-3200.pdf`).

Although automated language analysis connected with mental health conditions goes back at least as far as the 1990s, it has not been a major focus for computational linguistics compared with other application domains. However, recently there have been two research

activities on this topic: one shared task on detecting the Big-5 personality traits,[1] and another on identifying emotions in suicide notes.[2]

However, no single determinant, including mental illness, is enough on its own to cause suicide. Rather, suicide typically results from the interaction of many factors, for example: mental illness, marital breakdown, financial hardship, deteriorating physical health, a major loss, or a lack of social support. All these situations are highly emotionally intensive cases, and the language used to express them can indicate strong emotions such as feeling of worthlessness, guilt, helplessness and self-hatred that characterize major depression. Other signals can be loss of interest or pleasure in normally enjoyable activities, long-lasting negative feeling towards a particular cause, and increased negative affects. It is also well established that people who suffer from major depression tend to focus their attention on unhappy and unflattering information, to interpret ambiguous information negatively, and to keep pervasively pessimistic beliefs (Rude et al., 2004).

We would like to highlight how our research can benefit these main concerns in the clinical psychology area. Although little previous work addresses detecting emotions in suicide notes (Desmet and Hoste, 2013; Homan et al., 2014) and predicting depression (Choudhury et al., 2013), to the best of our knowledge there is no work that would also explain what evoked the emotions felt by the people suffering from depression or those at risk of suicide.

We believe our research in emotion and emotion stimulus detection can be used in detecting and predicting affective disorder in individuals, with the possibility of explaining the reason behind it. This can complement and extend traditional approaches to diagnosis.

Our findings and methods can be useful in identifying the onset of major depression, for use by healthcare agencies; or on behalf of individuals, enabling those who suffer from depression to be more proactive about their mental health. We also believe our research paves the way for further investigation of the relation of emotions and depression as well

[1] `http://mypersonality.org/wiki/doku.php?id=wcpr13`

[2] Participants were asked to find emotions in suicide notes. The data come from a collection of over 1,000 notes that were written by people who have died by suicide. The data are hand-annotated. This track is supported by NIH Shared Task 2010 - Analysis of Suicide Notes for Subjective Information (Pestian et al., 2012).

as the relation of emotions and their stimuli with the risk of committing suicide, and how to detect depression based on the emotion expressed in text which will result in depression prevention and detection of those at risk of committing suicide. That will not only save billions of dollars but can also save thousands of lives.

That being said, there are no publicly available clinical data on people who suffer from depression, or those at the risk of committing suicide; therefore, we focus our research on detecting emotions and their stimuli in blog data. Blog data are a source largely used to report one's feelings and express oneself through the Web, so they are publicly available, and have the most similar structure to the data that people who suffer from depression publish on social media and World-wide Web.

Considering the motivation of our research, we would like to highlight some of our objectives in this regard.

- Classify sentences into fine-grained emotion categories that are most commonly used in NLP research, called basic emotions.
- Address psychological aspects of emotions and explaining how they can be useful in computational approaches to emotions.
- Apply deeper methods, considering syntactic and semantic relations in text rather than only shallow methodologies such as corpus-based and lexically motivated methods.
- Detect what evokes a particular emotion, which is named emotion stimulus detection in our work.

1.2 Contributions

In this section, we would like to list the current limitations in the emotion analysis research area. As our intended contribution, we will also explain how we are planning to alleviate these limitations.

- Even though the cognitive structure of emotions is addressed from a theoretical perspective (Ortony et al., 1988), it has been barely acknowledged in computational linguistics. There is still a large gap between the theoretical research on psychology of

emotions and emotion studies in computational linguistics. We believe computational linguists should at least move in that direction. In our research, we aim to contribute towards filling the gap by considering both theoretical and computational aspects of emotions. Computationally, we first start with simpler methods such as lexical and corpus-based approaches, followed by deeper methods that consider syntactic and semantic relations in text. Using structural, syntactic and semantic features, we detect emotion stimulus.

- Most of the previous work in emotion analysis is tied to sentiment and polarity classification. In computational linguistics, emotion analysis tends not to be addressed independently, rather it has usually been considered as a sub-area of polarity detection in sentiment analysis. Most of the methods in emotion classification are adopted from polarity analysis. However, in our research, we address emotions as independent states that have their own characteristics, which makes it necessary to process them differently.

- Prior work on text-based emotion recognition has been limited by available datasets either for supervised machine learning approaches or evaluation purposes; therefore the research in this area is limited to the type of datasets that are available. The limitation in resources has also affected the direction of the research on emotion recognition. In this thesis, we go beyond the existing data and resources. We study emotion stimulus, which is quite a new field in emotion analysis; therefore we collect data and build a new dataset using FrameNet.[3]

- Previously, the emotion detection problem has been oversimplified by assuming few discrete emotion classes, while there has been very little attempt, if any, to determine what stimulus causes the emotions. Hence, it is of limited use in explaining when a particular emotion will be felt. We believe that going beyond detecting emotions only and addressing emotion stimuli builds a framework for more thorough and deeper

[3]In a quite recent work, Neviarouskaya and Aono (2013) built a small dataset with 523 sentences for 22 emotions. However, we found it too sparse to be used for machine learning-based cause detection methods.

analysis of emotions.

- We study emotion stimulus detection in sentences with explicitly expressed emotions as well as sentences with implicit emotion expression. We showcase that using emotion stimulus can help detect the emotion expressed in sentences with implicit emotion expression.
- We explain how a clinical psychology application can benefit from our research.

1.3 Outline of the thesis

The thesis is organized as follows:

- Chapter 2 discusses the theoretical definition of emotions as well as the existing methodologies to study emotions from that perspective. We also mention different theories on basic emotion categories.
- Chapter 3 builds a background for the work presented in the remainder of the thesis. We mention previous work on computational analysis of emotions as well as the various knowledge resources used in automatic emotion analysis.
- Chapters 4, 5 and 6 cover our main methodologies and experiments. While chapter 4 addresses explicit *Emotion Expression* classification into Ekman's six emotion classes (Ekman, 1992), we discuss *Emotion Stimulus* detection in chapter 5. Finally in chapter 6 we study emotion stimulus in regard to the emotion that is expressed in the sentence.
- Chapter 4 presents experiments on fine-grained emotion classification. It starts with explaining a simple method and moves toward our more sophisticated methods. First, the usefulness of external knowledge resources, such as polarity and emotion lexicons, in automatic emotion detection task is demonstrated. Then, we provide a description of the more elaborate features chosen for characterizing emotional content based on syntactic structure of sentences, as well as the machine learning techniques adopted for emotion classification.

- Chapter 5 presents techniques used for detecting emotion stimulus based on FrameNet's definition of emotions. In this chapter, we also explain the process of building a dataset, as well as our experiments on that data.
- In Chapter 6 we use the emotion stimulus models that were built in the previous chapter to deepen the analysis on the data used in Chapter 4. We also address the relation of implicit and explicit emotion expression with emotion stimulus.
- Chapter 7 concludes the thesis. In this chapter, we also discuss the obtained results and proposed future work.

Chapter 2

Theory of emotions and emotion analysis

Emotion is one of the most pervasive aspects of human experience. Even though it was more of an interest to researchers in social and behavioural sciences, it is becoming a multi-disciplinary research area, studied in computational linguistic as well. In this chapter, we will address emotions from a theoretical perspective. We explain different approaches in the theoretical study of emotions in psychology. We also discuss questions such as: what are emotions? how many emotions do we have? what is the cognitive structure of emotions? and what causes an emotion?

2.1 Emotion theory

Research on emotion in psychology reveals three major frameworks to study emotions: (1) the discrete emotion approach, (2) the dimensional approach, and (3) the cognitive appraisal approach (Watson and Spence, 2007).

The first method which is widely used is category-based. The *category approach* groups emotions into discrete categories based on their similarities. Researchers who adopt the discrete emotion approach use nominal, ordinal or interval scales to indicate emotions. A nominal scale contains terms that best describe the emotion experienced. However, there is no agreement on the list of nominal emotions. The theory most often applied in computational linguistic research is Ekman's (1992) theory. Ekman classifies emotions into six

classes which he calls basic emotions. Basic or primary emotions are considered fundamental because they represent survival-related patterns of responses to events in the world that have been selected for over the course of our evolutionary history. They are also defined as emotions with universally accepted distinctive facial expressions. According to Plutchik (1962), each fundamental, basic or primary emotion fulfills a specific adaptive role in helping organisms deal with key survival issues posed by the environment. These emotions are also considered to be fundamental because all other emotions are thought to be derived from them. Plutchik argues that there are eight basic and prototypical emotions, while Izard (1971) defines ten categories as basic emotions. There is obviously no agreement on the number of basic emotions, which is one of the drawbacks of this approach. Examples of some of most common emotions among the suggested categories are Sadness, Happiness, Anger, and Fear.

The ordinal scales define whether the respective emotion was experienced a little, somewhat or strongly, and the interval scale is an analog scale to indicate how much an emotion has been experienced. While the results obtained with this approach are highly plausible and easily interpretable, there are serious comparability problems, because different sets of emotion labels are used across different studies and the same as nominal scale, yet there is no agreement on the list of emotion labels to use. Finally, there is no attempt to determine what causes the similarities between emotions within each category; therefore, it is of limited use in explaining the cause of an emotion and when a particular emotion will be felt.

The other two methods are not as much used in computational linguistics, because it is less trivial to implement them computationally. However, the progress of theoretical emotion studies shows the shortcomings in the computational aspect of emotion studies. So we found it worthwhile to discuss them and later we will explain in more detail the potential of using them in NLP.

The second method, the *dimensional approach*, is mostly used in the humanities. Many dimensional theorists limit emotions to two valences (the degree of attraction or aversion that an individual feels toward a specific object or event) and arousal (body excitement) dimensions (Russell, 1983). This method is simple, straightforward and generally quite

reliable in obtaining self-reports of emotional feelings. However, one of the main shortcomings of dimensional approaches based on valence and arousal is that both dimensions are quite ambiguous. Thus, there is much overlap which makes the dimensional approach limited in its ability to distinguish between emotions of similar valence and arousal levels, such as the highly negative emotions of shame, fear and anger. For example, both very fearful and very angry persons would be characterized by a similar region of the two-dimensional space – negatively valenced and highly aroused. It is often not clear whether a valence judgement (pleasant or unpleasant) concerns the appraisal of the nature of the stimulus object, event or rather the feeling induced by it. Thus, additional dimensions such as goal conduciveness and coping potential have been advanced and illustrated as orthogonal axes. However, this method is still limited and misses some emotions such as *fear* (Scherer, 2005).

All that being said, to define emotions, we need a more comprehensive method which addresses the shortcomings of previous approaches. This brings us to the *cognitive appraisal approach*. Ortony et al. (1988) claim there is an essential and profound cognitive basis for emotions. Cognitive appraisal theories address three issues: what are the underlying characteristics inherent in events that are evaluated or appraised; what, if any, emotions are experienced as a result of this appraisal process; and what are the behavioural responses to the experienced emotions.

Appraisals differ from dimensions in that they are interpretations of characteristics of events that combine to cause particular emotions, while dimensions are inherent aspects of emotions themselves. For example, winning a sporting event may not always be interpreted positively, hence eliciting pride; players on a team that cheated to win may instead feel guilt. The winners on a team that cheated may interpret the outcome as undesirable or unfair; it is the interpretation of the event that causes the feeling of guilt. The cognitive appraisals approach is claimed to be more sophisticated than the dimensions approach in both purpose and scope (Ortony et al., 1988) and it is a promising avenue for studying emotions in computational linguistics. To back up this claim, in section 2.3 we will explain the cognitive structure of emotions in more detail. However, the next section will first discuss the definition of emotions from a theoretical perspective.

2.2 What are emotions?

At the time when affective and emotional phenomena are addressed inter-disciplinarily, the need for defining features of the different types of affective phenomena becomes essential (Ortony et al., 1988). However, the vague nature of emotions with overlaps between the categories makes it hard to have a universal and invariable definition of them. Even though the term is used very frequently, to the point of being extremely fashionable these days, the question *What is an emotion?* rarely generates the same answer. Here, we will mention some definitions by most well-known emotion theoreticians.

Scherer (2005) defines emotions as interrelated, synchronized changes in the states of all or most of the organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism. In defining emotions, we might also wonder *how emotions can be distinguished from other affective phenomena such as feelings, moods or attitudes.* Scherer refers to a few key emotion properties such as event focus, appraisal driven, etc. to address this question. Here the main characteristics of emotions are described in detail.

1. Event focus: One of the main features of emotions is that they are generally elicited by stimulus events. That means something happens to the organism to stimulate or trigger a response after having been evaluated for its significance.
2. Appraisal driven: Normally we do not get emotional about things that we do not care about. There are two kinds of appraisal, intrinsic and extrinsic. Intrinsic appraisal evaluates the feature of an object or person, independently of the current needs and goals of the appraiser, based on generic preferences. Extrinsic appraisal evaluates events and their consequences with respect to the needs, desires or goals of the appraiser.
3. Response synchronization: It defines whether emotions prepare appropriate responses to events; the response patterns must correspond to the appraisal analysis of the presumed implications of the event.
4. Rapidity of change: Emotional states are rarely steady states. Rather, emotion processes undergo constant modification allowing rapid readjustment to changing circum-

stances or evaluations.

5. Behavioural impact: This feature of emotions is due to their strong effect on emotion-consequent behaviour, which can have a strong impact on communication and social interaction.
6. Intensity: Intensity of the response patterns and the corresponding emotional experience is relatively high, suggesting that this may be an important design feature in distinguishing emotions from moods, for example.
7. Duration: In contrast to intensity, emotion duration must be relatively short in order not to tax the resources of the organism and to allow behavioural flexibility, whereas low-intensity moods can be maintained for much longer periods of time.

Nonetheless, some of these various sides of an emotion are unlikely to manifest themselves. While characteristics such as a *stimulus event*, *extrinsic object* or *person appraisals*, *emotion intensity* and *duration* can be identified in text, other features such as *intrinsic emotion appraisals* and *emotion rapidity of change* may not have any visible indicators in text.

Moreover, we are more interested in properties that cause a particular emotion to be felt, while *response synchronization*, *rapidity of change* and *behavioural impact* mainly address the characteristics of the emotions after they have been sensed. Finally, the two last properties, *intensity* and *duration*, are important because they are the main characteristics that differentiate emotions from mood and attitudes. However, that is beyond the scope of our work. Considering these emotion properties, we would like to differentiate emotions from feelings, attitudes and moods based on those characteristics.

Feelings are considered as one of emotion's components. In fact feelings integrate the central representation of appraisal-driven response organization in emotion.

Relatively long-lasting beliefs and predispositions towards specific objects or persons are generally called *attitudes*.

On the other hand, *moods* are considered as diffuse affect states. They may often emerge without apparent cause that could be clearly linked to an event or specific appraisals. Moods are generally of low intensity and show little response synchronization, but may last over hours or even days. Examples are being cheerful, gloomy, listless or depressed.

While Scherer presents an analysis of emotions from a broader theoretical and psychological perspective, we are interested in the properties of emotions that can be indicated by text. *Event-focus* and *appraisal driven* are the two main properties that can be found in explanations of emotions in text. We call the second one *agent* because it is an object or a person that causes a particular emotion.

FrameNet (Fillmore et al., 2003), a linguistic resource, defines emotion elements in the emotion frame as follows: An *experiencer* has a particular emotional *state*, which may be described in terms of a specific *stimulus* that provokes it, or a *topic* which categorizes the kind of stimulus. Rather than expressing the experiencer directly, it may have in its place a particular *event* (with participants who are Experiencers of the emotion) or an *expressor* (a body-part of gesture which would give an indication of the experiencer's state to an external observer). There can also be a *circumstance* under which the response occurs or a *reason* why the stimulus evokes the particular response in the Experiencer.

Fillmore et al. define six core frame elements for an emotion and six non-core elements (Table 2.1). Of these, the emotion stimulus is the person, event, or state of affairs that evokes the emotional response in the experiencer. These emotion properties are the common properties with FrameNet's definition of emotions stimulus in the emotion frame and the two properties of *event-focus* and *appraisal driven* defined by Scherer. We found them interesting and we will discuss them in more detail in Chapter 5.

Another definition of emotions is by Ortony et al. (1988) who define emotions as valenced (positive-negative) reactions to events, agents, or objects with their particular nature being determined by the way in which the eliciting situation is construed. That is the base of what is called the cognitive structure of emotions that is explained in more detail in the next section.

Core	Event	The occasion or happening in which Experiencers in a certain emotional state participate.
	Experiencer	The person or sentient entity who experiences or feels the emotions.
	Expressor	It marks expressions that indicate a body part, gesture or other expression of the Experiencer that reflects his or her emotional state.
	State	The abstract noun that describes a more lasting experience by the Experiencer.
	Stimulus	The person, event, or state of affairs that evokes the emotional response in the Experiencer.
	Topic	The general area in which the emotion occurs. It indicates a range of possible Stimulus.
Non-Core	Circumstances	The condition(s) under which the Stimulus evokes its response.
	Degree	The extent to which the Experiencer's emotion deviates from the norm for the emotion.
	Empathy-target	The individual or individuals with which the Experiencer identifies emotionally and thus shares their emotional response.
	Reason/ Explanation	The explanation for why the Stimulus evokes a certain emotional response.
	Manner	It is any description of the way in which the Experiencer experiences the Stimulus which is not covered by more specific frame elements. Manner may also describe a state of the Experiencer that affects the details of the emotional experience.
	Parameter	A domain in which the Experiencer experiences the Stimulus.

Table 2.1: The elements of the emotion frame in FrameNet.

2.3 Cognitive structure of emotions

Many independently developed, yet highly convergent, perspectives have been proposed regarding what underlying appraisals cause emotions (Frijda, 1987; Ortony et al., 1988; Scherer, 1988). Based on the cognitive structure of emotions, several dimensions, namely valence or pleasantness-unpleasantness, certainty, controllability, and agency or responsibility are found in most or all analyses. They are defined as follows.

Outcome desirability/pleasantness: Outcome desirability refers to the initial cognitive appraisal of whether the outcome of a situation/event is good or bad (positive or negative) with respect to personal well-being. There are two schools of thought regarding what underlies this central appraisal. Theorists such as Lazarus (1991) believe that this appraisal is an evaluation of a person-environment relationship, while others like Frijda (1987) believe it to be a motivational response to the relationship, with the motivation pertaining to a personal goal. Watson and Spence (2007) define outcome desirability as the combination of evaluations and motivational responses and it is the overall evaluation of how positive or negative (desirable/undesirable) a situation is relative to a personal benchmark. Outcome desirability has a very large effect on the attempts to categorize emotions. However, positive/negative appraisals alone are not particularly diagnostic when attempting to distinguish between specific emotions; other appraisals combine with outcome desirability to evoke specific emotions.

Agency: The next most influential event characteristic that determines an emotion is agency (Ortony et al., 1988; Craig A. Smith, 1985). The causal agent is who or what had control over the stimulus event. The agent could be oneself, someone else or a circumstance. Agency has been found to be more relevant in situations involving negative emotions than positive emotions and in response to failure rather than success. Agency is also related to controllability. While some theorists separate these two, we use a broad definition of agency by Watson and Spence (2007) that encompasses both the agent and their perceived control over the event being appraised. The appraised event may be referred to as self-caused, other-caused or circumstance-caused. Although outcome desirability accounts for a large share of explained variance by putting emotions into the two broad categories of desirable

and undesirable, it is agency-related appraisals that have the greatest effect on which specific emotions emerge within the desirable/undesirable emotion class. For example, Sadness and Disgust are both certain negative emotions; however, disgust is caused by situations that are perceived to be under the control of someone else while sadness is caused by situations that are believed to be controlled by circumstance (or are uncontrollable).

Certainty: Certainty represents the perceived likelihood of a particular event. Past events are certain ("I failed an exam"), future events are uncertain. How certain someone is about an outcome will influence how they feel about it. High levels of uncertainty are strongly associated with the emotions of hope and fear. Certainty is therefore an important appraisal for determining emotions felt, but is more relevant to decision making when anticipated decision outcomes are the focus.

Fairness: Fairness deals with how morally appropriate one perceives an event to be. Fairness appraisal can be strongly linked to agency in terms of its ability to explain people's behavioural responses to stimulus events. So, due to its high correlation with agency, we will not consider fairness as a separate feature.

Attention: Attention refers to the focusing of one's consciousness and receptivity, which may or may not be voluntary. Arguably, representing attention as an appraisal contradicts the essence of appraisal theory because one must attend to a stimulus prior to appraising it. Thus, attention precedes the appraisal process and it is not considered as one of the appraisal features.

Coping potential: Coping potential is associated with one's perceived ability to deal with, or to change, a situation. There is not enough supporting evidence to include coping as a necessary appraisal; instead, it is best thought of as a consequence of an emotion.

As a summary, based on the cognitive appraisals explained, we have the following propositions to explain differences between basic emotions based on the first three main appraisals.

1. Joy and distress (sadness):

    - Joy is being pleased about a desirable event.

    - Distress is being displeased about an undesirable event.

2. Hope and fear:

   - Hope is being pleased about the prospect of a desirable event or being pleased about an uncertain desirable event.

   - Fear is being displeased about the prospect of an undesirable event or being displeased about an uncertain undesirable event.

3. Anger and disgust (compound emotions):

   - Anger is being displeased about an undesirable event caused by someone else's action.

   - Disgust is disliking an unappealing object or situation.

4. Shame and guilt:

   - Shame and guilt are being displeased about one's own blameworthy action/undesirable event caused by self.

Here, we have addressed some emotions based on Ortony et al.'s (1988) theory, but there is still no consensus on the number of basic emotions among different theories. For example, Ortony et al. reject surprise as an emotion because they believe it can arise in the absence of valenced reaction, and it is rather a cognitive state having to do with unexpectedness. In the next section, we address this issue from different theorists' point of view.

2.4 Emotion categories

A number of theories have been proposed, each with its own set of basic emotions, though with significant overlap. But before mentioning the suggested basic emotions, we would like to discuss what it means for an emotion to be basic. Does it mean that it is important for the organism and its survival, or it is basic because it occurs more frequently relative to other emotions? Is it that it is common between all cultures or it is because it appears at a very early age? Ortony et al. (1988) argue that the latter can make emotions basic if one can show that emotions that develop later are somehow compounded out of basic emotions that

appear earlier. Although Ortony et al. avoid the notion of basic emotions, they treat some emotions as more basic than others, because those emotions have less complex specification and eliciting conditions. Therefore, they propose to have a hierarchical structure in which there are two basic kinds of affective reaction (positive and negative) at the top level and the other emotions will be at the next levels based on their degree of differentiation from the first level. We can thus think of the less differentiated emotions as being more basic. Ortony et al.'s theory is well justified, but they do not define a list of basic emotions. So, we cannot directly apply their method in the first part of our work.

On the other hand, many other theorists such as Ekman, Plutchik and Izard take a combinatorial view, meaning that the primary emotions can be mixed and will result in other emotions. Ekman (1992) argued for six basic emotions with universally accepted distinctive facial expressions: joy, sadness, anger, fear, disgust and surprise. Based on the relation to adaptive biological processes, Plutchik (1962) added trust and anticipation to Ekman's set. Izard (1971) defined ten basic emotions: anger, contempt, disgust, distress, fear, guilt, interest, joy, shame and surprise. He proved that each of the fundamental emotions can be represented in a unique pattern of facial activity or facial behaviour and is associated with a corresponding set of symbols or verbal labels. Additionally, he assumed that the fundamental emotions are innate and universal phenomena.

Among these three mentioned theories the six basic emotions defined by Ekman are common to all three theories. They are most frequently adopted by NLP researchers. Therefore in the first part of our work, we categorize sentences into six basic emotions defined by Ekman.

In the next chapter, we discuss emotions from the perspective of computational linguistics. We mention the main existing methodologies as well as the resources and data that are available.

2.5 Summary

In this chapter, we addressed theoretical aspect of emotions. We explained different takes on the study of emotions in psychology. We explained emotions and emotion causes from a psychological and cognitive perspective. We discussed what emotions are, and what are their main characteristics. We also studied the cognitive structure of emotions as well as different emotion categories that are proposed by various scientists. In the next chapter, we will address the related work in computational analysis of emotion and sentiment. We will explain the main computational-linguistic methods of emotion analysis as well as the available resources and data.

Chapter 3

Related work in computational analysis of emotion and sentiment

In this chapter, we explain the main computational-linguistic methods of emotion and sentiment analysis as well as important available resources and data.

Some of the emotion theorists mentioned in Chapter 2, such as Ortony et al. (1988), try to lay a foundation for a computationally tractable model of emotion. That shall be done by discussing how emotion characterizations contribute to some goals of Artificial Intelligence (AI). Applying emotions in the context of AI is not a question of "can computers feel" or "can computers have emotions", but AI endeavours to understand and reason about emotions or different aspect of emotions, Ortony et al. say. They believe that if computers are to be able to reason about emotions, we need a system of rules and representations about the elicitation of emotions.

Recognition, interpretation, and representation of affect have been investigated by researchers in the field of affective computing (Picard, 1997). She considers a wide range of modalities such as affect in speech, facial display, posture and physiological activity. Due to the huge amount of text data available on the Internet, such as blogs, emails and chats, which often are full of emotions, recently there has been a growing interest in automatic identification and extraction of sentiment, opinions and emotions in text. Besides, textual data take less space and are easier to transfer on the Web, so they have a high potential to

be used for sharing ideas, opinion and emotions.

Sentiment analysis and opinion mining constitute the field of study which analyzes peoples' opinions, sentiments, evaluations, attitudes and emotions occurring in written language (Wilson et al., 2005). Sentiment analysis has become one of the most active research areas in natural language processing (NLP), and a major application of data mining, Web mining and text mining. This wave of activity is connected with the rapid growth of the social media – in particular reviews, discussion fora, blogs, micro-blogs, Twitter and social networks – and the ensuing easy access to a mass of subjective and emotional data recorded in the digital form.

Another reason why sentiment analysis is such an active area of NLP research is that its applications have spread to multiple domains, from consumer product reviews, health care and financial services to social events and political elections (Liu, 2012). Here are some of the best known applications: classifying positive and negative movie reviews (Pang et al., 2002; Turney, 2002), opinion question-answering (Yu and Hatzivassiloglou, 2003; Stoyanov et al., 2005), summarizing customer reviews (Hu and Liu, 2004), and performing opinion piece recognition, which is a form of text categorization and genre detection Wiebe et al. (2004).

Early research has mainly focused on determining the presence of sentiment in the given text, and on determining its polarity – the positive or negative orientation, a finding which is useful but insufficient in most real-life applications. Practicality often dictates a more in-depth and fine-grained analysis. The analysis of sentiment must therefore go beyond differentiating positive from negative emotions toward giving a systematic account of the qualitative differences among individual emotions (Ortony et al., 1988). There has been progress in research on polarity and sentiment analysis, but not as much in automatic recognition of emotion in text.

In order to recognize and analyze affect in written text – seldom explicitly marked for emotions – NLP researchers have come up with a variety of techniques, including the use of machine learning, rule-based methods and the lexical approach (Neviarouskaya et al., 2011; Alm, 2008; Aman and Szpakowicz, 2007; Katz et al., 2007; Wilson et al., 2009; Strapparava

and Mihalcea, 2008; Chaumartin, 2007; Kozareva et al., 2007). In the following sections, we will explain each methodology in more detail.

Please note that relevant literature review is also present in later chapters.

3.1 Rule-based methods

Advanced rule-based linguistic approaches that target textual affect recognition at the sentence level are applied by Chaumartin (2007) and Neviarouskaya et al. (2011). Chaumartin developed a rule-based system relying on the lexicon from WordNet-Affect (Strapparava and Valitutti, 2004) and SentiWordNet (Esuli and Sebastiani, 2006), and applied it to affect sensing in news headlines. Neviarouskaya et al. used hard-coded rules to recognize affect in sentences of varying complexity, including simple, compound, complex (with complement and relative clauses), and complex-compound sentences.

Rule-based methods have not been as popular as the use of machine learning. They require a substantial manual rule-creation effort, an expensive process with weak guarantee of consistency and coverage, and are likely very task-dependent. The set of rules for an affect analysis task (Neviarouskaya et al., 2011) can differ drastically from what underlies other tasks such as rule-based part-of-speech taggers, discourse parsers, word sense disambiguation and machine translation (Ghazi et al., 2012).

3.1.1 Lexical semantics and keyword spotting

Olveres et al. (1998), in one of the first works in emotion detection in text, use a simple natural language parser for keyword spotting, phrase length measurement and emoticon identification. They apply some rules to construct emotion scores based on the parsed text and contextual information.

Another common take on emotion analysis is to use a lexicon with information about which words and phrases carry which emotion. The study of emotions in lexical semantics was the theme of a SemEval 2007 task (Strapparava and Mihalcea, 2007), carried out in an unsupervised setting (Strapparava and Mihalcea, 2008; Chaumartin, 2007; Kozareva et al.,

2007; Katz et al., 2007). The participants were encouraged to work with WordNet-Affect (Strapparava and Valitutti, 2004) and SentiWordNet (Esuli and Sebastiani, 2006) to classify news headline sentences into six emotions. Word-level analysis, however, will not suffice when affect is expressed by phrases which require complex phrase- and sentence-level analyses: words are interrelated and they mutually influence their affect-related interpretation. On the other hand, words can have more than one sense, and they can only be disambiguated in context. Let us highlight several issues.

- An emotion word may have different orientation in different application domains – like the word *cool* when used to describe a car or someone's demeanour.
- A sentence containing an emotion word may express no emotion. Consider the vacuous polite *afraid* in "I'm afraid it's going to rain."
- Many sentences without explicit emotion words can also convey emotions. Consider the rather negative emotion in "My life fell apart when my mum died.".
- Words are interrelated and they mutually influence their affect-related interpretation. However, emotion word lists do not take into account negation, syntactic relations and semantic dependencies in the sentence.

As a result, the emotion conveyed by a word in a sentence can differ drastically from the emotion of the word on its own. To admit that, also according to a linguistic survey done by Pennebaker and Francis (2001), across all of the studies described by them people usually use fewer emotion words, even though they may express affective content significantly. This indicates that affective lexicons might not be sufficient to recognize affective information in texts, and raises the suspicion that keyword spotting methods might not perform well for this objective either.

3.2 Statistical methods

The existence of annotated corpora with rich information about opinions and emotions allowed the development and evaluation of NLP systems which exploit such information. In

particular, statistical NLP and machine learning approaches have become the methods of choice for constructing a wide variety of practical NLP applications (Wiebe et al., 2005).

3.2.1 Supervised methods

Such methods have been previously applied to corpus-based features, mainly unigrams, combined with lexical features (Pang et al., 2002; Alm et al., 2005; Katz et al., 2007; Aman and Szpakowicz, 2007; Wilson et al., 2009). Pang et al.'s (2002) paper was among the first to describe a classification of movie reviews into positive and negative. The authors showed that using unigrams as features in classification performed quite well with Naïve Bayes and with Support Vector Machines (SVM). In general, these two supervised machine learning algorithms have long been the methods of choice for sentiment recognition in text. SVM has been shown to outperform Naïve Bayes consistently in our particular task (Aman and Szpakowicz, 2007; Ghazi et al., 2010a). The weakness of those methods was that they disregarded negation, syntactic relations and semantic dependencies. Later, there were more sophisticated methods that apply syntactic relations that also consider negation (Ghazi et al., 2012). And recently, with growing research on semantic parsers, common-sense knowledge bases and semantic relations, applying deeper emotion analysis methods is inevitable. This approach, applied as part of our work, is discussed in Chapter 5.

There are a few drawbacks of such supervised machine-learning methods: they require large annotated corpora for meaningful statistics and good performance, processing may take time, and the manual annotation effort is inevitably high.

3.2.2 Unsupervised methods

Unsupervised statistical machine learning methods such as clustering try to find hidden structure in unlabelled data so they group items that seem to fall naturally together (Witten and Frank, 2005). We are not aware of any previous work that would apply clustering methods to emotion analysis. The main reason could be the high number of emotion classes that make it a difficult task for clustering methods.

To the best of our knowledge, previous methods that are known as unsupervised included

happiness	sadness	anger	disgust	surprise	fear	total
398	201	252	53	71	141	1116

Table 3.1: Distribution of labels in the WordNet-Affect lexicon.

some sort of supervision by either using lexicons or linguistic specification of the text to build rules manually. The first group uses a lexicon with information about the emotion carried by each word (Strapparava and Mihalcea, 2008; Chaumartin, 2007; Kozareva et al., 2007; Katz et al., 2007) that were explained in section 3.1.1. There are also methods that rely on linguistic rules (Neviarouskaya et al., 2011) that were explained in section 3.1.

In the next section, we explain the available resources, such as datasets, lexicons and common sense knowledge bases that are used or can be applied in the above-mentioned tasks.

3.3 Available resources

Supervised statistical methods typically require training data and test data, manually annotated with respect to each language-processing task to be learned. Due to the contributions of many researchers, several subjectivity and emotion datasets and lexicons have been built. Most of them are in the public domain. In this section, we explain several datasets and lexicons as well as common-sense knowledge bases that are available.

3.3.1 Lexicons

WordNet-Affect. It is a set of words which indicate the presence of a particular emotion. WordNet-Affect (Strapparava and Valitutti, 2004) contains six lists of words corresponding to the six basic emotion categories. It is the result of assigning a variety of affect labels to synsets in WordNet, which was built automatically. Table 3.1 shows the distribution of words in WordNet-Affect.

The Prior-Polarity Lexicon. The prior-polarity subjectivity lexicon (Wilson et al., 2009) contains over 8000 subjectivity clues collected from a number of sources. To create this

Neutral	Negative	Positive	Both
6.9%	59.7%	31.1%	0.3%

Table 3.2: Distribution of labels in the Prior-Polarity Lexicon.

joy	sadness	anger	disgust	surprise	fear	trust	antici- -pation	pos.	neg.
689	1191	1247	1058	534	1476	1231	839	2312	3324

Table 3.3: Distribution of labels in the NRC-Emotion lexicon.

lexicon, the authors began with the list of subjectivity clues extracted by Riloff (2003). The list was expanded using a dictionary and a thesaurus, and adding positive and negative word lists from the General Inquirer (`www.wjh.harvard.edu/~inquirer/`). Words are grouped into strong subjective and weak subjective clues; Table 3.2 presents the distribution of their polarity.

The NRC Emotion lexicon. This lexicon (Mohammad and Turney, 2010) is a list of words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The manual annotation was constructed by crowd-sourcing, using Amazon's Mechanical Turk. Each word-sense pair had at least three annotators, and the lexicon was created by taking the union of emotions associated with all the senses of a word. Table 3.3 displays the distribution of the emotion and polarity of words. The main difference between this lexicon and WordNet-Affect is that each word in this lexicon can belong to multiple emotion classes as well as both sentiment classes. As a result, while there are 6195 emotion words and their emotion class pairs for six emotion classes in the lexicon, it only has 3462 emotion word types.

The Intensifier Lexicon. It is a list of 112 modifiers (adverbs) (Neviarouskaya et al., 2010). Two annotators gave coefficients for intensity degree – strengthening or weakening, from 0.0 to 2.0 – and the result was averaged.

hp	sd	ag	dg	sr	fr	ne	total
536	173	179	172	115	115	2800	4090

Table 3.4: Distribution of labels in Aman's dataset; the labels are happiness, sadness, anger, disgust, surprise, fear, no emotion.

3.3.2 Datasets

Aman's Dataset. In this dataset, each sentence is tagged by its dominant emotion, as mixed-emotion if the emotion conveyed cannot be attributed to any basic category, or as no-emotion if it does not include any emotion (Aman and Szpakowicz, 2007). The annotation is based on Ekman's six emotions at the sentence level. This dataset was created from blog data collected from the Web. First, a set of seed words for the six emotion categories were defined. Next, the seed words for each category were fed to the blog search engine BlogPulse, and blog posts containing one or more of those words were retrieved. As a result total of 173 blog posts were collected. For each sentence in the corpus, two annotators were required to label the predominant emotion category in the sentence, the intensity level of the emotion, as well as the emotion indicators in the sentence. The sentences with mixed emotions were removed, so the rest of the dataset contains 4090 annotated sentences, 68% of which were marked as non-emotional. Table 3.4 shows the details of Aman's dataset.

Alm's Dataset. It is a sentence-annotated corpus for large-scale exploration of affect in classical tales (Alm, 2008). A subset involves sentences marked by high agreement. The affects considered are the same as Ekman's six emotions, except that the *surprise* class is subdivided into positive and negative surprise. A version of the dataset merged *angry* and *disgusted* instances and combined the positive and negative *surprise* classes.

The ISEAR Dataset. ISEAR (International Survey on Emotion Antecedents and Reactions) is a dataset that was collected over a period of many years during the 1990s. A large group of psychologists all over the world collected data in the ISEAR project, directed by Wallbott and Scherer (1986). The participants were asked to report situations in which they had experienced all of seven fundamental emotions: joy, fear, anger, sadness, disgust,

shame and guilt. In each case, the questions covered the way the participants had appraised the situation and how they reacted. The final data set thus contained reports on seven emotions, each by close to 3000 respondents in 37 countries on five continents. The dataset contains 7,666 such statements, which include 18,146 sentences and 449,060 running words.

Neviarouskaya's Dataset. Neviarouskaya et al. (2010) extracted 700 sentences from the Weblog Data Collection in order to evaluate their rule-based emotion recognition algorithm. Three independent annotators labelled the sentences with one of nine emotion categories (Anger, Disgust, Fear, Guilt, Interest, Joy, Sadness, Shame and Surprise), or neutral, and a corresponding intensity value. Additionally, they interpreted these fine-grained annotations using three polarity-based categories (positive emotion, negative emotion, and neutral) by merging Interest, Joy, and Surprise in a positive emotion category, and Anger, Disgust, Fear, Guilt, Sadness, and Shame in a negative emotion category. Based on the annotators' agreement, they have a set of 656 sentences that at least two annotators agreed on, among which all three annotators completely agreed on 249 sentences.

While accumulating affect database entries, they also collected 364 emoticons, both of American and Japanese style, and the 337 most popular acronyms and abbreviations, both emotional and non-emotional. Moreover, they included in the database 112 modifiers which influence the strength of related words and phrases in a sentence.

Text Affect. This is a dataset that was available for SemEval 2007 (Strapparava and Mihalcea, 2007). It is annotated based on the emotion and valence of news headlines separately. For the SemEval 2007 task, two datasets were available; a development dataset consisting of 250 annotated headlines, and a test data set which is 1000 annotated news headlines. In this dataset, each sentence is represented by six numerical values that each one indicates the strength of each emotion. The interval for the emotion annotations was set to [0, 100], where 0 means the emotion is missing from the given headline and 100 represents maximum emotional load. The interval for the valence annotation was set to [-100,100], where 0 represents a neutral headline, -100 represents a highly negative and 100 corresponds to highly positive headline.

3.3.3 Knowledge-base resources

There are some well-known and widely used knowledge bases in NLP, namely, lexical knowledge bases such as WordNet (Fellbaum, 1998) and FrameNet (Fillmore et al., 2003) as well as common-sense knowledge bases such as ConceptNet (Liu and Singh, 2004), and SenseNet (Chen et al., 2006).

Here we explain the FrameNet knowledge base because it was used in our work. We also explain two emotion common-sense knowledge bases that were created using WordNet: ConceptNet and SenseNet.

Lexical knowledge bases

A lexical knowledge base is a general repository of knowledge about lexical concepts and their relationships. It is a repository of computational information about concepts that contains information derived from machine-readable dictionaries, the full text of reference books, the results of statistical analyses of text usages, and data manually obtained from human world knowledge (Amsler, 1984).

FrameNet FrameNet is a lexical database of English that is both human- and machine-readable, based on annotating examples of how words are used in actual texts. It is a dictionary of more than 10,000 word senses, with 170,000 manually annotated sentences that show the meaning and usage of words.

FrameNet is based on a theory of meaning called Frame Semantics, deriving from the work of Fillmore et al. (2003). The basic idea is that the meanings of most words can best be understood on the basis of a semantic frame: a description of a type of event, relation, or entity and the participants in it. For example, the concept of cooking typically involves a person doing the cooking (Cook), the food that is to be cooked (Food), something to hold the food while cooking (Container) and a source of heat (Heating-instrument). In FrameNet, this is represented as a frame called Apply-heat, and the Cook, Food, Heating-instrument and Container are called frame elements (FEs). Words that evoke this frame, such as fry, bake, boil, and broil, are called lexical units (LUs) of the Apply-heat frame.

FrameNet has more than 1,000 semantic frames that are linked by a system of frame relations, which relate more general frames to more specific ones and provide a basis for reasoning about events and intentional actions.

Common-sense knowledge bases

Common-sense knowledge bases are databases containing world knowledge including millions of basic facts and information that an ordinary person is expected to know (Liu and Singh, 2004). They are useful to capture deep semantic information effectively and construct a large-scale knowledge base efficiently.

SenticNet Cambria et al. (2010) believe that use of common-sense computing can significantly enhance computers' emotional intelligence. They introduce a method, which they call Sentic Computing, in which they use a common-sense repository to infer affective states. For that purpose, they create a collection of commonly used polarity concepts called SenticNet.

SenticNet is a publicly available resource for opinion mining that contains around 5700 polarity concepts. The authors combine the domain-general knowledge in ConceptNet with the affective knowledge contained in WordNet-Affect. Then they assign values to key affective concepts and apply an operation that spreads their values across the ConceptNet graph.

To evaluate SenticNet, the authors compare it with SentiWordNet's capacity of detecting opinion polarity over a collection of 2,000 patient opinions, of which 57% are labelled as negative, 32% as positive and the rest as neutral. They claim that they can identify positive opinions with higher F-measure value of 67% versus 49% in SentiWordNet. However, they do not mention their results for negative and neutral opinions. This result for a task with baseline accuracy of 57% (the percentage of the largest class, negative opinions) is not very promising.

EmotiNet The EmotiNet knowledge base (Balahur et al., 2011c) is a resource for the detection of emotion from text based on common-sense knowledge of concepts, their interaction and their affective consequence. The core of the resource is built from a set of self-reported

affective situations and extended with external ontologies such as ReiAction[1] and family relations ontology. Due to lack of coverage, they also extended the EmotiNet knowledge base by adding new actions similar to the ones included in the core using existing knowledge bases – VerbOcean (Chklovski and Pantel, 2004), core WordNet and WordNet-Affect to extract similar verbs that are in the same synset. The general idea behind their work is to model situations as chains of actions and their corresponding emotion using an ontological representation. However, the evaluation of their knowledge based on 859 examples of ISEAR dataset (Balahur et al., 2011a), shows lack of coverage and consistency in their methodology.

3.4 Summary

In this chapter, we explained the main computational-linguistic methods of emotion and sentiment analysis, as well as the available resources and data. We addressed the related work in computational analysis of emotion and sentiment. We discussed methodologies such as rule-based and statistical methods as well as existing data sources, namely, data sets, lexicons and knowledge-base resources. Considering these methodologies and data sources, in the next chapter we will address our first set of experiments on recognizing expressions of emotions in text.

[1] http://www.csee.umbc.edu/~lkagal1/rei/ontologies/ReiAction.owl

Chapter 4

Recognition of emotions in text

Note. Parts of this chapter have been published earlier in (Ghazi et al., 2010a,b, 2012, 2014).

4.1 Motivation

In this thesis, we deal with assigning fine-grained emotion classes to sentences in text. On the face of it, our task is strongly related to polarity analysis, but categorization into distinct emotion classes is more difficult. That is not only because emotion recognition in general requires deeper insights but also because there are similarities between emotion which make clean classification a challenge. Particularly notable in this regard are anger and disgust, two emotion classes which even human annotators often find hard to distinguish (Aman and Szpakowicz, 2007).

As explained in Section 2.4, a number of theories have been proposed, each with its own set of basic emotions, though with significant overlap. Ekman (1992) argued for six basic emotions with universally accepted distinctive facial expressions: joy, sadness, anger, fear, disgust and surprise. Plutchik (1962) added trust and anticipation to Ekman's set. Izard (1971) defined ten basic emotions: anger, contempt, disgust, distress, fear, guilt, interest, joy, shame and surprise. More recently, Parrot (2001) proposed six primary emotions: love, joy, surprise, anger, sadness and fear.

In this chapter, we categorize sentences into six basic emotions defined by Ekman (1992), a set most frequently adopted by NLP researchers. There also may, naturally, be no emotion in a sentence; that is tagged as neutral/no-emotion.

Sentiment can be sought in text units at three levels (Wilson et al., 2009). While much work in sentiment analysis has targeted documents, more has been done with sentences. The third level of sentiment analysis is phrase-level, which to the best of our knowledge has not been used in emotion recognition research due to its complication and lack of data.

Document-level analysis assumes that each document expresses a single sentiment. While it is likely for a document to express a single polarity, it is quite rare to find a document with only one emotion. On the other hand, one emotion is quite feasible in a sentence, though often more than one emotion is conveyed. For example, the sentence "Although the service is not that great, I still love this restaurant." has a positive tone, but it is not entirely positive. Our work focusses on a dataset from which sentences with mixed emotions have been removed, so we can assume that each sentence only indicates one emotion.

As noted in chapter 3, to recognize emotions in text, NLP researchers have come up with a variety of techniques, including the use of machine learning, rule-based methods and the lexical approach (Neviarouskaya et al., 2011; Alm et al., 2005; Alm, 2008; Aman and Szpakowicz, 2007; Katz et al., 2007; Wilson et al., 2009; Strapparava and Mihalcea, 2008; Chaumartin, 2007; Kozareva et al., 2007).

In this chapter, we discuss the application of various methods to the recognition of emotions in text. The methodologies range from shallow methods working directly on the surface features of text, namely lexical and corpus-based approaches explained in Section 4.2, to deeper methods using sophisticated linguistic analysis, discussed in Section 4.3.

4.2 Lexical and corpus-based emotion recognition

As mentioned before, one of the tasks at SemEval 2007 (Strapparava and Mihalcea, 2007) was carried out in an unsupervised setting; the emphasis was on the study of emotion in lexical semantics (Strapparava and Mihalcea, 2008; Chaumartin, 2007; Kozareva et al., 2007; Katz

et al., 2007). The participants in the SemEval 2007 workshop took a linguistic approach, using enriched lexical resources such as SentiWordNet and WordNet-Affect. They also used statistics gathered from Web search engines (Kozareva et al., 2007) based on the hypothesis that groups of words which co-occur with a given emotion word are highly likely to express the same emotion.

Strapparava and Mihalcea (2008) exploited the co-occurrence of words in the text with the words which have explicit affective meaning. They used the WordNet-Affect list as the direct affective words, and implemented a variation of Latent Semantic Analysis to yield a vector space model which allows for a homogeneous representation of words. These tasks were tested on the emotion classification of news headlines extracted from news Web sites.

Aman and Szpakowicz (2007; 2008) have shown that the best results on a dataset annotated with emotions are achieved by a combination of corpus-based unigram features and lexically inspired features using Support Vector Machine (SVM). In general, two supervised machine learning algorithms, SVM and Naïve Bayes, have long been the methods of choice for sentiment recognition at the text level. SVM has been shown to outperform NaïveBayes consistently in emotion classification (Pang et al., 2002; Aman and Szpakowicz, 2007; Ghazi et al., 2010a).

4.2.1 Data representation

A few datasets were discussed in Section 3.3.2. Alm's dataset is a sentence-annotated corpus exploring affect in children's story tales. ISEAR is a dataset that was collected by psychologists all over the world, who asked participants to report situations in which they had experienced all of defined seven major emotions. Neviarouskaya's dataset extracted 700 sentences from the Weblog Data Collection in order to evaluate their rule-based emotion recognition algorithm. Text Affect is a dataset made available for SemEval 2007, annotated based on the emotion and valence of news headlines. Finally, there is Aman's dataset, used in our experiments.

The primary consideration behind Aman's dataset was that the data should be rich in emotion expressions and the data should contain ample instances of all the emotion categories

hp	sd	ag	dg	sr	fr	ne	total
536	173	179	172	115	115	800	2090

Table 4.1: Distribution of labels in Aman's modified dataset; the labels are happiness, sadness, anger, disgust, surprise, fear, no emotion.

considered in the author's research (Aman and Szpakowicz, 2007). Such data can be found in personal texts such as diaries, emails, and blogs. While some personal texts might not be publicly available, blogs are online personal journals containing owner's reflections and comments that cover everyday topics as well as reflections on personal relationships and global issues. Blogs make good candidate for emotion study, as they are likely to be rich in emotion content. Also blogs offer real-world examples of emotion expression in text, unlike Alm's dataset that is built from children's story tales, and ISEAR dataset that was collected from participants of the project.

The applications of our work deal with real-world text – often containing noise, such as misspellings and slang, for instance, in affective interfaces and communication systems. Another consideration in selecting Aman's dataset was that, unlike Text Affect dataset which was built on news headlines, such text does not conform to the style of any particular genre per se, thus offering a variety in writing styles, choice and combination of words, as well as topics. Thus, the methods learned for this dataset would be general and widely applicable. Last but not least, we found it a good choice because it is a decent-size dataset.

However, the original dataset is highly imbalanced, with no-emotion sentences by far the largest class and with merely 3% in the fear and surprise classes. This prompted us to remove 2000 of the no-emotion sentences. This brought the number of no-emotion sentences down to 38% of all the sentences, and thus reduced the imbalance. The details of the modified Aman's dataset are shown in Table 4.1.

In this section, we use three sets of features. The first set is corpus-based, so we need no other external resources; the other two sets are lexical features. The lexically motivated approach requires lexical-semantic resources. In our experiments, we use three lexicons: the Prior-Polarity subjectivity lexicon (Wilson et al., 2009), WordNet-Affect (Strapparava

and Valitutti, 2004), and a list of words derived from Roget's Thesaurus (Jarmasz and Szpakowicz, 2003).

We evaluate the performance of different lexicons by comparing them with unigram features. We also find the most appropriate set of features among them.

Bag-of-Words

The first experiment is a corpus-based classification which uses unigrams. The unigram models have been widely used in text classification, and shown to provide good results in sentiment classification tasks (Kennedy and Inkpen, 2006; Aman and Szpakowicz, 2007; Ghazi et al., 2010a). In this experiment, we use unigrams from the corpus, selected using feature selection methods from Weka (Hall et al., 2009).

Lexical features

The second experiment involves classification with features derived from the polarity lexicon. The features here are the tokens that are common between the Prior-Polarity lexicon and the chosen dataset – 796 tokens in total.

In the last experiment, we use a combination of the lists of emotion words from Roget's Thesaurus (collected by Aman and Szpakowicz (2008)) and WordNet-Affect. There are 1264 tokens in common between the combined lexicons and dataset's tokens.

4.2.2 Sentence-level emotion classification

Text classification quite often deals with high dimensionality of the feature space. Many learning algorithms do not scale to a high-dimensional feature space. SVM has been shown to give good performance in text classification experiments because it scales well to the large numbers of features (Yang and Liu, 1999; Pang et al., 2002; Aman and Szpakowicz, 2007). We also experimented with other classifiers: NaïveBayes and Decision Tree. As expected, their results on Bag-of-Words features were substantially lower than SVM, with around 23% and 25% lower accuracy respectively. Consequently, we use SVM as a machine-learning algorithm to be applied to the chosen features.

4.2.3 Experiments and evaluation

As a result of the emotion classification experiments using SVM – the SMO algorithm in Weka (Hall et al., 2009)[1] – and setting *10-fold cross validation* as the testing option, we get the accuracy of 65.55% for the first set of features, 62.67% for the second set and 57.30% for the last feature set.

Based on the accuracy of the three experiments, unigram features outperform the other two types. Even though the unigram model is substantially better than the third one, its difference with the second feature set merits a discussion. Here, we would like to point out parameters – other than the accuracy of the results – which should be considered in choosing the list of features. The number of features is one of the main parameters. It is important because having fewer features reduces training times, makes a model more general-purpose by reducing overfitting, and it is simpler to understand and explain the model; therefore we are not interested in having better results by adding more features. What is more interesting is to find a list of fewer but more meaningful features which could contribute more to learning. In our experiments, the number of polarity features is much smaller than the unigrams. Also, by inspecting the features manually, we noticed that they appear to be more meaningful. For example, among the unigram features, we have proper nouns such as names of people and countries. It is also possible to have misspelled tokens in unigrams, while the Prior-Polarity lexicon features are well-defined words usually considered as polar. Besides, we used domain- and corpus-independent lexicons, which result in having domain- and corpus-independent lexical features.

Nonetheless, the lexical approach has some shortcomings, of which two are the most obvious: the emotion conveyed by a word in a sentence can differ drastically from the emotion of the word on its own; a sentence containing an emotion word may not express any emotion.

In the next section, we explain a method that takes the context of the emotion word into account so it detects the emotion of the sentence considering the context rather than only

[1]We use the polynomial kernel with normalized training data settings. It is the default setting which works well in our experiments.

the emotion word.

4.3 Syntax-based emotion recognition

A set of words labelled with their prior emotion is an obvious place to start on the automatic discovery of the emotion of a sentence, but it is clear that context must also be considered. It may be that no simple function of the labels on the individual words captures the overall emotion of the sentence; words are interrelated and they mutually influence their affect-related interpretation. It happens quite often that a word which invokes emotion is used in a neutral sentence, or that a sentence with no emotion word carries an emotion. This could also happen among different emotion classes.

In this section, we focus on disambiguating the contextual emotion of words, taking a sentence as the context. Our method combines several way of tackling the problem. First, we find keywords listed in WordNet-Affect and select the sentences which include emotion words from that lexicon. Next, we study the syntactic structure and semantic relations in the text surrounding the emotion word by presenting a set of features which enable us to take the contextual emotion of a word into account. We explore features important in emotion recognition, and we consider their effect on the emotion expressed by the sentence. For that we assess the performance of different features sets across multiple classification methods. We show the features and a promising learning method which significantly outperforms two reasonable baselines. We group our features by the similarity of their nature. That is why another facet of our evaluation is to consider each group of the features separately and investigate how well they contribute to the result. The experiments show that all features contribute to the result, but it is the combination of all the features that gives the best performance. Finally, we use machine learning to classify the sentence, represented by the chosen features, based on the contextual emotion of emotion words which occur in that sentence.

We evaluated our results by comparing our method applied to our set of features with Support Vector Machine (SVM) applied to Bag-of-Words, which was found to give the best

performance among supervised methods (Yang and Liu, 1999; Pang et al., 2002; Aman and Szpakowicz, 2007; Ghazi et al., 2010a). We show that our method is promising and that it outperforms both a system which works only with prior emotions of words, ignoring context, and a system which applies SVM to Bag-of-Words. We also evaluate the usefulness of our features for this task by comparing the performance of each group of features with the baseline. We find that all the features help the performance of the classifier somewhat, and that certain features – such as part-of-speech and negation – improve the results significantly.

4.3.1 Data representation

Training and test data

The main consideration in the selection of data for emotion classification task is that the data should be rich in emotion expressions. That is why we chose for our experiments the corpus of blog sentences annotated with emotion labels, produced by Aman and Szpakowicz (2007). As explained in Section 3.3.2, each sentence is tagged by its dominant emotion, or as no-emotion if it does not include any emotion. Table 3.4 shows the details of the chosen dataset.

For our particular task, we cannot use the whole dataset. Our main consideration is to classify a sentence based on the contextual emotion of the words (known as emotion-bearing in the lexicon). We use for this purpose two of the emotion lexicons described in Section 3.3.1, namely WordNet-Affect and the NRC Emotion lexicon; we use them separately to choose only the sentences which contain at least one emotion word according to the lexicon.

In the NRC Emotion lexicon, however, each word can belong to multiple emotion classes. This actually happens quite often among negative emotion classes. For example, the word "hate" belongs to four negative emotion classes: anger, disgust, fear and sadness. As a result, if we choose sentences which contain at least one emotion word according to the NRC Emotion lexicon, we will have 1010 sentences. Only 361 of those sentences will have one emotion word from exactly one emotion class (201 of them are *Happy* words). Part of our experiments are based on sentences with only one emotion word, so this makes the NRC Emotion lexicon not suitable for our task, particularly for negative emotion classes. Even

	hp	sd	ag	dg	sr	fr	ne	total
part 1	196	64	64	63	36	52	150	625
part 2	51	18	22	18	9	14	26	158
part 1+part 2	247	82	86	81	45	66	176	783

Table 4.2: Distribution of labels in the portions of Aman's dataset used in our experiments, which we call part 1 (one emotion word), part 2 (more emotion words), and part 1+part 2; the labels are happiness, sadness, anger, disgust, surprise, fear, no emotion.

so, we use it for comparison purposes and to evaluate our features.

In effect, to prepare the dataset for our task, we use the WordNet-Affect lexicon. In the dataset, we only choose sentences which contain at least one emotion word according to WordNet-Affect. This gives us 783 sentences, 625 of which contain only one emotion word. The details appear in Table 4.2.

Features

As in any other application of supervised machine learning, the key in sentiment classification is the engineering of a set of effective features. Terms and their frequency are the most common features in traditional text classification, and they have been shown highly effective for sentiment classification as well (Liu, 2012). They cannot, however, detect relations and semantic dependencies. We use those most common features only for comparison purposes.

The features used in our experiments were motivated both by the literature (Wilson et al., 2009; Choi et al., 2005) and by the exploration of contextual emotion of words in the annotated data. The counts of the feature values were based on the occurrences in the sentence of emotion words from the lexicons. For ease of description, we group the features into four distinct sets: emotion-word features, part-of-speech (POS) features, sentence features and dependency-tree features. We will describe some of the features as multi-valued or categorical features. In practice, however, all features have been binarized.

Emotion-word features. This set of features is based on the emotion words themselves. Although the prior emotion and polarity of the emotion word in the sentence are insufficient

for emotion analysis, they *are* useful for simple sentences such as "I am happy", where the prior and contextual emotion of the sentence are exactly the same and there are no emotion shifters and influencers. These features have also been widely adopted and found useful in detecting sentiment and emotions:

1. the emotion of a word according to WordNet-Affect (Strapparava and Valitutti, 2004).
2. the polarity of a word according to the prior-polarity lexicon (Wilson et al., 2009).
3. the presence of a word in a small list of modifiers (Neviarouskaya et al., 2010).

POS features. Words in certain grammatical classes have been shown to be more effective in recognizing the emotion of a sentence. For example, adjectives and adverbs are important indicators of emotions (Mohammad and Turney, 2010). We use the Stanford tagger (Toutanova et al., 2003), which gives every word in a sentence a Penn Treebank tag.

The WordNet-Affect emotion lexicon lists the POS of the words along with their corresponding emotion. That is why we use the POS of the emotion word itself, both according to the emotion lexicon and to the Stanford tagger.

The POS of neighbouring words in the same sentence are an important factor. We chose a [-2, +2] window, as suggested by the literature (Choi et al., 2005).

Sentence features. At present, we only have one feature in this group: the number of words in the sentence.

Dependency-tree features. These features capture the various types of relationships which involve the emotion word. The feature values come from a dependency parse tree of the sentence, obtained by parsing the sentence and then converting it to dependencies. We work with the Stanford parser (Marneffe et al., 2006). The dependencies are all binary relations: a grammatical relation holds between a governor (head) and a dependent (modifier). Figure 4.1 shows an example of a dependency parse tree and its dependencies.

According to Mohammad and Turney (2010), adverbs and adjectives are some of the most emotion-inspiring terms. This is not surprising given that adjectives qualify nouns and adverbs qualify verbs. That is why we have decided to consider only a handful of the 52 different dependencies. In order to keep the number of features small, we only chose the negation, adverb and adjective modifier dependencies.

Your query

It was the best summer I have ever experienced.

Tagging

It/PRP was/VBD the/DT best/JJS summer/NN
I/PRP have/VBP ever/RB experienced/VBN ./.

Parse

```
(ROOT
  (S
    (NP (PRP It))
    (VP (VBD was)
        (NP
          (NP (DT the) (JJS best) (NN summer))
          (SBAR
            (S
              (NP (PRP I))
              (VP (VBP have)
                (ADVP (RB ever))
                (VP (VBN experienced)))))))
    (. .)))
```

Typed dependencies

nsubj (summer-5, It-1)
cop (summer-5, was-2)
det (summer-5, the-3)
amod (summer-5, best-4)
root (ROOT-0, summer-5)
nsubj (experienced-9, I-6)
aux (experienced-9, have-7)
advmod (experienced-9, ever-8)
rcmod (summer-5, experienced-9)

Figure 4.1: A graphical representation of the Stanford parser's POS tags, parse tree and dependencies for the sentence "It was the best summer I have ever experienced."

After parsing the sentence and getting the dependencies, we count three dependency-tree Boolean features for the emotion word.

1. If the word is in a "neg" dependency (negation modifier): true when there is a negation word which modifies the emotion word.
2. If the word is in a "amod" dependency (adjectival modifier): true when the emotion word is (i) a noun modified by an adjective, or (ii) an adjective modifying a noun.
3. If the word is in a "advmod" dependency (adverbial modifier): true when the emotion word (i) is a non-clausal adverb or the head of an adverbial phrase which serves to modify the meaning of a word, or (ii) has been modified by an adverb.

We also have several modification features based on the dependency tree. These Boolean features capture different types of relationships involving the emotion word. We list the feature name and the condition on the emotion word w which makes the feature true.

1. Modifies-positive: w modifies a positive word from the prior-polarity lexicon.
2. Modifies-negative: w modifies a negative word from the prior-polarity lexicon.
3. Modified-by-positive: w is the head of the dependency, which is modified by a positive word from the prior-polarity lexicon.
4. Modified-by-negative: w is the head of the dependency, which is modified by a negative word from the prior-polarity lexicon.
5. Modifies-intensifier-strengthen: w modifies a strengthening intensifier from the intensifier lexicon.
6. Modifies-intensifier-weaken: w modifies a weakening intensifier from the intensifier lexicon.
7. Modified-by-intensifier-strengthen: w is the head of the dependency, which is modified by a strengthening intensifier from the intensifier lexicon.
8. Modified-by-intensifier-weaken: w is the head of the dependency, which is modified by a weakening intensifier from the intensifiers lexicon.

4.3.2 Sentence-level emotion classification

Our main goal is to evaluate the usefulness of the features described in Section 4.3.1 in recognizing the contextual emotion of an emotion word in a sentence. To evaluate those features, we investigate their performance, both together and separately. In the experiments, we use the Aman's emotion dataset presented in Section 3.3.2. Next, we represent the data with the features presented in Section 4.3.1. Those features, however, were defined for each emotion word based on its context; we will therefore proceed differently for sentences with one emotion word and sentences with more than one emotion word.

- In sentences with one emotion word, we assume that the contextual emotion of the emotion word is the same as the emotion assigned to the sentence by the human annotators; therefore all the 625 sentences with one emotion word are represented with the set of features presented in Section 4.3.1 and the sentence's emotion will be considered as the contextual emotion of these words.
- For sentences with more than one emotion word, the emotion of the sentence depends on all emotion words and their syntactic and semantic relations. We have 158 sentences where no value can be given to the contextual emotion of their emotion words, and all we know is the overall emotion of the sentence.

We will, therefore, have two different sets of experiments. For the first set of sentences, the data are all annotated, so we will take a supervised approach. In the first set of experiments, we also investigate the performance of our features separately. In these experiments, we group the features defined in Section 4.3.1 based on their nature and we assess the usefulness of each group of the features individually. For their evaluation, we use t-test to find out which group contributes to the results significantly. We also look at how the data representation features are related to classes learnt by the machine learning algorithm.

For the second set of sentences, we combine supervised and unsupervised learning. We train a classifier on the first set of data and we use the model to classify the emotion words into their contextual emotion in the second set of data. Finally, we propose an unsupervised method to combine the contextual emotion of all the emotion words in a sentence and

calculate the emotion of the sentence.

For evaluation, we report precision, recall, F-measure and accuracy to compare the results. We also define two baselines for each set of experiments to compare our results with. The experiments are presented in the next three subsections.

4.3.3 Experiments on sentences with one emotion word

In these experiments, we only use the sentences which include exactly one emotion word. We first explain the baselines and then the results of our experiments.

Baselines We develop two baseline systems to assess the difficulty of our task. The first baseline labels the sentences the same as the emotion of the most frequent class, which is a typical baseline in machine learning tasks (Aman and Szpakowicz, 2007; Alm et al., 2005). This baseline will result in 31% accuracy.

The second baseline labels the emotion of the sentence the same as the prior emotion of the only emotion word in the sentence. The accuracy of this experiment is 51%, remarkably higher than the first baseline's accuracy. The second baseline is particularly designed to address the emotion of the sentence only via the prior emotion of the emotion words. It will allow us to assess the difference between the emotion of the sentence based on the prior emotion of the words in the sentence and the emotion which we determine when we consider the context and its effect on the emotion of the sentence.

Learning experiments In this part, we use two classification algorithms, Support Vector Machines (SVM) and Logistic Regression (LR), and two different set of features, the set of features from Section 4.3.1 and Bag-of-Words (unigrams). Unigram models have been widely used in text classification and shown to give good results in sentiment classification tasks.

In general, SVM has long been a method of choice for sentiment recognition in text. For the classification, we use the SMO algorithm (Platt, 1998) from Weka (Hall et al., 2009), setting *10-fold cross validation* as a test option. We compare applying SMO to two sets of features: (i) Bag-of-Words, which are binary features defining whether a unigram exists in a

sentence,[2] and (ii) an extended set of features that we described in section 4.3.1 and we will call them "our features".

We also compare those two results with the third experiment: apply SimpleLogistic (Sumner et al., 2005) from Weka to our set of features, again setting *10-fold cross validation* as the test option. Logistic regression is a discriminative probabilistic classification model which operates over real-valued vector inputs. It is relatively slow to train compared to the other classifiers. It also requires extensive tuning in the form of feature selection and implementation to achieve state-of-the-art classification performance. Logistic regression models with large numbers of features and limited amounts of training data are highly prone to over-fitting (Alias-i, 2008). Besides, logistic regression is slow and it is known to only work on data represented by a small set of features. That is why we do not apply SimpleLogistic to Bag-of-Words features. On the other hand, the number of our features is relatively low, so we find logistic regression to be a good choice of classifier for our representation method. The classification results are shown in Table 4.3.

The results of both experiments using our set of features significantly outperform (on the basis of a paired t-test, p=0.005) both the baselines and SVM applied to Bag-of-Words features. We get the best result, however, when we apply logistic regression to our feature set. The number of our features and the nature of the features we introduce make them an appropriate choice of data representation for logistic regression methods.

4.3.4 Feature set evaluation experiments

In this section, we evaluate the contribution of each group of features to the logistic regression classification results. We perform two series of evaluations. In the first set of evaluations, we use the regression coefficient of the features. The coefficient indicates the nature of the relationship between a particular independent variable and the dependent variable, while a negative value implies an inverse relationship. As a result of applying logistic regression on the sentences with one emotion word, we get the coefficients that are all shown in Table 4.4 (we only show non-zero values). Although the coefficient values are not a measure

[2]We use unigrams from the corpus, selected using "StringtoWordVector" filter from Weka.

		Precision	Recall	F	Accuracy
SVM + Bag-of-Words	Happiness	0.59	0.67	0.63	
	Sadness	0.38	0.45	0.41	
	Anger	0.40	0.31	0.35	
	Surprise	0.41	0.33	0.37	
	Disgust	0.51	0.43	0.47	
	Fear	0.55	0.50	0.52	
	Non-emo	0.49	0.48	0.48	
	Macro-average	0.47	0.45	0.46	50.72%
SVM + our features	Happiness	0.68	0.78	0.73	
	Sadness	0.49	0.58	0.53	
	Anger	0.66	0.48	0.56	
	Surprise	0.61	0.31	0.41	
	Disgust	0.43	0.38	0.40	
	Fear	0.67	0.63	0.65	
	Non-emo	0.51	0.53	0.52	
	Macro-average	0.58	0.53	0.55	58.88%
Logistic Regression + our features	Happiness	0.78	0.82	0.80	
	Sadness	0.53	0.64	0.58	
	Anger	0.69	0.62	0.66	
	Surprise	0.89	0.47	0.62	
	Disgust	0.81	0.41	0.55	
	Fear	0.71	0.71	0.71	
	Non-emo	0.53	0.64	0.58	
	Macro-average	0.70	0.61	0.65	66.88%

Table 4.3: Classification experiments on the dataset with one emotion word in each sentence. Each experiment is marked by the method and the feature set.

of significance *per se*, we use them to roughly estimate the relation of features with each emotion class. Here, we highlight those relations between the coefficients and each emotion class which we find interesting.

First, we consider the features related to the prior emotion and the prior polarity of the emotion word and their relation with the emotion classes. We will assess the relation of part-of-speech and dependency features later. In the *Sadness* class, there is a positive relation between the emotion class and two features, 'sad' and 'negative'. The 'sad' feature indicates whether the prior emotion of the emotion word is sadness; the 'negative' feature shows whether the prior polarity of the emotion word is negative based on the lexicon.

In the *Happiness* class, as expected, there is a inverse relation between the 'negative' feature and this class, while it has a direct relation with 'positive' and 'happy' features.

In the *Fear* class, other than a direct relation with the 'fear' feature, it also has a inverse relation with the 'surprise' feature. This probably indicates the relation of the two emotion classes, as it can be seen in the sentence "She was shocked to see him with a gun".

The *Disgust* class has a direct relation with 'negative', 'anger', and 'disgust' features and a inverse relation with 'sad'. This is, however, in agreement with the fact that Disgust is considered as a challenging and difficult emotion class, which the human annotators often find hard to distinguish from *Anger* (Aman and Szpakowicz, 2007).

The *Anger* class is also in the inverse relation with the 'fear' feature, for which we have not found any particular reason.

One interesting relation in the *Surprise* class is its inverse relation with the 'neg' feature, which may seem odd if we consider surprise as a negative emotion. A closer look into the dataset, however, shows many sentences such as "my night last night turned out amazing", which definitely carry a positive surprise.

An investigation of the relation of POS features with each emotion class shows that there is a strong connection, but we do not see any regular pattern worth noting.

Last but not least, we only find very few relations between 'dependency' features and emotion classes, which are mainly the polarity and negation modifiers. Particularly the two dependency features, 'neg' and 'modified-by-positive', can be highlighted for the *No-Emotion*

Sadness	[negative]	0.90
	[sad]	2.11
	[adj]	-1.04
	[RP-0]	-1.14
	[VB-0]	-0.93
	[modifies-negative]	0.71
Happiness	[negative]	-0.89
	[positive]	0.66
	[happy]	1.92
	[IN-0]	-1.62
	[DT-1]	0.64
	[DT-2]	0.81
	[neg]	-1.19
Fear	[fear]	2.83
	[surprise]	-0.87
	[CC–2]	1.23
	[FW-1]	1.93
	[RB-2]	-1.34
	[modifies-negative]	0.86
	[length]	0.03
Disgust	[negative]	0.74
	[anger]	1.30
	[disgust]	2.81
	[sad]	-0.63
	[verb]	0.54
	[NNS–1]	1.63
	[TO–1]	-1.03
Anger	[negative]	0.69
	[anger]	2.44
	[fear]	-1.00
	[EX–2]	2.58
	[WRB–2]	2.08
	[UH-1]	2.36
	[PRP-2]	1.00
Surprise	[negative]	-0.63
	[surprise]	1.29
	[VB-0]	-0.83
	[IN-1]	0.93
	[TO-1]	1.38
	[CC-2]	1.31
	[modified-by-negative]	-0.74
NonEmo	[verb]	1.17
	[IN-0]	1.07
	[JJ-1]	0.71
	[NNS-1]	-1.27
	[PRP-1]	-0.78
	[neg]	1.54
	[modified-by-positive]	-0.50

Table 4.4: The coefficients of features with each class. A positive coefficient value indicates a direct relation of the feature and the class, and a negative value implies an inverse relationship.

class. Also, the *Happiness* class has a inverse relation with the 'neg' feature that indicated whether the emotion word is modified by a negation word. We also can see that none of the intensity-related features can be seen among the features related with emotion classes; the main reason could be the size of the list (only 112 words) used as an intensifier lexicon, which barely covers any of the words in our dataset. The sparseness of the "dependency" features is problematic. Our observations suggest that very few features in the "dependency" group can occur together in one sentence.

In the second set of experiments meant to evaluate the features, we go more deeply into investigating the significance of the features. For these experiments, we group the features by similarity. The groups are listed below.

1. Prior polarity indicates whether the prior polarity of the emotion word is positive or negative.
2. Lexical includes all the features from the lexicons which we use: the prior emotion, the prior polarity, the intensity and the part-of-speech of the emotion word from the emotion lexicon.
3. Part-of-speech is the same as the part-of-spech features explained in Section 4.3.1.
4. Dependency includes all the features defined as dependency-tree features in Section 4.3.1, except the intensity-related features which form a separate group.
5. Negation includes two negation features. First, is the prior polarity of the emotion word negative? Second, is the emotion word modified by a negative word?
6. Intensity groups all the features based on the intensifier lexicon. It includes the feature which defines the presence of the emotion word on the intensifier list, and the features which identify the dependency between the emotion word and the words from the intensity lexicon.
7. Length is the length of the sentence in words.

We perform a series of experiments in which different sets of features are added to the prior-emotion features and new classifiers are trained. Each group of features is applied separately, and the accuracy of the method is calculated. Afterwards, we compare their

Features	Significant ($p < 0.001$)	Significant ($p < 0.01$)	Not significant	No change
Prior Polarity			54.42%	
Lexical	57.25%			
POS	60.25%			
Dependency		55.20%		
Negation		55.99%		
Intensity			53.86%	
Length				53.44%

Table 4.5: T-test results of feature comparison, based on the accuracy of the classifier on each feature set.

accuracy with the baseline of 53.44%, which is the classifier's result only on the prior-emotion feature of the emotion word.

As shown in Table 4.5, all groups of the features except length contribute positively to the classification result. POS and lexical features, with the accuracy of 60.25% and 57.25% respectively, have the strongest effect and significantly outperform the baseline (on the basis of a paired t-test, p=0.001). While dependency and negation features, with the accuracy of 55.20% and 55.90%, improve the results significantly (paired t-test, p=0.01), the improvements from the prior polarity, intensity and length of the sentence are not significant.

4.3.5 Experiments on sentences with more than one emotion word

In this set of experiments, we work with sentences with more than one emotion word, and we combine supervised and unsupervised learning to place each sentence in one of six emotion classes or the non-emotion class. Here, we have two-step classification. The first step is to learn the classification model from the sentences with one emotion word, explained in Section 3.3.2. Next, we can use the model thus created to classify each emotion word, represented by its contextual features, into one of k emotion classes via majority voting.

However, we followed a slightly different procedure. Instead of producing a simple classification of each test sample (the emotion word and its context features), our classifier can produce a class probability vector that gives the probability of the example belonging to each of the k classes. We sum these class probability vectors for each of emotion classes, to obtain the voted class probability vector.[3] Finally the sentence is put in the emotion class with the maximal voted probability.

The process is shown in Figure 4.2.

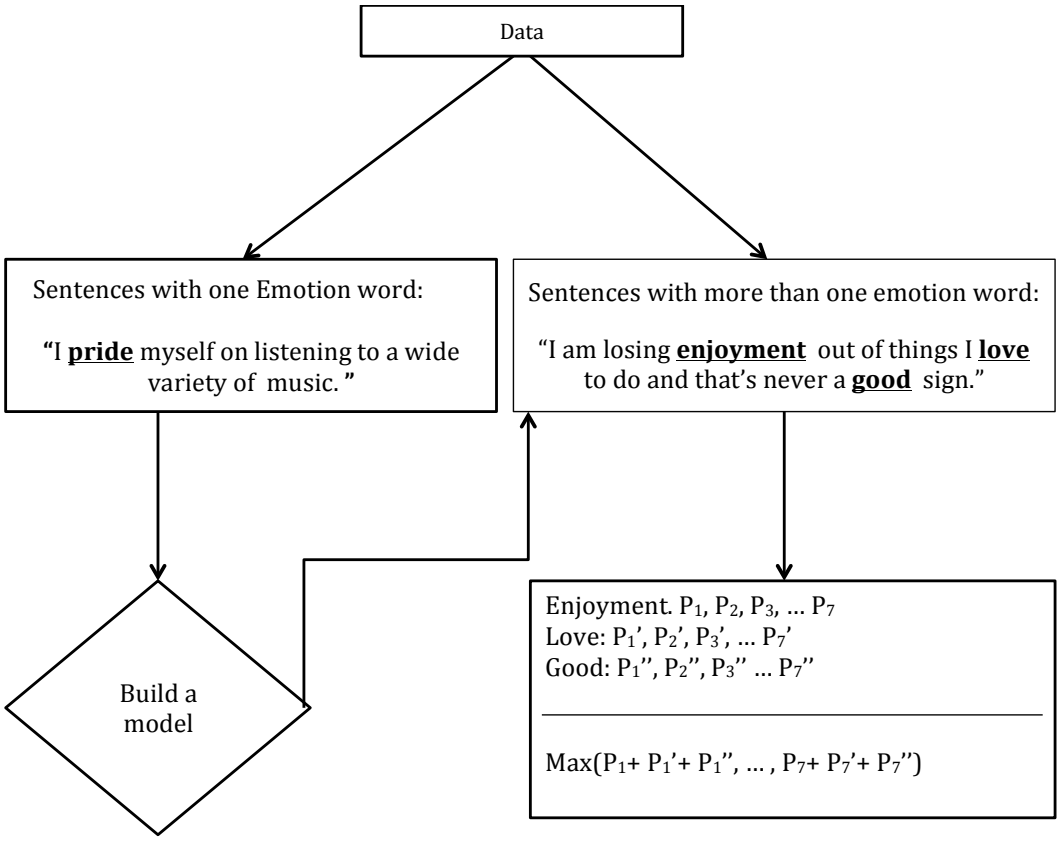

Figure 4.2: Syntax-based emotion recognition for sentences with more than one emotion word. Pk is the probability that the example belongs to emotion class k (k = 1 to 7).

[3]We could normalize the resulted vector so that the probabilities sum to 1, but it does not affect the final result which is choosing the class with maximum probability.

Baselines We develop two baseline systems. The first baseline labels all the sentences the same, as the emotion of the most frequent class, giving 32% accuracy. The second baseline labels the emotion of the sentence the same as the most frequently occurring prior emotion of the emotion words in the sentence. In the case of a tie, we randomly pick one of the emotions. The accuracy of this experiment is 45%. Again, as a second baseline we choose a baseline that is based on the prior emotion of the emotion words so that we can compare it with the results based on contextual emotion of the emotion words in the sentence.

Learning experiments We first train the classifier on the part of the dataset which only includes sentences with one emotion word. For sentences with more than one emotion word, we represent each emotion word and its context by the set of features explained in Section 4.3.1. For each emotion word, we get the probability distribution of emotion classes (calculated by the logistic regression function learned from the annotated data in the previous step). We add up the probabilities of each class for all emotion words and select the class with the maximum probability.

The result, shown in Table 4.6, is compared using supervised learning, SVM, with Bag-of-Words features, explained in the previous section, with *10-fold cross validation* as a test option.[4]

The results in Table 4.6 suggest that the result of learning applied to our set of features significantly outperforms (on the basis of a paired t-test, p=0.005) both baselines and the result of SVM algorithm applied to Bag-of-Words features.

4.3.6 Discussion and evaluation

In these set of experiments, we showed that our method and the features significantly outperform the baselines and the SVM result applied to Bag-of-Words. For the final conclusion,

[4]Support vector machine by nature is not a probabilistic classifier; therefore we do not apply SVM to our features in this set of experiments. However, one can use the option that fits logistic regression models to the outputs of the support vector machine. It is a post-hoc probability score wrapped around the final SVM, which some scientists criticize as not being really probabilistic.

		Precision	Recall	F	Accuracy
SVM + Bag-of-Words	Happiness	0.52	0.60	0.54	
	Sadness	0.35	0.33	0.34	
	Anger	0.30	0.27	0.29	
	Surprise	0.14	0.11	0.12	
	Disgust	0.30	0.17	0.21	
	Fear	0.44	0.29	0.35	
	Non-emo	0.23	0.35	0.28	
	Macro-average	0.32	0.30	0.31	36.71%
Logistic Regression + unsupervised + our features	Happiness	0.63	0.71	0.67	
	Sadness	0.67	0.44	0.53	
	Anger	0.50	0.41	0.45	
	Surprise	1.00	0.22	0.36	
	Disgust	0.80	0.22	0.34	
	Fear	0.60	0.64	0.62	
	Non-emo	0.37	0.69	0.48	
	Macro-average	0.65	0.47	0.54	54.43%

Table 4.6: Classification experiments on the dataset with more than one emotion word in each sentence. Each experiment is marked by the method and the feature set.

we add one more comparison. As we can see from Table 4.6, the accuracy result of applying SVM to Bag-of-Words is really low. Because supervised methods scale well on large datasets, one reason could be the size of the data we use in this experiment; therefore we try to compare the results of the two experiments on all 758 sentences with at least one emotion word.

For this comparison, we apply SVM with Bag-of-Words features to all of 758 sentences and we get an accuracy of 55.17%. Considering our features and methodology, we cannot apply logistic regression with our features to the whole dataset. We calculate its accuracy by counting the percentage of correctly classified instances in both parts of the dataset, used in the two experiments, and we get an accuracy of 64.36%. We also compare the results with the baselines. The first baseline, which is the percentage of most frequent class (Happiness in this case), results in 31.50% accuracy. The second baseline based on the prior emotion of the emotion words results in 50.13% accuracy. It is notable that the result of applying LR to our set of features is still significantly better than the result of applying SVM to Bag-of-Words and both baselines; this supports our earlier conclusion. It is hard to compare the results mentioned thus far, so we have combined all the results in Figure 4.3, which displays the accuracy obtained by each experiment.

We also looked into our results and assessed the cases where the contextual emotion is different from the prior emotion of the emotion word. Consider the sentence "Joe said it does not happen that often so it does not bother him." The emotion lexicon classifies the word "bother" as *angry*. This is also the emotion of the sentence if we only consider the prior emotion of words. In our set of features, however, we consider the negation in the sentence, so the sentence is classified as no-emotion rather than angry. Another interesting sentence is the rather simple "You look like her I guess." Based on the lexicon, the word "like" is in the happy category, while the sentence is no-emotion. In this case, the part-of-speech features play an important role and they catch the fact that "like" is not a verb here. It does not convey a happy emotion, and the sentence is classified as no-emotion.

We also analyzed the errors and found some common errors due to:

1. limited coverage of the emotion lexicon, as in the sentence "I'm tired of broken promises

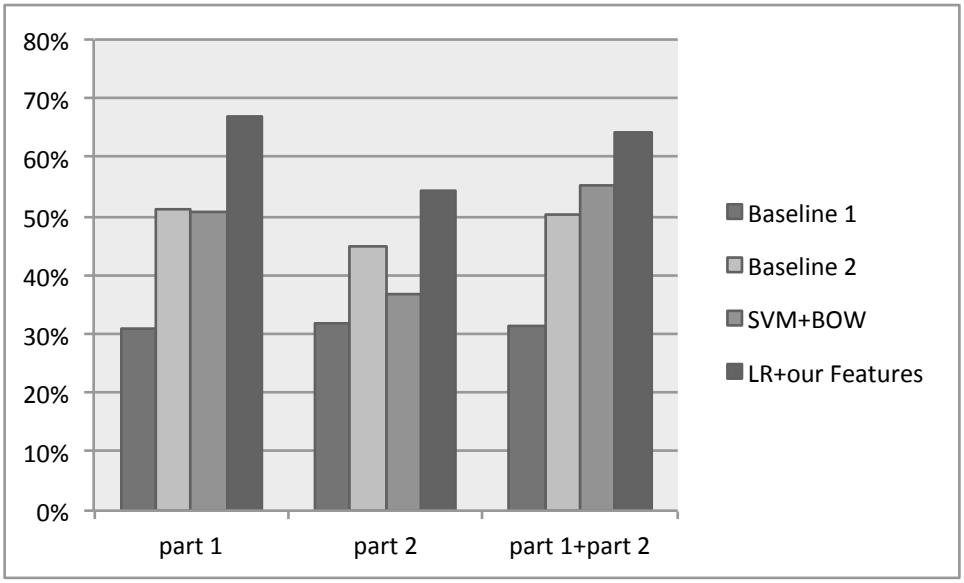

Figure 4.3: The comparison of accuracy results of all experiments for sentences with one emotion word (part 1), sentences with more than one emotion words (part 2), and sentences with at least one emotion word (part 1+part 2).

and empty dreams." – it expresses sadness, but contains no emotion-bearing word according to WordNet-Affect;

2. complex or incomplete sentences which cause the parser to fail or return incorrect data, resulting in incorrect dependency-tree information – examples are the text "Good to love life." or the complex (anger-bearing) sentence "The second reason I was pissed was because not only was my friend is dying from this disease and has had it the entire time that we have been friends but to my knowledge he has not taken the appropriate care of himself that a person with HIV should be taking."

We will discuss the first issue in the next chapter.

4.3.7 Experiments on all emotion-bearing sentences

In the previous experiments, we noticed a tight relation between the number of words in each emotion list in the emotion lexicon and the number of sentences derived for each emotion class. Since this suggests that a larger lexicon could have a greater coverage of emotion

expressions, we used the NRC Emotion lexicon which covers more emotion words. As explained in Section 3.3.2, however, we could not use the NRC Emotion lexicon to train our classifier to learn the contextual emotion of an emotion word. That was due to the lack of sentences with one emotion word (belonging to one emotion class). For example, according to the NRC Emotion lexicon the word "hate" belongs to four negative emotion classes: anger, disgust, fear and sadness. This is not necessarily because the word is ambiguous. It can be due to the quality of Turkers' work who not specialists and might not be precise enough.

On the other hand, one of the weaknesses of our approach is the fact that we cannot use all the instances in the dataset. That is why we could not directly compare our results with the previous results achieved. Again, the main reason is the low coverage of the emotion lexicon we used. The other reason is the limitation of our method: we had only to choose the sentences with one or more emotion words.

To relax these restrictions, we use the NRC Emotion lexicon with a different method of choosing the cue word. In the previous set of experiments, our main focus was on sentences with one emotion word, in order to compare the prior emotion of an emotion word with its contextual emotion. For sentences with one emotion word, we used the only emotion word of the sentence as a cue word. The features were defined for each cue word based on its context. In this set of experiments, most of the sentences are either sentences with more than one emotion word or sentences which have no emotion word. That is why we also found it helpful to add a few sentence-based features. In the following experiments, based on the sentence and the number of emotion words in the sentence, we also have a different strategy of finding the cue word. We have three types of sentences.

1. Sentences with more than one emotion word. In these sentences, we will select the cue word based on a simple heuristic assumption that the emotion words which belong to fewer emotion classes are more informative than emotion words which belong to more emotion classes in the lexicon. Also, when there is more than one emotion word which belongs to the same number of emotion classes, we will choose one at random. For example, the sentence "I am sorry someone died but to mislead me like that is inexcusable." contains three emotion words according to the NRC Emotion lexicon.

The word "die" is marked with two classes, "mislead" and "inexcusable" with three,[5] so in this sentence we choose "died" as the cue word.

2. Sentences with one emotion word. In this case, we simply consider the only emotion word as a cue word. For example, in the sentence "I pride myself on listening to wide variety of music." we take "pride" as the cue word.
3. Sentences with no emotion word. In this case, we will take the root of the sentence (based on the dependency tree result) as a cue word. We can calculate all the features regarding the root word rather than the emotion word. The root, however, will be emotionally neutral: it belongs to none of the emotion classes in the lexicon. The values of the emotion-word features, explained in Section 4.3.1, will be zero. For example, in the sentence "It was the best summer I have ever experienced.", there is no emotion word according to the NRC Emotion lexicon, so the cue word is "summer", the root of the sentence.

For this set of experiments, we added the following sentence-based features.

1. The number of emotion words in each emotion class according to the NRC Emotion lexicon. This feature is especially useful for sentences with more than one emotion word.
2. The number of polar words for each polarity class based on the prior-polarity lexicon. This feature and the next one will give us more information about the sentence. They can be particularly useful for sentences with no emotion word.
3. The number of intensifiers in the sentence based on the intensifier lexicon.
4. Whether the sentence is in passive voice.
5. Whether there is an auxiliary verb in the sentence.
6. Whether there is a copula in the sentence.

The last three features were motivated by the work of Wiebe et al. (2005), who state that those features are useful. They found that a sentence tends to be more subjective if it has a

[5]"die: *fear* and *sadness*, "mislead": *anger*, *fear* and disgust, "inexcusable": *anger*, *disgust* and *sadness*

copula verb which connects the subject to the complement, and auxiliary verbs which express tense, aspect and modality. They also looked into their training data and found that the passive voice is often used to query someone about a specific opinion; therefore they found the active and passive voices to behave very differently. The features are computed from the dependency tree of the sentence. We add these new features to the previous features and we use them all to represent the data. We run the Logistic Regression classification algorithm, setting *10-fold cross validation* as a test option. Table 4.7 shows the results.

We want first to validate the application of the extended feature set with the new methodology: we compare its result to both baseline experiments explained in Section 4.3.2 and also the SVM result applied to Bag-of-Words. The first baseline labels the sentences the same as the emotion of the most frequent class, which result in 41.50% accuracy. The second baseline labels the emotion of the sentence the same as the most frequently occurring prior-emotion of the emotion words according to the NRC Emotion lexicon in the sentence. If there is a tie, we randomly pick one of the emotions. The accuracy of this experiment is 51.39%. Finally, SVM's application to Bag-of-Words results in 55.89% accuracy. By comparing these results with our result – the first experiment in Table 4.7 – we can see a significant improvement in accuracy.

This experiment is also designed to compare our features and our approach with the previous work on the whole dataset by (Aman and Szpakowicz, 2007). We have two concerns here: first the work in (Aman and Szpakowicz, 2007) used the whole original dataset (4090 sentences) with all the no-emotion sentences. Due to the lack of balance, we removed 2000 no-emotion sentences, which makes the sizes of the datasets different. So, we cannot directly compare these results.

On the other hand, no-emotion sentences are supposed to have fewer emotion words than other sentences. While we do not quite agree – we saw many no-emotion sentences with emotion words – a high percentage of sentences with no emotion words are no-emotion sentences. A closer look at the dataset shows many short sentences such as "Good night!", "Okay" and "Yeah" annotated as no-emotion sentences. So, given the different nature of no-emotion sentences, they need further investigation. Our experiments on the whole dataset explained

in Section 3.3.2 also emphasize our concern. When using the whole dataset including all no-emotion sentences, we saw a dramatic drop compared to only using the emotion-bearing sentences. That brings us to the last experiments, in which we apply our procedure to all the 1290 emotion-bearing sentences on the dataset from (Aman and Szpakowicz, 2007) using both the NRC Emotion lexicon and WordNet-Affect lexicon. The result are shown in Table 4.7.

We expected an improvement due to a larger lexicon, but the experiment using WordNet-Affect improves the result particularly in the negative emotion classes. That could be due to the fact that each word can belong to multiple emotion classes in the NRC Emotion lexicon, which is quite often among negative emotion classes. We cannot directly compare our results with Aman's results, because the datasets differ. F-measure, precision and recall for each class are reported on the whole dataset, but we only used the emotion-bearing sentences of that dataset. To show how hard this task is, and to see where we stand, the best results from (Aman and Szpakowicz, 2007) are repeated in Table 4.8.

4.4 Summary

In this chapter, we discussed how to detect the emotion expressed in a sentence using lexical and syntax-motivated features. We showed that our results are promising in detecting emotion. However, we believe that is not enough; we would like to go deeper and address the question what stimulus evokes the emotional response in the experiencer. We also would like to study instances that do not indicate any explicit emotion expression, such as those which do not include any emotion keyword. These two matters will be taken up in the next two chapters.

		Precision	Recall	F	Accuracy
1290 Emo-sentences + NRC lexicon	Happiness	0.66	0.93	0.77	
	Sadness	0.57	0.41	0.48	
	Anger	0.52	0.48	0.50	
	Surprise	0.69	0.30	0.42	
	Disgust	0.52	0.31	0.39	
	Fear	0.56	0.42	0.48	
	Macro-average	0.59	0.47	0.52	61.63%
1290 Emo-sentences + Word-NetAffect	Happiness	0.65	0.92	0.76	
	Sadness	0.63	0.50	0.56	
	Anger	0.67	0.49	0.56	
	Surprise	0.64	0.43	0.51	
	Disgust	0.61	0.36	0.45	
	Fear	0.73	0.52	0.61	
	Macro-average	0.65	0.54	0.59	65.04%

Table 4.7: Classification experiments on the emotion-bearing sentences in Aman's dataset, using the NRC Emotion lexicon and WordNet-Affect lexicon respectively.

		Precision	Recall	F
Aman's best results	Happiness	0.81	0.70	0.75
	Sadness	0.60	0.42	0.49
	Anger	0.65	0.44	0.52
	Surprise	0.72	0.41	0.52
	Disgust	0.67	0.49	0.57
	Fear	0.87	0.51	0.64
	Non-emo	0.59	0.62	0.60

Table 4.8: Aman's best results on the whole dataset with 4090 sentences, explained in Section 3.3.2.

Chapter 5

Recognition of emotion stimulus in text

Note. Parts of this chapter have been published earlier in (Ghazi et al., 2015).

5.1 Introduction

Emotion recognition, studied in the previous chapter, has focused on detecting the expressed emotion in text. A related challenging question, *why* the experiencer feels that emotion, has, to date, received very little attention. The task is difficult and there are no annotated English resources. FrameNet refers to the person, event or state of affairs which evokes the emotional response in the experiencer as emotion *stimulus*.[1] In this chapter, we explain how we automatically built a dataset annotated with both the emotion and the stimulus using FrameNet's *emotion-directed* frame. We address the problem both as an information extraction and classification problem. We build a Conditional Random Fields (CRF) learner and a Conditional Markov Model (CMM) learner, two sequential learning models to detect the emotion stimulus spans in emotion-bearing sentences, and we convert the problem to a classification problem with four classes: Begin-of-Stimulus, Continue-Stimulus, End-of-Stimulus, and Outside.

Causality is a semantic relation defined as "the relationship between cause and effect",[2]

[1] Most authors talk generically of *cause*.

[2] `http://www.oxforddictionaries.com/`

where the latter is understood as a consequence of the former. Causality detection used to depend on hand-coded domain-specific knowledge bases Kaplan and Berry-Rogghe (1991). Girju and Moldovan (2002) defined semantic constraints to rank possible causality, and machine learning techniques now prevail.

Girju (2003) described automatic detection of lexico-syntactic patterns which express causation. There are two steps: discover lexico-syntactic patterns, and apply inductive learning to detect automatically syntactic and semantic constraints (rules) on the constituent components. Chang and Choi (2006) extracted possible causal relations between noun phrases, via a bootstrapping method of causality extraction using cue phrases and word-pair probabilities. A simple supervised method in Bethard and Martin (2008) trained SVM models on features derived from WordNet and the Google N-gram corpus; giving the causal relation classifier temporal information boosted the results significantly. All that work, however, addresses causality in general, which is quite different from detecting the cause of an emotion. The causal relations between two parts of the text are sought. We look for the emotion stimulus when the emotions can be conveyed explicitly in the text or are implicit. Moreover, we show our work is different from general causal detection by applying a state-of-the-art discourse parser that detects discourse relations including causal relations. We show the two relevant relations, "Cause" and "Explanation" barely detect emotion stimulus.[3] However, we benefit from existing causal detection systems as features that are added to our learning model.

A more specific task in causality analysis, most similar to our task, is to identify sources of opinions. Bethard et al. (2004) used machine-learning techniques to identify propositional opinions and their holders (sources). That pioneering work had limited scope: it considered only propositional opinions (which function as the sentential complement of a predicate), and only direct sources. Choi et al. (2005), also among the pioneers in this field, viewed the problem as an information extraction task and tackled it using sequence tagging and pattern matching techniques simultaneously. They hypothesized that information extraction

[3] In the whole data it only detected 12 instances of *Cause* and 10 instances of *Explanation*. This shows the causal relations detected by this discourse parser cannot be used to detect emotion cause.

techniques would be well-suited to source identification because an opinion statement can be viewed as a kind of speech event with the source as the agent. Choi et al. (2006) identify two types of opinion-related entities, expressions of opinions and sources of opinions, along with the linking relation between them. Kim and Hovy (2006) analyzed judgment opinions, which they define as consisting of valence, a holder and a topic. All that work invokes interchangeably the terms *source of an opinion* and *opinion holder*. Although the source can be the reason which implies the opinion, it is mainly seen as the opinion holder – thus, it is a task very different than ours.

We focus on detecting the *emotion stimulus*. According to FrameNet (Fillmore et al., 2003), an *experiencer* has a particular emotional *state*, which may be described in terms of a specific *stimulus* provoking it, or a *topic* categorizing the kind of stimulus. Rather than expressing the experiencer directly, it may have in its place a particular *event* (with participants who are experiencers of the emotion) or an *expressor* (*e.g.*, a body-part of gesture which would give an indication of the experiencer's state to an external observer). There can also be a *circumstance* under which the response occurs or a *reason* why the stimulus evokes the particular response in the experiencer. Consider, for example, the sentence "In the Commons, Labour MPs unleashed their anger at the Liberal Democrats for promising to back the Government." "Labour MPs" is the *Experiencer*, "anger" is the expression of emotion, "the Liberal Democrats" is the *Emotion Stimulus*, and "for promising to back the Government" is the *Explanation*. We want our system to find the reason why Labour MPs were angry: to return the span *the Liberal Democrats* as the emotion stimulus.

Fillmore et al. (2003) define six core frame elements for an emotion and six non-core elements (Table 5.1). Of these elements, emotion stimulus seems to be the closest to saying *why* the experiencer feels that emotion. Therefore, here we focus on this aspect of emotion analysis.

We are particularly interested in the stimulus. That is because determining what provokes an emotion has such intriguing applications as consumer behaviour analysis or mental-health care. It would be very useful for systems which answer question such as "how [x] feels about [y]" or "why [x] feels [y]". It also has practical importance to text summarization, because

emotion expressions and emotion stimuli tend to be the most informative in an expressive sentence, so they can get higher weight in abstractive summarization.

We discuss emotion stimuli, a dataset we automatically built with emotion stimulus labels, and ways of detecting them. Section 5.2 covers the related work. Section 5.3 explains the process of collecting and annotating an emotion stimulus dataset using FrameNet data. Section 5.4 discusses the features and Section 5.5 explains the baselines for detecting emotion stimuli in emotion-bearing sentences as well as the experiments and the results.

5.2 Related work

Detecting emotion stimuli is a very new concept in sentiment analysis. Emotion-stimulus interaction is an eventive relation, which potentially yields crucial information in terms of information extraction. For example, we can predict future events or decide on the best reaction if we know what causes an emotion (Chen et al., 2010).

On the face of it, detecting emotion stimulus among the emotion frame elements might resemble Semantic Role Labelling (SRL). SRL is the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles (Gildea and Jurafsky, 2002). We did look into two SR labelers, namely Senna (`http://ml.nec-labs.com/senna/`) and Semafor (Das et al., 2014). Senna is generic; tags Arg0, Arg1, *etc.* would help little in our case. Semafor is specific to FrameNet and its frame elements. It works for common frame elements such as "Agent", but struggles with specific elements such as "Emotion stimulus". Also, Semafor is very computationally demanding and requires substantial time and memory. Moreover, neither tool is the state of the art, so their errors could carry into our system.

On the other hand, SRL relies on lexical and syntactic features. Often SRL tools' goal is to label text spans in the original text by assuming strong dependencies between the arguments and developing joint models to identify the various elements. However, this does not necessary apply to our task because different frame elements such as the emotion itself

Core	Event	The occasion or happening in which Experiencers in a certain emotional state participate.
	Experiencer	The person or sentient entity who experiences or feels the emotions.
	Expressor	It marks expressions that indicate a body part, gesture or other expression of the Experiencer that reflects his or her emotional state.
	State	The abstract noun that describes a more lasting experience by the Experiencer.
	Stimulus	The person, event, or state of affairs that evokes the emotional response in the Experiencer.
	Topic	The general area in which the emotion occurs. It indicates a range of possible Stimulus.
Non-Core	Circumstances	The condition(s) under which the Stimulus evokes its response.
	Degree	The extent to which the Experiencer's emotion deviates from the norm for the emotion.
	Empathy-target	The individual or individuals with which the Experiencer identifies emotionally and thus shares their emotional response.
	Reason/Explanation	The explanation for why the Stimulus evokes a certain emotional response.
	Manner	It is any description of the way in which the Experiencer experiences the Stimulus which is not covered by more specific frame elements. Manner may also describe a state of the Experiencer that affects the details of the emotional experience.
	Parameter	A domain in which the Experiencer experiences the Stimulus.

Table 5.1: The elements of the emotion frame in FrameNet.

are not necessarily explicitly stated in text. We believe our task should also consider emotion-specific features and semantic roles inspired by the cognitive structure of emotions (events, agents, *etc.*.). This distinguishes our work from pure SRL, so we define a separate problem which is not a subtask of SRL, so the same methods and tools need not apply.

Event-based emotion detection has been addressed in some previous research (Wu et al., 2006; Tokuhisa et al., 2008; Lu et al., 2010) but, to the best of our knowledge, only Chen et al. (2010), Lee et al. (2010a) and Lee et al. (2010b) have worked on emotion cause detection. They explore emotions and their causes, focusing on five primary emotions – happiness, sadness, fear, anger and surprise – in Chinese texts. They have constructed a Chinese emotion corpus, but they focus on explicit emotions presented by emotion keywords, so each emotion keyword is annotated with its corresponding causes, if they exist. In their dataset, the authors observe that most causes appear in the same clause of the representation of the emotion, so a clause might be the most appropriate unit to signal a cause. We find such granularity too large to be considered an emotion stimulus in English. Also, clauses were distinguished by punctuation: comma, period, question mark and exclamation mark. Just four punctuation marks are not enough to capture English clauses adequately.

Using linguistic cues, including causative verbs and perception verbs, the authors create patterns to extract general cause expressions or specific constructions for emotion causes. They formalize emotion cause detection as multi-label classification. Each instance may contain more than one label, such as "left-1, left-0", to represent the location of the clauses which are part of the cause. We have no evidence that their findings can be valid for English data. Also, keyword spotting is used to detect the emotion word in the sentence and try to find the cause of that emotion, but not all emotions are expressed explicitly by emotion keywords.

The other paper on automatically identifying linguistic contexts which contains possible causes of emotions is applied to Italian newspaper articles (Russo et al., 2011). This work was built on a previous work by Lee et al. (2010b), which is different in two main ways: 1) Russo et al. use data mining techniques to automatically induce the rules for sentential contexts in which an event cause phrase is expressed; and 2) they exploit a common-sense repository of

EK-ECE (emotion keyword - emotion cause event) noun couples for the identification of the correct ECE nouns. They use an Italian equivalent of WordNet-Affect and a dataset of 22469 words. They choose the noun-noun bigrams where at least one item has an associated sense corresponding to the "emotion" category and the other word corresponded to the WordNet-Affect label of the "emotion eliciting situation". As a result they built a list of 133 EK and 161 nominal lemmas of possible ECEs.

Using only noun-noun bigrams is a strong limitation for the following reasons: emotions are known to be mostly presented by adjectives and adverbs, while they are not considered here; also, it was shown that the main cause of emotions are verbal events, while here only nominal emotion eliciting situations are considered. Russo et al. used a pattern extractor that considered all combination of EK-ECE occurring in the same sentence and selected 47 patterns that were used among the features for the clustering and classifying. To find a set of rules for the emotion cause detection, they used a combination of clustering and classification. They used the EM algorithm to detect two data clusters, one which includes examples where the EK and the emotion cause co-occur in the same sentence and another where the emotion cause is not present or it occurs in a different sentence. The results of each clustering analysis are passed to the Weka PART rule-induction classifier. The PART classifier provided a total of 49 rules for the detection of EK-ECE contexts. However, the classifier only identifies whether a cause phrase occurs in the same sentence, but is not able to identify the noun which corresponds to the ECE.

Finally, they use a combination of EK-ECE Mutual Information (MI) values and a probability measure of eventivity,[4] to detect the ECE with the highest probability. In the end, their evaluation results seem low. The 50% accuracy in the sentences that EK-ECE couples co-occur in which are only 82 sentences, and 30.76% of 13 sentences that the EK- ECE couples do not co-occur in the same sentence is not really convincing that their method works.

[4] This measure is useful in case more than one ECEs is occurring in the context. It is derived from the probability of the ECE lemma being associated with Multi-WordNet top ontological classes, namely "event", "state", "phenomenon" and "act".

All in all, these prior studies (Chen et al., 2010; Lee et al., 2010a,b; Russo et al., 2011) seem to be a good start, particularly regarding the data analysis and interesting statistics it reveals for emotion causes. There is, however, no evidence that those findings can be valid for English data. This is rather practical research, and the problem is not very well defined yet. The theoretical background and fundamental research on this problem is missing. That work is more of a solution for an over-simplified version of a complicated problem.

In English text, Neviarouskaya and Aono (2013) analyze emotion causes in sentences, where emotions are explicitly indicated through emotion keywords. They developed a corpus of emotion causes containing 523 sentences for 22 emotions. That means that, on average, it has around 25 sentences for each emotion, which makes it a very sparse dataset. Based on the analysis of the corpus, they distinguish eight types of linguistic relations between emotion and its cause. They also define five rules to detect cause phrases introduced by propositions and conjunctions. To evaluate their method, they apply the rules that they defined as a result of the analysis of the corpus on the same data, which makes us wonder if it is a reliable evaluation. It is not surprising if those rules do well on their own dataset, but whether the method scales well to other data is questionable. Finally, their work contains interesting observations and hand-crafted rules, yet it is preliminary.

Mohammad et al. (2014) focus on the roles of experiencer, state and stimulus from FrameNet – see Table 5.1. They used Amazon's Mechanical Turk service to annotate 2000 tweets about the 2012 US presidential elections. However, they found that in most instances the experiencer of emotions in a tweet is the tweeter, so they limit their attention to automatic detection of the emotional state and the stimulus. They also see stimulus narrowly as one towards whom the emotion is directed and they define a limited list of the emotion targets (whom) based on the domain and the data they built; therefore their task is domain-dependent and limited in scope, with a definition of stimulus different from ours.

Some of the assumption and simplifications in the above-mentioned work (Wu et al., 2006; Tokuhisa et al., 2008; Lu et al., 2010) should be avoided:

- In the work which classifies text into emotions by means of considering events, the rules and patterns are mainly built manually. This really constrains the coverage of

the solution. An attractive alternative is to learn the patterns and rules automatically from a statistical representation based on semantically inspired features.

- In the event-based work, the event-emotion relation is considered as a one-to-one relation. This ignores the cases when there are multiple events in the text which contribute to the emotion felt. To take care of that, we need to analyze discourse structures, and to consider how events are related when there is more than one event. This could be either a sentence containing more than one event or a longer text with more than one sentence annotated with one emotion.
- Due to the complexity of this problem, its application in free text requires a detailed understanding and analysis of the text. Deep analysis of text, both syntactic and semantic, is required.

In the end, what stands out is the fact that, as far as we know, there has been no significant previous work on emotion cause detection in *English* data. This may be due to the relative complexity of English in expressing emotions in text, or to the limitation in existing resources and datasets either for supervised machine learning or for evaluation purposes.[5] Therefore, we have collected data and built a new dataset for detecting emotion stimuli. Using the dataset, we also explore different baselines to establish the difficulty of the task. Finally, we train a supervised information extraction model, as well as a classification model, which detect the emotion stimulus spans in emotion-bearing sentences.

5.3 Data collection and annotation

Fillmore et al. (2003) define 173 emotion-directed lexical units which correspond to different emotions. A lexical unit (LU) is a word/meaning pair (essentially a lexeme).[6] Typically, each sense of a polysemous word belongs to a different semantic frame. "Happy", "angry" and "furious" are examples of LUs in the Emotion-directed frame. For each of them, FrameNet

[5]We found the dataset built by Neviarouskaya and Aono (2013) too sparse for machine learning methods of cause detection.

[6]The term *lexical unit* will be henceforth abbreviated as LU.

annotators have labelled some sentences. We built a set of sentences marked with the emotion stimulus (cause), as well as the emotion itself. To collect a larger set of data, we used synonyms of emotion LUs to group the data into fewer basic emotions. In the manual synonym annotation task, we suggested Ekman's six emotions of *happiness*, *sadness*, *suprise*, *disgust*, *anger* and *fear* (Ekman, 1992). We also sugggested *shame*, *guilt* and *hope*, posited in the literature (Ortony et al., 1988), to consider if an emotion LU did not fit any of Ekman's emotion. We also allowed the annotators to propose a more appropriate emotion not on the list. In the end, we chose Ekman's emotions and *shame*. *Guilt* and *hope* were not found useful, but *confusion* was suggested as another emotion, with many synonym matches among the emotion LUs. This requires further study. *Confusion* is not considered in related research, and we need more evidence in the literature.

Some emotions fit a basic one; for example, *fury* clearly belongs to the *anger* category. On the other hand, *affront* is not so obvious. We will now discuss the dictionaries and thesauri we used to identify emotions and their synonyms. Then we will describe the manual annotation. Next, we will apply the first round of tagging the emotions with one of the seven classes we chose. We separate the emotion LUs with a strong agreement between sources and annotations from those with weaker agreement. For the latter set of LUs, we relax one of the conditions. We use two more emotion lexicons to break the ties and classify more LUs. Finally we build two datasets, one with the group of emotions with strong agreement and the other with the result of the second phase of annotation added.

5.3.1 Annotating emotion lexical units

We collected the synonyms from two trustworthy online sources.[7] Of the lexical units collected from FrameNet, some are not covered at all and some belong to two emotion classes at the same time. The two sources also do not always agree. To get a tie breaker, we resorted to manual annotation.

We grouped the list of 173 emotion lexical units from FrameNet into Ekman's six emotion

[7] `http://www.thesaurus.com/` and
`http://www.oxforddictionaries.com/definition/english-thesaurus`

classes by using the synonym list from the Oxford Dictionary and from thesaurus.com, joined the results and asked human annotators (fluent English speakers) to verify those annotations. We gave them an Excel table with each lexical unit displayed in a row and each emotion in a column. For each word, the corresponding emotions were marked. A word could be classified into:

I one of Ekman's six emotions (110 words);
II two classes at the same time (14 words);
III none of the six emotion classes (49 words).

For group I, the annotators indicated if they disagreed with the existing classification by crossing the existing mark and indicating the emotion they think is more appropriate (if there is one). For group II, they chose only one of the two emotions they thought was closer to the LU and crossed out the other one. For group III, they chose one of Ekman's six emotion classes plus the three suggested classes, *guilt*, *shame* and *hope*, and grouped the LU into one of them. Finally, there was a column for comments, where the annotators could write any other emotion they thought was more suitable as a synonym of the emotion LU.

Considering both thesauri and human annotators' tags, each lexical unit was tagged with the emotion that had the highest number of votes. We combined the result of the Oxford Dictionary and thesaurus.com, so we had four annotations in total. As a result, 100 out of 173 lexical units were labeled with high agreement (at least 3 out of 4 votes) – shown as strong agreement in Table 5.2.

For the other emotion lexical units, we did not have enough information to group them into one of the seven emotions. We used two more sources – the NRC emotion lexicon (Mohammad and Turney, 2010) and the WordNet-Affect lexicon (Strapparava and Valitutti, 2004) – to break some of the ties. In this step, the LUs were labeled with an emotion for which the number of votes was more than half of the votes (three in our case) and the difference of two votes for the two top emotion classes was at least 2. This added 24 LUs to our set. While the NRC emotion lexicon was not very useful,[8] many ties were broken

[8]It puts many words into multiple negative classes such as *sadness*, *anger*, *disgust*. For example, *Despair* is labeled as *anger*, *disgust*, *fear* and *sadness*.

	happiness	sadness	surprise	disgust	anger	fear	shame	total
Strong Agreement	21	31	12	3	18	13	2	100
Weaker Agreement	21	32	12	8	24	19	6	124
Dictionary and Thesaurus	17	30	11	7	17	25	13	120

Table 5.2: Distribution of lexical units for each emotion class based on human annotators' strong agreement, their weaker agreement, and finally dictionary and thesaurus annotation.

by WordNet-Affect. The result of this extended list is shown as weaker agreement in Table 5.2. In Table 5.2 we have also shown the result of only using the Oxford Dictionary and thesaurus.com.

The list of synonym lexical units that are grouped both by using 1) the online dictionary and thesaurus and 2) human annotators is presented in Tables 5.3-5.5.[9]

As shown in Table 5.5, the number of lexical units with no emotion class when only using dictionary and thesaurus is much higher than when we added human annotation. It is the opposite for lexical units with more than one emotion class. The main reason is that we encouraged human annotators more towards choosing the best option among suggested emotions rather than leaving a lexical unit with no emotion. We also noticed that the first set of results has many more *Shame* instances compare to results with human annotation; the reason can be that we built the first round of results with only Ekman's emotions which is what we presented to the annotators. It might have indirectly motivate them to choose Ekman's emotions and therefore fewer *Shame* instances.

The results of these annotations are used to build the dataset containing sentences with emotion stimulus tags. Section 5.3.2 presents the process of building the dataset.

[9]The words in boldface in Tables 5.3-5.5 get two out of three human annotator votes. Surprisingly, two of the annotators consistently agree on the emotion class of these words.

	Dictionary and Thesaurus	Human Annotators
Happiness	contented.a, delight.n, delighted.a, ecstatic.a, elated.a, elation.n, exhilarated.a, exhilaration.n, glee.n, gleeful.a, thrilled.a, gratification.n, gratified.a, happy.a, jubilant.a, overjoyed.a, pleased.a	contented.a, delight.n, delighted.a, ecstatic.a, elated.a, elation.n, exhilarated.a, exhilaration.n, glee.n, gleeful.a, gratification.n, gratified.a, thrilled.a, happy.a, jubilant.a, overjoyed.a, pleased.a, amused.a, amusement.n, excited.a, excitement.n, *relaxed.a*
Sadness	anguish.n, anguished.a, bitterness.n, blue.a, dejected.a, inconsolable.a,dejection.n, despondency.n, despondent.a, disconsolate.a, downcast.a, glum.a, glumness.n, grief-stricken.a, grief.n, heartbreak.n, heartbroken.a, low-spirited.a, lugubrious.a, miserable.a, mournful.a, sad.a, saddened.a, sadness.n, sorrow.n, sorrowful.a, woebegone.a, depressed.a, desolate.a, downhearted.a	blue.a, dejected.a, dejection.n, depressed.a, desolate.a, despondency.n, despondent.a, disconsolate.a, downcast.a, downhearted.a, glum.a, glumness.n, grief-stricken.a, grief.n, heartbreak.n, heartbroken.a, inconsolable.a, low-spirited.a, lugubrious.a, miserable.a, mournful.a, sad.a, saddened.a, sadness.n, sorrow.n, sorrowful.a, woebegone.a, crestfallen.a, anguish.n, anguished.a, devastated.a, **despair.n** , *crushed.a, tormented.a*
Fear	agitated.a, agitation.n, alarmed.a, anxious.a, concern.n, concerned.a, discouraged.a, discouragement.n, disheartened.a, dismay.n, dismayed.a, frightened.a, horrified.a, horror.n, nervous.a, perplexed.a, perplexity.n, perturbed.a, petrified.a, rattled.a, terror-stricken.a, upset.a, worried.a, disquiet.n, disquieted.a	alarmed.a, anxious.a, disquiet.n, disquieted.a, frightened.a, nervous.a, petrified.a, terror-stricken.a, concern.n, concerned.a, horrified.a, horror.n, worried.a, **dismay.n, dismayed.a, agitated.a, rattled.a, agitation.n, disheartened.a**

Table 5.3: The list of synonym lexical units for the emotion-directed frame in FrameNet (I).

Anger	affronted.a, anger.n, angry.a, cross.a, enraged.a, furious.a, incensed.a, infuriated.a, irate.a, mad.a, miffed.a, offended.a, resentful.a, riled.a, sore.a, vexation.n, vexed.a	affronted.a, anger.n, angry.a, enraged.a, furious.a, incensed.a, infuriated.a, irate.a, mad.a, miffed.a, offended.a, resentful.a, riled.a, outrage.n, fury.n, annoyance.n, irked.a, livid.a, peeved.a, **bitterness.n, cross.a, vexation.n, vexed.a, indignant.a, exasperated.a, exasperation.n**, *unsympathetic.a*
Surprise	astonished.a, astonishment.n, astounded.a, bewildered.a, bewilderment.n, disappointed.a, disappointment.n, stupefaction.n, stupefied.a, flabbergasted.a, nonplussed.a	astonished.a, astonishment.n, astounded.a, nonplussed.a, stupefaction.n, stupefied.a, startled.a, stunned.a, bewildered.a, bewilderment.n, flabbergasted.a, puzzlement.n, *baffled.a, bafflement.n, befuddled.a, disorientation.n, disoriented.a, flummoxed.a, mystification.n, mystified.a*
Disgust	appalled.a, fed up.a, fed-up.a, overwrought.a, revolted.a, disgruntled.a, disgruntlement.n	appalled.a, revolted.a, sickened.a, **revulsion.n, disgruntled.a, disgruntlement.n, fed up.a, fed-up.a**
Shame	abashed.a, ashamed.a, discomfited.a, discomfiture.n, disconcerted.a, disconcertion.n, embarrassed.a, embarrassment.n, flummoxed.a, flustered.a, humiliated.a, mortification.n, mortified.a	ashamed.a, mortified.a, **discomfited.a, embarrassed.a, embarrassment.n, humiliated.a**

Table 5.4: The list of synonym lexical units for the emotion-directed frame in FrameNet (II).

No Class	agonized.a, agony.n, amused.a, amusement.n, baffled.a, bafflement.n, befuddled.a, bored.a, boredom.n, covetous.a, crushed.a, devastated.a, disorientation.n, disoriented.a, embittered.a, excited.a, excitement.n, interest.n, irked.a, livid.a, mystification.n, mystified.a, peeved.a, puzzlement.n, relaxed.a, ruffled.a, sickened.a, stressed.a, sympathetic.a, sympathize.v, sympathy.n, tormented.a, traumatised.a, unsettled.a, unsympathetic.a, harried.a	agonized.a, agony.n
More Than One Class	stunned.a, crestfallen.a, annoyance.n, chagrin.n, chagrined.a, despair.n, displeased.a, displeasure.n, distress.n, distressed.a, exasperated.a, exasperation.n, fury.n, indignant.a, outrage.n, revulsion.n, startled.a	**despair.n, dismay.n, dismayed.a, agitated.a, rattled.a, bitterness.n, cross.a, vexation.n, vexed.a, indignant.a, exasperated.a, exasperation.n, revulsion.n, disgruntled.a, disgruntlement.n, fed up.a, fed-up.a, discomfited.a, embarrassed.a, embarrassment.n, humiliated.a, agitation.n, disheartened.a,** abashed.a, chagrin.n, chagrined.a, disappointed.a, disappointment.n, discomfiture.n, disconcerted.a, disconcertion.n, discouraged.a, discouragement.n, displeased.a, displeasure.n, distress.n, distressed.a, embittered.a, flustered.a, fury.n, harried.a, mortification.n, overwrought.a, perplexed.a, perplexity.n, perturbed.a, ruffled.a, sore.a, traumatised.a, unsettled.a, upset.a,
Not Enough Votes		*unsympathetic.a, relaxed.a, crushed.a, tormented.a, baffled.a, bafflement.n, befuddled.a, disorientation.n, disoriented.a, flummoxed.a, mystification.n, mystified.a,* covetous.a, stressed.a, interest.n, boredom.n, bored.a, sympathetic.a, sympathize.v, sympathy.n

Table 5.5: The list of synonym lexical units for the emotion-directed frame in FrameNet (III).

	happiness	sadness	surprise	disgust	anger	fear	shame	total
Dataset 1	227	98	53	27	168	129	34	736
Dataset 2	227	107	53	59	199	146	68	859

Table 5.6: Distribution of labels in the emotion stimulus datasets. Dataset 1 contains the synonym groups with strong agreement, Dataset 2 also includes those with weaker agreement.

5.3.2 Building the dataset

Using the list of grouped emotion synonyms, we collected FrameNet data manually labelled for each lexical unit. Next, we selected the sentences which contain emotion stimuli and we assigned each sentence to its corresponding emotion class. The distribution of the instances in the dataset appears in Table 5.6. Each instance is a complete sentence, 18 tokens on average, and contains one stimulus assigned to the emotion LU.[10]

We now give examples of emotion stimuli for each emotion class.

- Happiness: "He was a professional musician now , still sensitive and happy <stimulus> doing something he loved </stimulus>."
- Sadness: "He was miserable <stimulus> at being left out </stimulus>."
- Surprise: "His last illness was the most violent , and his doctors were astounded <stimulus> that he survived it </stimulus>."
- Disgust: "Woody Allen wrote and directed this comedy in which he stars as New Yorker Isaac Davis who is in love with his city but disgruntled <stimulus> with his job as a TV comedy writer </stimulus>."
- Anger: "Conner is still furious <stimulus> about losing Tuesday 's protest</stimulus>."
- Fear: "He was petrified <stimulus> of the clippers </stimulus> at first."

[10]All datasets discussed in this section are publicly available (`http://www.eecs.uottawa.ca/~diana/resources/emotion_stimulus_data/`).

	happiness	sadness	surprise	disgust	anger	fear	shame	total
No Stimulus 1	289	446	160	13	181	205	9	1303
No Stimulus 2	289	468	160	63	284	295	78	1637

Table 5.7: Distribution of labels in the emotion dataset with no stimulus tags.

- Shame: "It is to their obvious embarrassment <stimulus> that they are again together</stimulus>. "

As a complementary dataset, we also collected the sentences with no stimulus tag, yet containing expressions of one of the seven emotions. This dataset is much larger than the dataset with stimuli. This makes us wonder whether it is due to the existence of many emotion causes implicit in the context; or if it is because of other possible frame elements in the sentence we disregard, such as circumstances and explanation, which can indicate emotion stimuli; or if a sentence is not enough to always contain the emotion stimulus, so we should consider larger text portions (maybe the current sentence and the previous and next sentence). Nonetheless, we believe that building this dataset is useful in deciding on one of these three reasons. An example of an instance in this dataset that belongs to the non-stimulus class is "He laughed shortly , little amusement on his face .".

In the next section, we also extend the emotion-stimulus dataset by considering other frame elements such as circumstances, explanation, and topic, which can indicate emotion stimuli. The distribution of the emotion instances in this dataset is presented in Table 5.7.

5.3.3 Extending the dataset

The set we built, described in the previous section, is small and well-formed, and contains carefully built data annotated by humans. Here, we would like to study other emotion-directed frame elements, such as *Circumstances*, *Explanation*, *Topic* and *Empathy-target*, to investigate whether they can also indicate emotion stimuli to be added to our dataset.

Among the frame elements described in Table 5.1, we choose the frame elements whose definition is directly connected to stimulus. The frame elements and an example for each

one is as follows:

- Circumstances: the condition under which *stimulus* evokes its response. For example, in the sentence " *{When Anna left Inspector Aziz}* , she was much happier.", "When Anna left Inspector Aziz" is the circumstances that evokes happy emotion.
- Explanation/Reason: the explanation why the *stimulus* evokes a certain emotional response. In the sentence "They were very quiet and sad at first, *{because I would be leaving them}* , but I promised I would never forget them and would often return to visit them.", in this case, "because I would be leaving them" is the reason that explained why they were sad.
- Topic: the general area in which the emotion occurs that indicates a range of possible *stimulus*. The sentence "Floyd Patterson has expressed dismay *{over the current treatment}* .", for this sentence, "over the current treatment " is the topic that evokes *fear* emotion in the experiencer (Floyd Patterson).
- Empathy-target: the individual or individuals with which the experiencer shares their emotional response.[11] In the sentence "Jacob felt to the brim with sorrow *{ the woman whom he had loved}*.", the experiencer (Jacob) shares the *sad* emotion with "the women whom he had loved" which is the empathy-target.

Here, we will study the relation of stimulus frame element with these four other frame elements, by finding whether there is any overlap between them. We also calculate the number of instances where emotion stimulus co-occurs with one of the frame elements as well as detecting the number of cases that each frame element occurs alone. The results are presented in Table 5.8. Next, we look at the data and examples to find out if they represent emotion stimulus and whether they should be added to the dataset.

Some of our observations based on Table 5.8 are:

- Stimulus and Topic frame element never occur at the same time. This can indicate

[11]The Empathy-target's definition is not directly linked with stimulus, but the target is a person that the experience shares their emotional response with which can serve as the person evoking the emotional response.

		Circumstances	Reason	Topic	Empathy-target
happiness	alone	13	3	29	1
	with stimulus	0	0	0	0
sadness	alone	12	7	30	6
	with stimulus	0	1	0	0
surprise	alone	3	1	3	0
	with stimulus	0	0	0	0
disgust	alone	6	1	2	0
	with stimulus	0	0	0	0
anger	alone	21	4	21	1
	with stimulus	2	21	0	0
fear	alone	17	2	38	1
	with stimulus	1	2	0	0
shame	alone	10	2	5	2
	with stimulus	0	0	0	1
total	alone	82	19	128	11
	with stimulus	3	24	0	1

Table 5.8: The number of instances containing other frame elements in the emotion dataset occurring with or without stimulus.

that in cases that there is no specific reason for the emotion felt Topic can be a good indicator of the more generic reason.

- The number of instances where emotion stimulus co-occurs with one of the other four frame elements is 28. The majority of these overlaps, 21 cases, occur with the Explanation frame element within the *anger* class. This is not unexpected, because in anger instances there is a high chance that we would not only mention the primary stimulus of the emotion, we would also add more explanation.

  In the cases when the two frame element spans are consecutive spans, we merge them together and consider the bigger span as the emotion stimulus. Some example of the instances in this group are:

  - "He couldn't take home business associates for dinner because he was embarrassed <Empathy-target> for them </Empathy-target> <Stimulus> to see you </Stimulus>. (Shame)"
  - "The other day he slapped Liam across the face and got mad <Stimulus> at me </Stimulus> <Circumstances> when I complained </Circumstances>. (Anger)"
  - "In the Commons , Labour MPs unleashed their anger <Stimulus> at the Liberal Democrats </Stimulus> <Reason> for promising to back the Government </Reason>. (Anger)"

- There is only one instance with a combination of more than two frame elements, including Empty-target, Topic and Reason, which is discarded from the dataset. That instance is:

  - "Such old people may have little embarrassment <Empathy-target> with each other </Empathy-target> <Topic> about bodily functions </Topic>, perhaps less than some husbands and wives <Reason> because they were the stuff of taken-for-granted childhood </Reason>. (Shame)"

- Finally, in the other 240 instances that contain only one of the four frame elements, we consider that frame element as the emotion cause because our manual investigation verifies the assumption.

	happiness	sadness	surprise	disgust	anger	fear	shame	total
With stimulus	273	162	60	68	246	204	87	1100
No stimulus	243	413	153	54	237	237	58	1395

Table 5.9: Distribution of labels in the emotion stimulus dataset considering Topic, Explanation, Circumstances, and Empathy-target.

The distribution of labels in the extended dataset using extra frame elements is presented in Table 5.9. All our experiments are based on the dataset with weaker agreement explained in Table 5.6. We explain how we use the emotion stimulus dataset to build a supervised model to learn emotion stimulus spans in emotion-bearing sentences. The dataset we built indicates both the emotion expressed and the emotion stimulus in each sentence.[12] In the next section, we only detect emotion stimulus, and assume that the emotion expression is present (in our data, it is). In the next chapter, however, we will address both emotion expression and emotion stimulus detection at the same time.

5.4 Features and data representation

In this section, we explore different data representation methods. We not only address more generic features such as corpus-based features, we also consider syntax-based and semantic features that we found useful for this task.

Corpus-based features. Although we believe corpus-based methods may not suffice for emotion cause detection which is a hard task that needs more sophisticated analysis, we build our first set of experiments using corpus-based features for the sake of comparison and for exploring different features.

We use a set of corpus-based features built in MinorThird's text analysis package `http://minorthird.sourceforge.net/`. Among the features there are the lower-case version of each single word, and analogous features for tokens in a small window to either side of

[12]Although not used in our research, we also tag the *Experiencer*, one of the main frame elements in which research community has shown interest, and we believe it will be beneficial for future users of this dataset.

the word. Here we set the window size to three as suggested by the literature (Choi et al., 2005). Additional token-level features also include information whether the token is a special character such as a comma, and orthographic information. For example, there is a feature, the character pattern "X+", which indicates tokens with all capital letters. The features are grouped into positive and negative. *Positive* refers to the features built and weighted based on the tokens within the stimulus span. *Negative* are the features related to all the tokens outside of the targeted span.

An analysis of our learnt model and the feature weights shows that, for the positive tokens, the left-side token features have a higher weight than the right-side tokens. It is the opposite for the negative tokens. Also, the highest-weighted token features include "at", "with", "about", "that" and emotion words such as "delight", "concerned", "ashamed", "anger" for the left-side tokens. According to the ranked features, many function words are among the highest-weighted features. This indicates that this task is very structure-dependent.

A few examples showcase some of the shortcomings of this model by comparing what is learnt (underlined) and what the actual stimulus (italicized) is.

- "Colette works at marshalling our feelings of revulsion *{{at this}* voracious creature *who has almost killed the poor box thorn.}*" This example shows that, although these features might be useful to detect the beginning of the emotion stimulus, detecting the end of the span seems more challenging for them.
- "He was petrified *{of the clippers}* {at first}." In this case the model has learned that many emotion stimuli start with the word "at", so it chooses "at first" regardless of its semantic and syntactic role in the sentence.

Looking at the predicted labels and comparing them with the actual labels shows that we need deeper semantic and syntactic features.

Part-Of-Speech tags. Part-of-speech tags are the base of most tools detecting more complicated syntactic structures such as parsers and chunkers. Also, in Chapter 4, we showed POS features had the strongest effect and significantly outperformed the baseline. Although that was a different task, we believe it is worth exploring whether it will be useful in this

tasks as well. We use the OpenNLP part-of-speech tagger[13] and we label each part-of-speech with its corresponding tag.

Chunker. Text chunking divides a text into syntactically coherent segments like noun groups or verb groups, but does not specify their internal structure, nor their role in the main sentence. We use the OpenNLP chunker[14] to tag the data with the chunks, because we believe that the chance of an emotion stimulus starting or ending in the middle of a chunk is very low. A chunker should help improve the span precision and recall.

We found the chunks that were returned by OpenNLP chunker where mainly very short. Therefore, we build longer chunk phrases using a few manually-built rules:

- Noun Phrase

  NP + NP → NP

  NP + PP + NP → NP

  PP + NP → NP

- Verb Phrase

  VP + VP → VP

  VP + ADJP → VP

  VP + PRT → VP

In our experiments, we found applying these rules and building longer chunks results in more coverage. Although the shorter chunks were slightly better in token-based precision and recall, longer chunks were beneficial in improving span precision and recall, therefore we use longer chunks in our experiments that are shown in Section 5.5.

Here are examples with the actual and predicted emotion stimulus label, using the corpus-based model. Considering Chunker features helps reducing the kind of errors which these examples illustrate.

- "Their cheerfulness and delight *{{at still being}* *alive}* only made Charlie feel more

[13] `http://opennlp.apache.org/documentation/manual/opennlp.html#tools.postagger`

[14] `https://opennlp.apache.org/documentation/1.5.2-incubating/manual/opennlp.html#tools.chunker`

guilty." "Being alive" should be placed in one chunk, therefore the span will not end in the middle of a chunk.

- "Feeling a little frightened *{{of the dead body behind} him in the cart}* , he stopped for some beer at a pub , where he met Jan Coggan and Laban Tall." Again, "behind him in the cart" should be in one chunk, so by using a chunker we would know the predicted span was incorrect.

Clause. In English grammar, a clause is the smallest grammatical unit which can express a complete proposition. There are two different types of clauses, independent and dependent. An independent clause can stand alone as a complete sentence. Dependent clauses can be nominal, adverbial or adjectival. Noun clauses answer questions like "who(m)?" or "what?" and adverb clauses answer questions like "when?", "where?", "why?".[15] Although there might not be many cases when the whole clause is the emotion stimulus, there are some cases, as mentioned below, which make it worthwhile to look into clauses and considering them among the features.

To mark the clauses in a sentence, we use the OpenNLP parser.[16] As suggested in the literature (Feng et al., 2012), we use the *SBAR* tag, which represents subordinate clauses in the parse trees, to identify dependent clauses in a sentence. We use the *S* tag inside the sentence to indicate independent clauses. The output of the parser for the following examples shows how the emotion stimulus tag exactly aligns with the *SBAR* tag which indicates the subordinate clause.

- "I am pleased *{that they have responded very positively}*."[17]
- "I was so pleased {she lived until just after Sam was born}."[18]

As shown in the examples, the subordinate and independent clauses highly overlap. The relation of the clauses is rather hierarchical than flat. Therefore just indicating the span of each clause will not be very informative. As a result we use some simplification to split

[15] `http://www.learnenglish.de/grammar/clausetext.html`

[16] `http://opennlp.apache.org/documentation/manual/opennlp.html#tools.parser`

[17] The parse is "I am pleased [SBAR that [S they have responded very positively.]]"

[18] The parse is "I was so pleased [SBAR [S she lived until [SBAR just after [S Sam was born.]]]]"

each instance into the highest number of non-overlapping clauses. For example if there is a clause that has only one other clause inside, we discard the inner clause and we only indicate the longer clause. Also if there is a clause with more than one clause inside, we discard the longer clause and we only consider the sub-clauses.

Discourse Structure. Here, we use an RST-style discourse parser (RST = Rhetorical Structure Theory) by Feng et al. (2014) that is based on HILDA parser (Hernault et al., 2010) which offers hierarchical text organization. It is also not limited to sentence level analysis and it can handle more than one sentence documents. HILDA not only builds a tree structure construction, it also labels discourse relations. In RST, a text is first divided into non-overlapping text chunks, called elementary discourse units (EDUs). For instance, the following sentence from our dataset, indicating happy emotion with the stimulus emphasized:

> Her cheeks were already aglow from a combination of the heat and the exhilaration *of being the centre of attention* , and this had lent an extra , youthful radiance to her beauty .

can be segmented into EDUs shown here:

> [Her cheeks were already aglow from a combination of the heat and the exhilaration] 1 [of being the centre of attention ,] 2 [and this had lent an extra , youthful radiance to her beauty .] 3

Next, consecutive EDUs are put in relation with each other, using a pre-defined set of rhetorical, or discourse, relations. The final goal of the discourse parser is to produce a tree structure as a representation of how all units of the text relate to each other. Also in each relation the EDUs can be labeled as Nucleus of the relation or Satellite. The nucleus is considered as more prominent than the satellite. A more elaborate discourse parser's result of the above-mentioned sentence that shows the relations between the EDUs is shown here:

(Joint[N][N](Here N means Nucleus and S means Satellite)

(Elaboration[N][S]

-!Her cheeks were already aglow from a combination of the heat and the exhilaration!

-!of being the centre of attention ,!-)

-!and this had lent an extra , youthful radiance to her beauty .<P>!-)

Here, there is an Elaboration relation between first and second EDU, while the first one is nucleus and the second EDU is satellite. And there is a joint relation between the first two EDUs and the third one.

In our dataset, around 18% of the sentences were segmented as having exactly one EDU that contain both emotion expression and emotion stimulus in the same EDU.

Among the 82% of the sentences that contain more than one EDU, in about 70% of the instances the stimulus that was tagged from FrameNet occurs within the boundaries of an EDU, which indicates an EDU as a reasonable boundary of an emotion stimulus.

To represent the data using discourse parser, we traverse the parse tree from the root, we get the relation sequence and the type (Nuclei or Satellite) sequence for each EDU. So we convert the information of the above-mentioned parse tree to :

Relation: -JOINT-ELABORATION, Type: -NUCLEUS-NUCLEUS, Text: Her cheeks were already aglow from a combination of the heat and the exhilaration

Relation: -JOINT-ELABORATION, Type: -NUCLEUS-SATELITE, Text: of being the centre of attention ,

Relation: -JOINT, Type: -NUCLEUS, Text: and this had lent an extra , youthful radiance to her beauty.

We found “Elaboration”, “Attribution”, and “Joint” seen most in emotion stimulus containing EDUs. We also found the dominant type of EDUs that contain stimulus is Nucleus (in 192 cases), the second-ranked type is Satellite (84 cases), next Nucleus-Nucleus and Nucleus-Satellite (50 cases each). Moreover, there are only 12 EDUs having Cause relation and 10 EDUs with Explanation relation. This indicates that our task is different than general causal detection task and the causal relations in existing tools cannot be used to detect emotion stimulus.

As a result of applying these features, we did not see significant improvement. The reason might be that we consider the hierarchy of the types and relations while we do not have much

data to afford this. Therefore we simplify our features to three features in case of discourse type (Nucleus, Satellite, Mixed)[19] and only leaf relations in case of discourse relations.

Events. FrameNet's definition of emotion stimulus treats events as one of the main factors in detecting stimuli. That is why we use a tool to automatically detect events and add them to the features. The following examples show how events can be the main part of emotion stimuli.

- "I am desolate *{that Anthony has died}*."
- "I join the Gentleman in expressing our sorrow *{at that tragic loss}*."

Evita (Pustejovsky et al., 2010) is a tool which develops algorithms to tag mentions of events in text, tag time expressions, and temporally anchor and order the events. The EVENT tag is used to annotate those elements in a text which mark the semantic events described by it. Syntactically, events are typically verb phrases, although some nominals, such as "crash" in "killed by the crash", will also be annotated as events. Evita's event classes are *aspectual*, *I-action*, *I-state*, *occurrence*, *perception*, *reporting* and *state*. Here we indicate both the events and each event's class as our feature.

Emotion Keyword. Finally, to see the effect of emotion expression features on emotion stimulus detection, we label the emotion lexical units as a feature that we consider in our learning model.

In the next section, we build our learning models using these features. We also show the significance of applying each feature set.

5.5 Automatically detecting emotion stimulus

In this section, we use the data that we discussed and will address automatic detection of emotion stimulus span, assuming the sentences contain an emotion. However, we think it would be beneficial to first detect whether there is an emotion stimulus then to retrieve the emotion stimulus span. Therefore, in this section we build two models, one defining whether

[19]When this applies, the type sequence is a mix of Nucleus and Satellite

an instance contains an emotion stimulus or not and then a model that extracts the emotion stimulus span.

Statistical learning models help avoid the biases and insufficiency of coverage of rule-based and pattern-based detection methods. We build two supervised models, one for each task. We have built labelled data annotated with whether the instance contain an emotion stimulus or not. For instances that contain emotion stimulus the stimulus span is also tagged. So we have the privilege to explore supervised learning methods for both classification and information extraction parts. Of the datasets explained in Tables 5.6-5.7, we use the one with weaker agreement, with 859 instances for emotion stimulus and 1637 instances with no emotion stimulus, as it gives us more data.

One of the most common learning paradigms for performing such labelling tasks are Hidden Markov Models or probabilistic finite-state automata to identify the most likely sequence of labels for the words in any given sentence (Wallach, 2004). Such models, however, do not support tractable inference, and they represent the data by assuming their independence. One way of satisfying both these criteria is to use a model which defines a conditional probability over label sequences given a particular observation sequence rather than using a joint distribution over both label and observation sequences.

Conditional random fields (CRFs) (Lafferty et al., 2001) are a probabilistic framework for labelling and segmenting sequential data, based on a conditional model which labels a novel observation sequence x by selecting the label sequence y maximizing the conditional probability p(y | x). We use CRF from MinorThird (Cohen, 2004). This learning tool offers CRF among many more learning algorithms. That gives us an advantage to compare different sequential learning models. As a second learner, in addition to the CRF learner from MinorThird, we apply SVMCMM which is a combination of Support Vector Machine (SVM) and Conditional Markov Model (CMM). SVM is trained on the token-based features provided and CMM is trained on the output of the SVM which results in returning spans of text rather than individual tokens. MinorThird also allows error analysis: comparing the predicted labels with the actual labels by highlighting them on the actual text. It also ranks the features based on the weight, so we can see which features have contributed the most.

	Precision	Recall	F-measure
Bag-of-Words	0.75	0.75	0.75
Corpus-based	0.78	0.77	0.78
Bag-of-Words+ Our Features	0.82	0.78	0.80

Table 5.10: The results of detecting if an emotion stimulus exists in the sentence.

5.5.1 Binary class stimulus detection

In this part, we use a dataset with 859 instances having stimulus and 1637 instances with no stimulus. The problem is a two-class classification. As a typical machine learning baseline, one of our baselines is the majority class ratio, which is 65.58%.

Another baseline we built here is by applying Bag-of-Words which results in the F-measure value of 75.9%. We also use corpus-based features as well as our features explained in Section 5.4. We apply *sequence annotator learner* with specifying *CRF Learner* as the sequence classifier learner, and setting *10-fold cross validation*. The results of these experiments are shown in Table 5.10.

The results show that the special character and orthographic information of instances with stimulus is not very different than instances with no stimulus; therefore the corpus-based results is not substantially better than Bag-of-Words.[20]

Moreover, adding our features substantially improves the result. That indicates that the syntactic and discourse structure play an important role in differentiating sentences with stimulus and those with no stimulus.

In the next section, we work on extracting the stimulus span, assuming that it exists in the sentence.

[20]The difference is mainly because of the window size which is zero in case of the Bag-of-Words as it only contains unigrams. We set it to three for corpus-based so it considers three tokens to the right and three tokens to the left as well.

5.5.2 Stimulus span extraction

Baselines

To assess the difficulty of extracting emotion stimuli, we develop baseline systems which work with intuitive features. We also explore various features and their effect on stimulus detection results. The baselines we explain here set the ground for comparing our results and evaluating the performance of different models. First we look at the case when we would return the whole sentence as an emotion stimulus. In that case, although we have 100% recall, we only get 34.50% precision which means that the emotion stimulus spans are about 34% of the whole data.

One of the main properties of emotions is that they are generally elicited by stimulus events: something happens to the organism to stimulate or trigger a response after having been evaluated for its significance (Scherer, 2005). That is why events seem to be the most obvious indicators of emotions; we build our first two baselines upon events. We *are* aware that they are not the only emotion stimuli, but we believe them to be important enough.

One problem with using events as emotion stimuli is that detecting events is a challenging task, not a solved problem. The literature suggests two types of events: verbal and nominal, and verbal events are much more numerous (Chen et al., 2010). Our first baseline is based on Evita (Pustejovsky et al., 2010), a tool which detects both nominal and verbal events; as an emotion stimulus, we select an event in a sentence at random.

We noted earlier that not only events can be stimuli. FrameNet defines a stimulus as the person, event or state of affairs which evokes the emotional response in the Experiencer. For the third baseline, we recognize an emotion stimulus in a larger portion of the sentence, a phrase or a syntactically motivated group of words. We use the OpenNLP chunker, and we randomly select as an emotion stimulus any chunk with a verb. We also use the longer version of chunks that we created by modifying the result of OpenNLP chunker using few manually built rules. Although the longer chunks increase the precision and recall compared to shorter version of chunks, the result of this baseline is still very low.

Lee et al. (2010b) used as a baseline the clause with the first verb to the left of the emotion

keyword. In English, however, there are single-clause sentences such as "My grandfather's death made me very sad." or "I was surprised to hear the ISIS news." with both the emotion state and the stimulus. The whole sentence would be returned as an emotion stimulus. Even so, we believe that it is worth exploring and investigating how useful a clause will be in detecting emotion stimuli in English. As the next baseline, we select a random clause as the stimulus. In OpenNLP parse trees, we take the *S* and *SBAR* tags as indicators of independent and dependent clauses, respectively. Next, we randomly choose one of the clauses as the emotion stimulus.

As a result of applying discourse parser each instance is divided into EDUs. So, we also use a random EDU for each instance as an emotion stimulus.

Finally, we use Bag-of-Words as a typical baseline for all NLP tasks and for the sake of comparison. While the previous baselines were rule-based systems with simple heuristics, this baseline is built by applying CRF sequence learner from MinorThird to all the unigrams in the text.

The baseline results are presented in Table 5.11. In span detection problems, the evaluation measures can either be based on the number of matching tokens or be more strict and consider the exact spans and the number of exact matches. Naturally, the value of token-level measures is higher than span-level measures. That is why, to build the higher-bound baseline, we report the token-level precision, recall and F-measure.

The results indicate very low coverage in the first three baselines, while the clause, EDU and Bag-of-Words baselines are much higher. The reason can be that the data in FrameNet are well-formed and carefully collected. Considering that emotion stimulus spans take about 34% of the whole data and having a quick look at the instances show that the emotion stimulus tends to be longer than just a verb or a phrase. The length of the items in the first two baselines is one token on average, and the third baseline has two tokens on average, while the emotion stimulus spans are 6 tokens on average. Stimuli are long enough to say why a particular emotion was experienced. Also, random clause has the highest recall because they tend to be longer part of sentence. Therefore, as baselines, the random clause and Bag-of-Words experiments have higher coverage. We believe that, although the first three baselines'

	Precision	Recall	F-measure
Rndom verb	0.21	0.06	0.09
Evita events	0.26	0.04	0.07
Random Chunk	0.29	0.07	0.11
Random Longer Chunks	0.41	0.14	0.21
Random Clause	0.43	0.76	0.55
EDU	0.36	0.37	0.37
Bag-of-Words	0.53	0.48	0.50

Table 5.11: The results of the baselines at the token level. Therefore the results at span level are much lower. For example, for Bag-of-Words, the span-level evaluation scores are 0.353, 0.264 and 0.302.

results are really low when used as the only feature in a simple rule-based system, some of them are still interesting enough to keep as features in our machine learning methods. In the next section, we will discuss adding these features, and we will compare the results with the baselines.

Experiments and results

Here, we will use our features explained in Section 5.4 individually and together. *Individually*, because we want to evaluate each feature to see whether they affect the results substantially. *Together*, because we want to build the best model using these features and the two different machine learning methods.

We chose *sequence annotator learner* with specifying *CRF Learner* and *Conditional Markov Model (CMM)* as the sequence classifier learner, and setting *5-fold cross validation* as an evaluation option.[21] To build our stimulus extraction models, we also set the tagging reduction to "BeginContinueEndReduction" for CRF model.[22]

[21]The dataset is too small for 10-fold cross validation.

[22]The other options are "InsideOutside", "BeginContinueOutsideReduction".

	Learning Model	Token Precision	Token Recall	Token F	Span Precision	Span Recall	Span F
Corpus-Based	CRF	0.73	0.72	0.73	0.54	0.52	0.53
Corpus-Based + POS	CRF	0.73	0.79	0.76	0.63	0.64	0.62
Corpus-Based + chunker	CRF	0.68	0.72	0.70	0.57	0.53	0.55
Corpus-Based + clause	CRF	0.80	0.69	0.74	0.62	0.56	0.59
Corpus-Based + Event	CRF	0.78	0.71	0.74	0.59	0.53	0.56
Corpus-Based + Discourse structure	CRF	0.82	0.62	0.71	0.73	0.55	0.63
Corpus-Based +Emotion Expression	CRF	0.79	0.76	0.77	0.66	0.62	0.64
Our Features	CRF	0.81	0.79	0.79	0.68	0.64	0.66
	CMM	0.852	0.83	0.84	0.71	0.69	0.70
All Features	CRF	0.77	0.87	0.82	0.65	0.67	0.66
	CMM	0.85	0.86	0.85	0.73	0.71	0.72

Table 5.12: Results of detecting emotion stimulus using different types of features.

The results of these experiments are shown in Table 5.12. They show that each set of features improves the span learning model, while POS, clause-based features, and emotion expression features substantially improve both span and token level results. It also shows that the result of applying CMM substantially outperform the CRF method. Finally, the result of combining all features is higher than all the result of each features plus corpus-based (although significantly better than some and less significant from others).

A closer look at different features and their results shows that:

- Although the results of corpus-based experiment outperform all the baselines, we notice that the span precision and recall are much lower than at the token level. The reason

is that the syntactic structure of a sentence is not considered in this set of features. Also these features are very dependent on the data. As we do not have a big dataset to learn from, it can suffer from low recall because many tokens in a new instance might not have been seen before.

- POS features greatly help with improving span-level results and highly improves the recall in token-level assessment. The reason can be that using a sequence of POS is more generic that using the token itself, so it can catch more of the stimulus spans. And finally they substantially improve span results.
- Although using the chunker did not help with token-level results, it improved the span-level precision, which was the initial intuition of using chunks. Overall, the improvement of using the chunker was not substantial.
- The results show that clause-based features highly improve the precision in both token and span-level results. Adding these features substantially improves the result over only using corpus-based feature.
- Although Events do not substantially improve the results, they have a high weight in detecting the beginning of a span specially as the first token after the beginning of the stimulus span.
- Discourse structure features highly improve both token and span-level precision. They are also one of the feature sets that substantially improve the span-level result.
- Emotion expression is the most effective set of features in improving both token and span-level results. Based on the learnt model and the feature weights, among very high weighted features in this experiment are one token before the beginning of stimulus being emotion keyword. There is also a high negative relation with the emotion keyword being in the stimulus continue class and there is a high positive weight of emotion keyword being among Outside tokens.
- Finally, when we combine all the features together and compare with the experiment when we only use syntactic and semantic features without the corpus-based features (tagged as "our features"), we notice the results are not substantially different. "All Features" is slightly better in token-based results but nothing substantial. On the other

hand, "Our Features" set has a few advantages over the "All Features" set. First of all, the number of features is much lower and the features are much more meaningful; second, the span-level results are almost the same and the token-based results only slightly decrease; finally, this model is much less dependent on the data and the corpus. As we will use the model learnt here on another dataset in the next chapter, the model without corpus-based features would be preferable.[23]

When combining all feature sets, it is hard to know which group of features had the most affect on the overall result only by looking at the final results. Therefore, we look into the learnt models and the weight of each feature set. Here, we will highlight some of the features that have the highest positive or negative weights in the last two CRF models built. Comparing the features of the two models shows the result of syntactic and semantic features are quite consistent between the two models. However, in the "All Features" model, the corpus-based features are always among the highest weighted features and they dominate the results. Therefore we show the list of features of "Our Feature" model which are more meaningful. The results are shown in Tables 5.13-5.14.

In the Beginning of Stimulus class, the POS of the token itself is also very important with IN being the highest and and TO and VBG being affective but DT, JJ, PRP$ being negatively effective. Although we consider a window of three to both sides, the features show that the closer to the token, the higher the weight of features. For example, there is a high correlation between the first left token being an emotion keyword, while the correlation of the first left token being an event is negative. Also the discourse type of the tokens in the right of the beginning of the stimulus have a positive impact. Although the left side has higher weight in detecting the beginning of the stimulus, the right side tends to be more important to detect the end of a stimulus span.

In the End of Stimulus class, the weight of the end token being an event is positive, which happens in many cases such as the examples that were mentioned in Section 5.4. Other features that seem to be important in detecting the end of stimulus span are discourse

[23]When comparing the result of "Our Feature" with the previous experiments on individual features, they also include corpus-based features.

Begin	Token : [POS]	2.80
	[Chunk]	0.62
	Left 1 : [EmoKey]	2.87
	[POS]	1.43
	[Event]	-0.37
	Left 2: [POS]	0.61
	[EmoKey]	0.31
	[Discourse-Type]	-0.23
	Left 3 : [EmoKey]	0.66
	[POS]	0.64
	[Chunk]	0.35
	Right 0 : [Chunk]	1.10
	[POS]	1.09
	[Clause]	0.65
	Right 1 : [Discourse-type]	0.61
	[Clause]	0.58
	[POS]	0.52
	Right 2 : [Chunk]	0.72
	[POS]	0.54
	[Discourse-type]	0.46

Continue	Token : [POS]	0.68
	[Chunk]	0.55
	[EmoKey]	-0.67
	Left 1 : [EmoKey]	-0.87
	[POS]	0.74
	[Chunk]	0.45
	Left 2 : [EmoKey]	1.39
	[POS]	0.64
	[Chunk]	0.60
	Left 3 : [EmoKey]	1.15
	[Chunk]	0.74
	[Discourse-relation]	0.56
	Right 0 : [Chunk]	1.20
	[POS]	0.80
	[EmoKey]	- 0.54
	Right 1 : [Discourse-type]	2.04
	[POS]	0.73
	[EmoKey]	- 0.44
	Right 2 : [Event]	0.46
	[POS]	0.44
	[EmoKey]	- 0.37

Table 5.13: The highest-weighted features in "Our Features" experiment. A positive weight value indicates a direct relation of the feature and the class, and a negative value implies an inverse relationship. (Part I)

End	Token : [POS]	0.88
	[Event]	0.72
	[Chunk]	0.69
	Left 1 : [POS]	1.25
	[Chunk]	1.12
	[Clause]	0.90
	Left 2 : [EmoKey]	1.54
	[Clause]	1.10
	[Chunk]	0.81
	Left 3 : [Chunk]	0.64
	[EmoKey]	0.63
	[Clause]	0.59
	Right 0 : [POS]	2.06
	[Discourse-relation]	0.55
	[Chunk]	0.32
	Right 1 : [POS]	0.45
	[Discourse-relation]	0.33
	Right 2 : [Chunk]	0.36
	[POS]	0.33

Outside	Token : [POS]	4.04
	[Clause]	1.19
	[Discourse-relation]	1.10
	[EmoKey]	1.16
	Left 1 : [EmoKey]	- 0.73
	[Event]	0.51
	[POS]	0.47
	Left 2 : [Event]	0.40
	[POS]	0.29
	[EmoKey]	- 0.91
	Left 3 : [POS]	0.85
	[Discourse-relation]	0.38
	[EmoKey]	- 0.31
	Right 0 : [EmoKey]	1.00
	[POS]	0.64
	[Chunk]	0.60
	Right 1 : [POS]	0.79
	[EmoKey]	0.73
	[Chunk]	0.49
	Right 2 : [POS]	0.82
	[EmoKey]	0.80
	[Chunk]	0.41

Table 5.14: The highest-weighted features in "Our Features" experiment. A positive weight value indicates a direct relation of the feature and the class, and a negative value implies an inverse relationship. (Part II)

relation in the right side of the end and POS of the very first right token has the highest weight. Looking closer at the POS feature shows that the highest-weighted POSs in this cases are punctuation and dot as defining end of sentence or clause and it matches with what we observed from the dataset.

In the Continue class there is a negative correlation between Emotion Keyword and this class, which is exactly what we would expect as emotion keywords and emotion stimulus are mutually exclusive spans.

Finally in the Outside class, the weight of features related to the token itself is much higher than the tokens in their right or left side, while POS has the highest weight in detecting these tokens. Interestingly, we found the emotion keywords had negative weight for the tokens in the left and positive weight for the tokens in the right. We believe that can be an indicator of the fact that emotion keywords occur later in the sentence and tend to be in the right side of many Outside tokens.

Although the results in Tables 5.13-5.14 show that all the features were effective in the learnt model, some were more useful and some less. For example, Emotion Keywords and POS features are the highest weighted features in all the classes. The reason can be that emotion stimulus spans normally do not contain emotion keywords, so the location of emotion keyword in the sentence should be a good indicator of the location of the stimulus span. However, the problem can be when there is no emotion keywords in the sentence(implicit emotion expression) which will be addressed in the next chapter. POS tags are the base of any parser that we used, here for chunking and defining clauses. That can indicate that more basic features that will give our algorithm more freedom to learn each tag and the correlation of different features rather than using more advanced features.

Next, we will analyse the common errors and we will further discuss the results.

Discussion and error-analysis

We also look into the result of the combination of all these features and we notice some errors that need further attention. We would like to highlight them here.

1. Mixed emotions: In these examples there is normally more than one emotion expression and based on the emotion expression that is our target the emotion stimulus can be different. Some of the examples for this case are:

   "I 'm delighted that you were pleased *{with the response }*." which is an example of 'pleased' emotion. However, if we look for the stimulus for 'delighted' emotion, it is seems to be more like : "I 'm delighted *{that you were pleased with the response}*."

   "Even the Prince of Wales , depressed *{ at the thought of the ordeal to come}* , cheered up at the sight of the food ." In this case if we would also detect the stimulus for "cheer up" it would have been: "at the sight of the food ."

   Looking at our result shows that this is what causes some of the errors. For example in the sentence:

   Reduced dogs are often somewhat nervous {of people}, and yet may also become exited *{ to see you}*, when you have been out.

   Our method returns the first stimulus "of people" which is not wrong considering *nervous* being the emotion. However in the data the second stimulus is annotated as the stimulus.

2. Rare cases:

   In the example: "Life , you know , *{it}* 's miserable for them .", while the actual answer is "life" but "it" is marked as stimulus which is a pronoun referring to "life". In our model we do not consider coreference resolution, which can resolve "it" to "life". However, we believe this can be interesting future work.

   In the example: "This would normally push up her heart rate but with her pulse already racing due to the alcohol , dancing and general *{party}* excitement the increase is not as noticeable as normal .", having at adjective before the emotion expression to indicate stimulus is very rare in our data and our model has not seen enough cases like this to learn from.

We also go further and address the problem as a classification problem to get more detailed results. We convert the dataset to a four-class classification problem with Stimulus-Begin, Stimulus-Continue, Stimulus-End and Outside as classes. For each instance, the

tokens are labelled as beginning of the stimulus span (Stimulus-Begin), inside the stimulus span (Stimulus-Continue), end of the stimulus span (Stimulus-End), or Outside which indicates tokens that are outside of the stimulus span. So the Information Retrieval problem will be converted to a multi-class classification problem. Here, we use CRFLearner classifier from the classification package of MinorThird as the machine learning method which reports *Error rate* for evaluation. We only use corpus-based features and our syntactic and semantic feature without corpus-based features (shown as "our features"). We show the result of these experiments in Table 5.15. All the results shown in the table are error rate, so the smaller, the better the result.

Features	stimulus-begin Error rate	stimulus-continue Error rate	stimulus-end Error rate	Outside Error rate
Corpus-Based	25.95	41.51	41.27	8.7
Our Features	13.7	19.83	32.25	12.42

Table 5.15: Results of applying Sequential Learners on the data split to train and test using both corpus-based and our features.

These results show that all the feature sets are much better at detecting the beginning of the stimulus span and detecting negative tokens rather than the middle and end of emotion stimulus span. They also show that our features substantially improve the result of detecting beginning, the middle, and the end of the stimulus span compare to corpus-based features, while they are slightly worse in detecting negative tokens. Last but not least, we found detecting the end of an emotion stimulus more challenging and we believe it needs more attention for improvement in the future.

5.6 Summary

In this chapter, we framed the detection of causes of emotions as finding a stimulus frame element defined for the emotion frame in FrameNet. We have created the first ever dataset

annotated with both emotion stimulus and emotion statement; it can be used for evaluation or training purposes. We used FrameNet's annotated data for 173 emotion lexical units, grouped the lexical units into seven basic emotions using their synonyms and built a dataset annotated with both the emotion stimulus and the emotion. We applied sequential learning methods to the dataset. We explored syntactic and semantic features in addition to corpus-based features. We built 1) a model that detects whether the sentence contains an emotion stimulus and 2) an information extraction model that detects the emotion stimulus span. They both outperforms all our carefully built baselines.

We believe that an emotion stimulus and the emotion itself are not mutually independent. Although in this chapter we did not take the relation of emotion expression and emotion stimulus of the sentences into account, in the next chapter we would like to address the relation of emotion expression and emotion stimulus at the same time. We will use the two models that were built in this chapter to go deeper on the analysis of Aman's dataset that was used in Chapter 4. We also would like to investigate whether an emotion stimulus can help with detecting the emotion expressed in sentences where there is no explicit mention of the emotion felt.

Chapter 6

Emotion expression and emotion stimuli

An emotion stimulus is not independant of the emotion expressed in response to that stimulus. In this chapter we would like to address emotion stimulus in regard to the emotion that is expressed in the sentence. We will use the two models that were built in Chapter 5: detecting whether an emotion-bearing sentence contains the stimulus of that emotion; and extracting the emotion stimulus span. They will allow us to go deeper into the emotion analysis of Aman's dataset that was used in Chapter 4. We also investigate the differences and challenges of emotion stimulus detection when the emotion expression is explicitly mentioned versus when the emotion expression is implicit.[1]

In combining emotion expression detection with emotion stimuli there are four possibilities:

1. Group 1: Both emotion expression and emotion stimulus appear in the text. For example, the sentence "I feel happy that I saw you after so many years." contains both an explicit mention of the word *happy* and the emotion stimulus "I saw you after so many years". In this case, we first detect the emotion expression and then define whether there is an emotion stimulus. If so, we go further and detect the emotion stimulus span. If there is no emotion stimulus indicated for the emotion expressed, that will mean it is an instance of the second group.

[1]In this context, *explicit* means it is clearly stated in the text while *implicit* means it is implied from the context but it is not readily apparent or visible.

2. Group 2: There is only an emotion expression without any explicit emotion stimulus related to the emotion expression. For example, the sentence "He sounded quite sad." only contains the emotion expressed (sad) with no explicit mention of emotion stimulus related to the emotion expression. We apply the model that we built in Chapter 5, so we indicate whether an instance contains an emotion stimulus or not. The instances that do not contain an emotion stimulus belong to this group.

3. Group 3: The third group, which is the most challenging group, is that there is only emotion stimulus in the text while the emotion expression is absent. For example, in the sentence "I passed my exam." there is no explicit mention of the emotion felt while the sentence contains the emotion stimulus that normally evokes the *happy* emotion. To address this group, among the instances that get labelled with emotion stimulus, we separate the sentences with implicit emotion expression from the sentences with explicit emotion expression and we study them further.

4. Group 4: Finally, neither emotion expression nor emotion stimulus exists in the text, which is not in the scope of our work, so we simply ignore this group.

In this chapter, we use Aman's dataset that was explained in Chapter 4, except that we discard the sentences tagged as no-emotion (Group 4). That is because we are looking for emotion stimulus, so when a sentence carries no emotion, emotion stimulus detection is no more valid.

Finally, we evaluate the result of the emotion stimulus span detection model on both explicit and implicit data. The reason is that we believe that explicit and implicit data have different characteristics and addressing them separately expresses more on the benefits and shortcomings of our model.

Next, we explain the process of splitting the dataset into *explicit* and *implicit* instances by using different list of emotion keywords. After applying the stimulus extraction models to the data, we do further analysis based on the above-mentioned groups that each instance can belong to.

6.1 Data with explicit and implicit emotion expressions

In our dataset, we define as *explicit* those instances that explicitly contain the emotion expression. We use a list of emotion keywords to group the instances which contain at least one emotion keyword as explicit data. On the other hand, an instance is considered *implicit* if it does not contain any emotion keyword.

Here, we use two different lexicons to split the data into instances with explicit or implicit emotion expression:

- WordNet-Affect which is a list of 1116 words grouped into six emotion classes explained in section 3.3.1.
- The list of synonym lexical units that we built in section 5.3 for each emotion class.

The result of splitting Aman's datasets into Explicit and Implicit sets using these two lists is displayed in Table 6.1. We also indicate the average length of each part of the dataset, which shows that instances with implicit emotion expression tend to be much shorter than instances with explicit emotion expression.

In the next section, we explain the result of emotion stimulus detection on this dataset with breakdown on different parts of the dataset, whether they are explicit, implicit or strongly explicit instances.[2]

6.2 Emotion stimulus results

The first part of the experiments is to indicate whether there is an emotion stimulus in the sentence. Here we use the model that we built in section 5.5.1. As a result of applying this model to the whole dataset, 374 sentences have been labelled as containing the emotion stimulus (Group 1 and 3) and 242 sentences as not containing emotion stimulus (Group 2).

[2]Strongly explicit instances contain at least one emotion keyword from a small carefully built list of synonyms for each emotion class.

		Explicit	Average-Length	Implicit	Average-Length
WordNet	Happiness	248	85	288	56
	Fear	67	125	48	78
	Anger	89	110	90	80
	Sadness	85	87	88	65
	Disgust	84	93	88	60
	Surprise	45	105	70	68
	total	618	100	672	68
FrameNet	Happiness	40	90	496	68
	Fear	21	105	94	105
	Anger	36	107	145	92
	Sadness	25	91	148	73
	Disgust	5	87	166	76
	Surprise	6	132	108	80
	total	133	102	1157	82

Table 6.1: The distribution of sentences with explicit emotion expression and sentences with implicit emotion expression and their average length in terms using two different emotion lexicons.

For the rest of the data, our model cannot confidently classify them into one of the classes. The group of sentences that belong to group 2 were addressed in chapter 4; they are ignored in this chapter, since they do not contain any emotion stimulus. The distribution of the emotion classes of the results is presented in Table 6.2. This is not surprising considering that 21.93% of the dataset are very short sentences.[3] Also the blog data that were labelled with the emotion expressed were not built for this purpose and many instances might naturally not carry emotion stimulus.

Now that the sentences which contain emotion stimulus have been defined, we apply the emotion stimulus span detection to those sentences which results in 300 emotion stimulus spans extracted. Among those 300 spans, 211 belong to sentences with an explicit emotion

[3]We define short sentences as sentences with five or fewer tokens.

	hp	sd	ag	dg	sr	fr	total
Has-Stimulus	155	59	40	53	24	43	374
No-Stimulus	92	32	33	39	29	17	242

Table 6.2: Distribution of sentences with emotion stimulus versus sentences with no emotion stimulus in Aman's dataset; the labels are happiness, sadness, anger, disgust, surprise, fear.

expression (Group 1) and 89 spans belong to sentences withan implicit emotion expression (Group 3). The distribution of the emotion classes of the results based on the sentences with explicit or implicit emotion expression is presented in Table 6.3.

The result shows that 70.33% of the emotion stimulus spans that are detected are from sentences that contain explicit emotion keywords. That can be due to various reasons:

1. As mentioned, part of the sentences that are categorized as sentences with implicit emotion are too short to contain an emotion stimulus.

2. The FrameNet data that our model is built on contain sentences with explicit emotion expressions. Although the model that we built was not based on the emotion expression, and the emotion expression is only one of many features that consider syntactic and semantic structure of the sentence, still our model is more biased towards sentences with more or less the same structure and detects more emotion stimulus spans in sentences with explicit emotion.

3. Generally, detecting emotion stimulus in sentences with an implicit emotion expression is harder due to the structure of these sentences and the fact that they do not contain clues such as the emotion keyword. For example, in the sentence "Octoberfest was awesome" the emotion stimulus is at the beginning of the sentence while there are not many examples in sentences with explicit emotion where the emotion stimulus occurs at the beginning of the sentence. Also in the sentence "Had a lovely birthday yesterday with Alex and Christine." the whole sentence is the emotion stimulus. This makes it hard for our model to detect the stimulus span because it has not seen a similar case in the training data to learn from.

	hp	sd	ag	dg	sr	fr	total
WNAffect Explicit	75	27	39	24	12	34	211
WNAffect Implicit	22	12	22	12	10	10	89

Table 6.3: Distribution of emotion stimulus span in Aman's explicit and implicit dataset; the labels are happiness, sadness, anger, disgust, surprise, fear.

4. The experiencer of the emotion also plays an important role in detecting the emotion stimulus span. Aman's dataset is built in such a way that the experiencer of the emotion is not necessarily the writer of the sentence. Therefore detecting the stimulus span can be very tricky. For example in the sentence "My dad's birthday was a blast, he got lots of fabulous gifts." the stimulus can be both "My dad's birthday" or "lots of fabulous gifts" based on the experiencer of the emotion. This makes detecting the emotion experiencer an interesting future work in the field of emotion detection.

Here are two examples of the emotion stimulus detected as a result of applying our emotion stimulus detector:

- I love <stimulus>my nap and story time with booboo every afternoon.</stimulus>
- It was wonderful to <stimulus> be away from home</stimulus> but it is indeed wonderful to be back.

Although the results are correct, in the second sentence we see there are two stimulus spans in the same sentence where "to be back" is the second stimulus span. However, our model only detects the first one, which can be due to the fact that it is built on a dataset that only contains one stimulus per sentence. That again shows that supervised machine learning model is not independent of the training data. Therefore, there are two directions of future work. One is to extend the training dataset with samples that have larger variety of structures, *e.g.*, containing data with both explicit and implicit expression of emotion. The second future direction can be to explore more semantic features, which are less dependent on the structure of the sentence.

6.3 Emotion stimulus evaluation

To evaluate the result of the 300 emotion stimulus that were detected, we asked two human annotators to label each emotion stimulus spans with one of three labels, namely, "correct", "not correct", and "partially-correct".[4]

The inter-annotator agreement was calculated using the kappa measure (Cohen, 1960). Cohen's kappa is popularly used to compare the extent of consensus between two raters in classifying items into known mutually exclusive categories. The average inter-annotator agreement was 0.659, which is considered good. We also consider two levels of agreement, strong agreement and weak agreement. Strong agreement shows the percentage of instances that both annotators labelled as correct or non-correct, which results in 62.33% agreement. For weak agreement correct and partially-correct are considered the same, so it is the percentage of instances that their labels exactly match or one annotator labelled it as partially-correct while the other as correct; this results in 78.00% agreement. Although there is a good agreement among the annotators according to the kappa measure, we found weak agreement is significantly higher than strong agreement, which shows the difficulty of detecting the exact emotion stimulus span even for human annotators.

Another interesting observation from the human annotation result was that for some sentences the annotators did not quite agree with the emotion label assigned to the sentence, which made it harder for them to agree or disagree with the emotion stimulus label of those sentences. For example, the sentence "Last night my Mom asked me if I am scared of flying and going on our trip." is labelled with the *fear* emotion; although the annotators disagree with the emotion label, one of the annotators indicated if "Scared" is considered the emotion expression, "flying and going on our trip" will be its corresponding emotion stimulus. This emphasizes that the emotion expressed in a sentence and the emotion's stimulus are very dependent on each other. So, knowing the emotion expressed helps in detecting emotion

[4]One of the annotators is a native English speaker who has completed graduate level education. The second annotator is pursuing her postgraduate studies and has completed a Master's degree in English language. She is fluent in English and her competency is confirmed by International English Language Testing Systems exam.

stimulus and vice versa.

We also went further and investigated whether it would be possible to define the emotion expressed only by looking at the emotion stimulus. We found two major types of stimulus, "person" and "events", which align with the emotion stimulus definition from FrameNet. In the cases when the stimulus is a person it is not possible to recognize the emotion expressed just by looking at the stimulus. For example, in the sentence "I am so proud of my sister" the stimulus is "my sister" and the emotion expressed is *happy*; it is hard to indicate what emotion was expressed only by looking at the stimulus.

On the other hand, in the sentences "I was scared to see him with a knife." the emotion stimulus is "see him with a knife", which is an event. When the emotion stimulus is an event or a state of affairs, or it contains the explanation of the stimulus that caused that emotion, the emotion stimulus can be an indicator of the emotion expressed. This is the case regardless of whether the emotion is explicitly expressed or it is implicitly conveyed in the sentence. For example, if this sentence were "I saw him with a knife." instead, we could still imply that the emotion expressed in this sentence is "fear" due to the emotion stimulus which has not changed which is "saw him with a knife". This makes us believe that using emotion stimulus to detect the emotion expressed is an interesting direction of research that we think can benefit from our work.

There also are some sentences in which the emotion stimulus is hidden in the meaning of the sentence. For example, consider the sentence "How about you ignore us, we'll ignore you, you all can go to heaven and we can all be infidels and suffer through our evil." From the context we can tell that the experiencer is angry due to "conflict in beliefs with a religious person/people". However there is no span of the text that we can choose as the emotion stimulus span and the *anger* feeling is more expressed in a sarcastic way. These are among more complicated instances that need further investigation in the future.

Finally, we report two sets of evaluation as a result of the emotion stimulus annotation by two human annotators, namely, exact-match and partial-match accuracy. *Exact-match* only considers the instances that are labelled by both annotators as correct, while *partial-match* treats partially-correct labels the same as correct. We consider as a true positive any

	hp	sd	ag	dg	sr	fr	all
Exact-Match	43.87	69.23	40.98	50.0	45.45	43.18	47.33
Partial-Match	61.22	84.6	57.37	75.0	68.18	72.7	69.84

Table 6.4: The result of extracting emotion stimulus in Aman's dataset; the labels are happiness, sadness, anger, disgust, surprise, fear.

instance that one of the annotators has labelled correct and the other annotator has labelled as partially-correct, or that they have both labelled as partially-correct or correct. The result of the emotion stimulus evaluation for each emotion class is displayed in Table 6.4.

The results show that the partial-match accuracy is significantly better than the exact match as it relaxed the requirement of exact span boundary matching. This is not surprising because it was also hard to reach consensus among human annotators on the exact span of the stimulus.

We also found that data played an important role. For example among the 133 instances that we grouped as strongly explicit instances,[5] 91 sentences are classified as sentences that contain emotion stimulus with 61.56% exact-match accuracy and 73.67% partial-match accuracy; this shows the exact-match accuracy result of these instances is substantially higher than the result of all instances. This confirms that the nature of the dataset with 52.09% implicit and 21.93% very short sentences makes it very challenging for our task.

On the other hand, the result of emotion stimulus detection on the instances with explicit emotion expression using WordNet-Affect shows that the accuracy of the emotion stimulus extraction in the first group is 47.86% (70.14% partial-match accuracy) and the accuracy of emotion stimulus extracted in sentences with implicit emotion expression is 44.94% (68.53% partial-match accuracy). Interestingly, we do not see a significant difference between the accuracy of sentences with an explicit emotion expression versus sentences with an implicit emotion expression. Although in the first phase, which classified sentences into two classes of containing and not containing a stimulus, being a sentence with explicit emotion expres-

[5]Those are the instances that contain at least one emotion keyword from the list of synonym lexical units of emotion frame in FrameNet which we presented in section 5.3.

sion versus implicit emotion expression played an important role, in the second phase, which extracts the emotion stimulus span, this does not make a big difference. Therefore, if we improve the first phase to recall more instances that contain emotion stimulus particularly for sentences with implicit emotion expression our methodology is very promising. It can particularly benefit the emotion detection of instances with implicit emotion expression because the quality of the stimulus spans detected in them is as good as instances with explicit emotion expression.

6.4 Summary

In this chapter, we discussed the relation of emotion stimulus and emotion expression for both sentences with explicit and implicit expression of emotions; therefore we focus on two groups of sentences: a) sentences that contain both emotion expression and emotion stimulus, b) sentences that only contain emotion stimulus with no explicit mention of emotion expression.

We applied the models that were built in a previous chapter on Aman's dataset. We first defined whether a sentence contains an emotion stimulus or not. Next, we detected the emotion stimulus span in sentences classified as containing emotion stimulus. We discussed the difference of the structure of our training dataset with the testing dataset which caused our model to only classify about half of the sentences as whether they contained emotion stimulus or not. However, the results of the emotion stimulus spans detection which were evaluated by two human annotators were quite promising. We reported the evaluation result of sentences with explicit emotion expression separately, then the sentences with implicit emotion expression. Our results showed that, although our model retrieves more stimulus spans in the sentences with an explicit emotion expression, the accuracy of the spans detected is equally good for both group of sentences with or without an explicit emotion expression.

We also showcased how an emotion stimulus that is an event or a state of affairs, or contains the explanation of the stimulus that caused the emotion can indicate the emotion expressed. This is particularly beneficial in sentences with implicit emotion expression, because most of the research performed in the field of emotion analysis depends on using emotion

keywords, therefore only addressing sentences with explicit emotion expression (*e.g.*, emotion keyword spotting using WordNet-Affect lexicon or NRC Emotion lexicon, using the emotion keywords as a feature in machine learning methods, or using the emotion keyword and its syntactic and semantic relations to the context around it to detect the contextual emotion). Nonetheless, an emotion is not always expressed through the use of emotion-bearing keywords. According to a linguistic survey by Pennebaker and Francis (2001), only 4% of the words used in written texts carry affective content. So, many emotional statements with no emotion keyword can be ignored in current research.

To the best of our knowledge, the little work on detecting implicit expressions of emotion which have either used common-sense knowledge-base (Balahur et al., 2011b) or applied rule-based approach (Udochukwu and He, 2015) suffers from low coverage and domain-specificity.

Therefore, we believe detection of emotions that are not explicitly stated in the text (that has received very little attention to date) is the next main challenge of emotion detection area, which needs deeper understanding of text. In this chapter we showed that detecting emotion stimulus is not only important in detecting why a particular emotion is felt but it can also be beneficial in detecting emotions that are implicitly expressed.

As a future work, we would like to extend the emotion stimulus training dataset to retrieve more emotion stimulus spans specially in sentences with implicit emotion expression. So, the emotion stimulus spans will be used to indicate the implicitly stated emotion.

Chapter 7

Conclusion

This thesis addressed the task of automatic classification of texts by the emotions expressed, as well as the detection of emotion stimuli that defines what provokes the emotion.

The presentation began with an introduction of the problem and an explanation of our motivation for addressing this problem. We explained how emotions play an important role both as the causes and the symptoms of depression; therefore our research can benefit health care agencies; or on behalf of individuals, enabling those suffering from depression to be more proactive about their mental health by detecting the emotion they have expressed and explaining why the particular emotions were felt.

Next, we studied psycholinguistically- and linguistically-motivated features that are important in recognizing emotions. We argued that there is a large gap between theoretical research on emotions and emotion studies in computational linguistics. We addressed psycholinguistic aspects of emotions and explained how they can be useful in computational approaches to emotions.

The rest of the work was organized in two parts, emotion expression and emotion stimulus.

In the first part, we explained the methodologies and datasets used in detecting the emotion expressed. We discussed the existing methodologies that lack considering syntactic and structural relations in detecting emotion. We claimed that lexical and corpus-based methods are not sufficient for automatic discovery of the emotion expressed in a sentence. We studied linguistically-motivated features that are important in recognizing emotions. We

explained the necessity of having more sophisticated syntactic and semantic-based methods. We described the features chosen for characterizing emotional content based on syntactic and semantic structure of sentences as well as the machine learning techniques adopted for emotion classification.

In the second part, we looked at the problem of detecting emotion stimuli from both the theoretical and computational perspective. We mentioned the dearth of research in the emotion stimulus detection area and the shortcomings of research in closely related areas such as event-based emotion recognition. We also suggested statistical methods to avoid the biases and lack of coverage of manual rule and pattern detection. That is why we created a new emotion dataset annotated with emotion stimulus in addition to the emotion itself.

To create the dataset, we used FrameNet's annotated data for 173 emotion lexical units, grouped the lexical units into seven basic emotions using their synonyms. We applied sequential learning methods to the dataset using syntactic and semantic features in addition to corpus-based features. We built 1) a model that detects whether the sentence contains an emotion stimulus and 2) an information extraction model that detects the emotion stimulus span. They both outperformed all our carefully built baselines.

Next, we addressed the relation of emotion expression and emotion stimulus at the same time. We used the two models that were built to go deeper into the analysis of the dataset that we had used to detect the emotion that was expressed. We also studied the behaviour of sentences with explicit or implicit expression of the emotion separately. We argued that sentences with implicit emotion expression need more attention. We investigated whether an emotion stimulus can help with detecting the emotion expressed in sentences where there is no explicit mention of the emotion felt and we showed that emotion stimulus can be a good indicator of an implicitly stated emotion.

Next, we would like to highlight our main contribution as a result of this thesis.

7.1 Contributions

We would like to discuss how we lifted some of the limitations in the field of emotion detection that were explained in section 1.2.

In our research we tried to bridge the gap between the theoretical research on psychology of emotions and emotion studies in computational linguistics by discussing emotions and their specifications from both perspectives. As a result, we applied deeper methods considering psycholinguistically- and linguistically-motivated features in the text. We also discussed how emotions have their own challenges and characteristics, which makes emotion analysis an independent research area rather than a sub-area of polarity detection in sentiment analysis.

Most of the prior work on emotions in text is limited to the available datasets to the point that the limitation in resources has also affected the direction of the research on emotion recognition. In this thesis, we went beyond the existing datasets and resources. We built our own dataset that is labelled with both the emotion expressed and the emotion stimulus using FrameNet's manually labelled data for emotion lexicons.

Using the dataset we built, we studied emotion stimulus. It is quite a new field in emotion analysis – there has been very little attempt, if any, to determine what causes the emotions. We believe that going beyond detecting emotions only and addressing emotion stimuli builds a framework for more thorough and deeper analysis of emotions.

We also studied the relation of emotion stimulus and emotion expression for both sentences with explicit and implicit expression of emotions. Our results showed that the accuracy of the spans detected is equally good for both group of sentences with or without explicit emotion expression. We also showcased how an emotion stimulus that is an event or a state of affairs, or contains the explanation of the stimulus that caused the emotion can indicate the emotion expressed. This is particularly beneficial when the emotion expression is implicit.

To explain how our research can be beneficial, we mentioned the area of mental health and the usefulness for automatic analysis of self-reported text from patients. We expressed our interest particularly in depression and suicide, the leading causes of death among people of all ages. We explained that suicide notes and depression-indicating data are highly emotionally

intensive. That makes it intriguing to apply our research to them, not only to detect the emotions expressed but also to explain the reason behind them. To the best of our knowledge, the suicide notes dataset collected for an i2b2 shared task of 2011,[1] and the depression indicating data collected by Choudhury et al. (2013) are not publicly available due to the sensitivity of the data. However, as an interesting possible application of our work we would like to elaborate on it.

7.2 Application

Suicide is a significant health problem, yet many questions regarding suicide remain unanswered. One of the most frequently asked questions is related to motive: "Why did that person commit suicide?" Motivations for committing suicide is explored in psychology and psychiatry by studying suicide notes.

It is estimated that 2530% of suicides are accompanied by a note. According to Olson et al. (2011), the most common reasons why people contemplating suicide choose to write a suicide note include one or more of the following:

- To ease the pain of those known to the victim by attempting to dissipate guilt.
- To increase the pain of survivors by attempting to create guilt.
- To set out the reason(s) for suicide.
- To express thoughts and feelings that the person felt unable to express in life.
- To give instructions for disposal of the remains.
- Occasionally, to confess acts of murder or some other offence.

Considering the reasons why people choose to write suicide notes, there are two major type of data in those notes. This is also reflected in the dataset that was built for NIH shared tash. The first group consists of factual data such as *information* on the properties and belongings and *instructions* to ask other people to do something. An example of information is "my cash is up under the books." while an instruction would be "Say goodbye to her for me."

[1] `https://www.i2b2.org/NLP/Coreference/Call.php`

The second group is emotional context that contains feelings and emotions expressed, as well as the reason for committing suicide. This is equivalent to emotion stimulus addressed in our research. For example, in the sentence "I leave the curse of God to Jane Johnson who has lied about my wife and caused all this misery.", the emotion expressed is *anger* and the stimulus is "Jane Johnson" with "who has lied about my wife" as an explanation of the stimulus. Another example is "I know he don't love me any more and wants to keep hurting me." which also indicates *sadness*. The stimulus here is "he don't love me any more".

The second group is the main group that was targeted in our research and our work can be applied to detect the emotion expressed as well as the emotion stimulus. Even so, we believe our methodology in classifying sentences into having stimulus versus not having stimulus can be beneficial in separating the first group from the second group. Our work can help identify instances with implicit emotion expression versus instances that just indicate information or instruction by classifying them into has-stimulus and no-stimulus classes.

To apply our work in this application, as mentioned in the previous section, the very first problem to overcome is lack of accessible public dataset due to the sensitivity of the data. So one necessary first future step is to collect suicide notes publicly available on the Web.

Assuming that we have collected the data, there will be several steps.

- We will label part of the data with information, instruction, emotions and their stimuli for evaluation purposes.
- We will split the data into those sentences which do and those which do not indicate an explicit emotion expression.
- To the sentences with an explicit emotion expression, we will apply our models built in chapter 4 to detect the contextual emotion expressed and next apply the models built in chapter 5 to detect the emotion stimulus if there is one in the sentence.
- To the sentences with no explicit emotion expression, we will first apply the model build in chapter 5 to define if there is a stimulus in the sentence or not. If there is no stimulus, we assume that data is information, instruction or neither emotional nor those two categories. If there is a stimulus, we will fist define the emotion stimulus span using the second model built in chapter 5, next use the emotion stimulus as one

of the main features to detect the emotion expressed.

- Finally, we would like to evaluate our results against the labelled data. Olson et al. (2011) explored motivations for committing suicide, especially with regard to cultural differences, by analyzing suicide notes written by Native Americans, Hispanics and Anglos in New Mexico. Five categories emerged describing motivation: (a) a feeling of alienation, (b) a feeling of failure or inadequacy, (c) a feeling of being psychologically overwhelmed,; (d) the desire to leave problems behind, and (e) reunification in an afterlife. We found it would be very interesting to group the automatic emotion stimulus detected by our model and verify whether they match the clinical study by Olson et al. on the motivations for committing suicide.

Here, we explained an application and how it can benefit from our research. Next, we would like to discuss future work.

7.3 Future work

The work presented in this thesis can be pursued further in several directions.

We were limited by the specification of the datasets that were used, so we based our work on sentence-level emotion and emotion stimulus detection. However, in Chapter 5, we still considered discourse structure of the text. It is an important feature in longer text particularly when taking adjacent sentences into account. We believe it is necessary to consider adjacent sentences especially in the case of emotion stimulus detection. Naturally there might be cases when the emotion is expressed in one sentence and the stimulus is indicated in an adjacent sentence. Therefore, extending our work to consider adjacent sentences to detect emotion stimulus is an interesting future direction.

The emotion stimulus dataset we built in this work is a well-formed dataset that contains carefully built data annotated by humans. The data also contain sentences with explicit expression of the emotion, which makes it not ideal in detecting emotion stimulus in sentences with implicit emotion expression. To study the emotion stimulus problem thoroughly, we would like in the future to extend our dataset by using it as seed samples and apply semi-

supervised bootstrapping methods to add instances of other existing emotion datasets which do not have emotion stimulus labels.

This work is one of the very first research projects on detecting emotion stimulus. We believe, therefore, that there is still some room for further improvement of emotion stimulus detection. Here are some of the shortcomings that need further attention:

- Applying coreference resolution, particularly important in cases when the stimulus is an agent; returning a pronoun as an emotion stimulus without coreference resolution will not be as informative.
- Addressing cases where there is more than one emotion expressed.
- Using other semantically motivated features such as agent-specific features as well as semantic role labels to make our models less structure-dependent and more semantic-driven.
- According to FrameNet, a stimulus is the person, event or state of affairs which evokes the emotional response in the Experiencer. We believe whether the stimulus is a person, event or state of affairs makes a difference to their specification and consequently in the methodology to detect them; therefore addressing them separately based on their own specifications could improve emotion stimulus detection.

The dataset we created for emotion stimulus is also labelled with *Experiencer*, one of the main frame elements in which research community has shown interest. This dataset allows addressing experiences, emotion expressed and the emotion stimulus at the same time using joint classification models, which not only makes the analysis more thorough and detailed but also provides a possibility to connect each emotion expressed with its stimulus and experiencer. This is particularly interesting in cases where there is more than one emotion expressed in the text, which makes it a valuable future path to take.

Last but not least, we believe that the detection of emotions not explicitly stated in the text is the next main challenge of the emotion detection area, which needs further attention. We showed that detecting emotion stimulus is not only important in determining why a particular emotion is felt, but it can also be beneficial in detecting emotions that

are implicitly expressed. However, our preliminary results show that we need to improve the emotion stimulus recall in sentences with implicit emotion expression. All in all, using emotion stimulus to detect implicitly expressed emotions is a promising future direction.

Bibliography

Inc. Alias-i. Lingpipe 4.1.0., October 2008. URL `http://alias-i.com/lingpipe`.

Cecilia Ovesdotter Alm. The Role of Affect in the Computational Modeling of Natural Language. *Language and Linguistics Compass*, 6(7):416–430, 2012.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from Text: Machine Learning for Text-based Emotion Prediction. In *HLT/EMNLP*, pages 347–354, 2005.

Ebba Cecilia Ovesdotter Alm. Affect in text and speech, 2008.

Saima Aman and Stan Szpakowicz. Identifying Expressions of Emotion in Text. In *Proceedings of 10th International Conf. Text, Speech and Dialogue*, pages 196–205. Springer-Verlag, 2007.

Saima Aman and Stan Szpakowicz. Using Roget's Thesaurus for Fine-grained Emotion Recognition. In *Proceedings of Third International Joint Conf. on Natural Language Processing IJCNLP 2008*, pages 296–302, 2008.

Robert A. Amsler. Lexical Knowledge Bases. In Yorick Wilks, editor, *COLING*, pages 458–459. ACL, 1984.

Alexandra Balahur, Jesus M. Hermida, and Andres Montoyo. Detecting Emotions in Social Affective Situations Using the EmotiNet Knowledge Base. In *ISNN (3)*, pages 611–620. Springer, 2011a.

Alexandra Balahur, Jesús M. Hermida, and Andrés Montoyo. Detecting Implicit Expressions of Sentiment in Text Based on Commonsense Knowledge. In *Proceedings of the 2Nd Work-*

shop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA '11, pages 53–60, Stroudsburg, PA, USA, 2011b. Association for Computational Linguistics.

Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo, and Rafael Muñoz. EmotiNet: A Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories. In *Proceedings of the 16th international conference on Natural language processing and information systems*, pages 27–39, Berlin, Heidelberg, 2011c. Springer-Verlag.

Steven Bethard and James H Martin. Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations. In *Proceedings of ACL08 HLT Short Papers*, pages 177–180. Association for Computational Linguistics, 2008.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic Extraction of Opinion Propositions and their Holders. In *In 2004 AAAI spring symposium on exploring attitude and effect in text*, pages 22–24, 2004.

Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl. Sentic Computing: Exploitation of Common Sense for the Development of Emotion-Sensitive Systems. In *COST 2102 Training School*, pages 148–156. Springer, 2010.

Du-Seong Chang and Key-Sun Choi. Incremental Cue Phrase Learning and Bootstrapping Method for Causality Extraction using Cue Phrase and Word Pair Probabilities. *Information Processing and Management*, 42(3):662–678, 2006.

Francois-Régis Chaumartin. UPAR7: A Knowledge-based System for Headline Sentiment Tagging. In *Proceedings of 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 422–425, 2007.

Ping Chen, Wei Ding, and Chengmin Ding. SenseNet: A Knowledge Representation Model for Computational Semantics. In *IEEE ICCI*, pages 434–439. IEEE, 2006.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. Emotion Cause Detection with Linguistic Constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 179–187, 2010.

Timothy Chklovski and Patrick Pantel. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, 2004. Association for Computational Linguistics.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 355–362, 2005.

Yejin Choi, Eric Breck, and Claire Cardie. Joint Extraction of Entities and Relations for Opinion Recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 431–439, 2006.

Munmun De Choudhury, Scott Counts, Eric Horvitz, and Michael Gamon. Predicting Depression via Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. AAAI, 2013.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

William W. Cohen. Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data. http://minorthird.sourceforge.net, 2004.

Phoebe C. Ellsworth Craig A. Smith. Patterns of Cognitive Appraisal in Emotion. *Journal of Personality and Social Psychology*, 48(4):813–38, 1985.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. Frame-Semantic Parsing. *Computational Linguistics*, 40(1):9–56, 2014.

Bart Desmet and Veronique Hoste. Emotion Detection In Suicide Notes. *Expert Systems with Applications*, 40(16):6351–6358, 2013.

Roger Detels. *The Scope and Concerns of Public Health*. Oxford University Press, 2009.

Paul Ekman. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3):169–200, 1992.

Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of 5th Conf. on Language Resources and Evaluation LREC 2006*, pages 417–422, 2006.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Song Feng, Ritwik Banerjee, and Yejin Choi. Characterizing Stylistic Elements in Syntactic Structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1522–1533, 2012.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 940–949, 2014.

Charles J. Fillmore, Miriam R.L. Petruck, Josef Ruppenhofer, and Abby Wright. FrameNet in Action: The Case of Attaching. *IJL*, 16(3):297–332, September 2003.

Nico H. Frijda. Emotion, Cognitive Structure, and Action Tendency. *Cognition and Emotion*, 1(2):115–43, 1987.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Hierarchical Approach to Emotion Recognition and Classification in Texts. In *Canadian Conference on AI*, pages 40–50, 2010a.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Hierarchical Versus Flat Classification of Emotions in Text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 140–146, Stroudsburg, PA, USA, 2010b. Association for Computational Linguistics.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Prior versus Contextual Emotion of a Word in a Sentence. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 70–78, 2012.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Prior and Contextual Emotion of Words in Sentential Context. *Computer Speech & Language*, 28(1):76–92, 2014.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Detecting Emotion Stimuli in Emotion-Bearing Sentences. In *Computational Linguistics and Intelligent Text Processing*, pages 152–165. Springer International Publishing, 2015.

Daniel Gildea and Daniel Jurafsky. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288, 2002.

Roxana Girju. Automatic Detection of Causal Relations for Question Answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering - Volume 12*, MultiSumQA '03, pages 76–83, 2003.

Roxana Girju and Dan Moldovan. Mining Answers for Causation Questions. In *AAAI symposium on*, 2002.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11: 10–18, November 2009.

Hugo Hernault, Helmut Prendinger, David duVerle, and Mitsuru Ishizuka. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3), 2010.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. In *Proc. Tenth ACM SIGMOD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, 2004.

Carroll E. Izard. *The Face of Emotion.* New York: Appleton-Century-Crofts, 1971.

Mario Jarmasz and Stan Szpakowicz. Roget's thesaurus and semantic similarity. In *In: Proceedings of the RANLP-2003*, pages 212–219, 2003.

Randy M. Kaplan and Genevieve Berry-Rogghe. Knowledge-based Acquisition of Causal Relationships in Text. *Knowledge Acquisition*, 3(3):317–337, 1991.

Phil Katz, Matthew Singleton, and Richard Wicentowski. SWAT-MP: the SemEval-2007 Systems for Task 5 and Task 14. In *Proceedings of 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 308–313, 2007.

Alistair Kennedy and Diana Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22:2006, 2006.

Soo Min Kim and Eduard Hovy. Identifying and Analyzing Judgment Opinions. In *Proceedings of HLT/NAACL-2006*, pages 200–207, 2006.

Zornitsa Kozareva, Borja Navarro, Sonia Vázquez, and Andrés Montoyo. UA-ZBSA: A Headline Emotion Classification through Web Information. In *Proceedings of 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 334–337, 2007.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Richard S. Lazarus. *Emotion and Adaptation.* Oxford University Press, USA, 1991.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. A Text-driven Rule-based System for Emotion Cause Detection. In *Proceedings of the NAACL HLT 2010 Workshop on*

Computational Approaches to Analysis and Generation of Emotion in Text, CAAGET '10, pages 45–53, 2010a.

Sophia Yat Mei Lee, Ying Chen, Shoushan Li, and Chu-Ren Huang. Emotion Cause Events: Corpus Construction and Analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010b. European Language Resources Association (ELRA).

Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers, 2012.

Hugo Liu and Push Singh. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 22(4):211–226, 2004.

Cheng-Yu Lu, Shian-Hua Lin, Jen-Chang Liu, Samuel Cruz-Lara, and Jen-Shin Hong. Automatic Event-level Textual Emotion Sensing Using Mutual Action Histogram Between Entities. *Expert Syst. Appl.*, 37(2):1643–1653, 2010.

Marie-Catherine De Marneffe, Bill Maccartney, and Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proc. LREC 2006*, pages 449–454, 2006.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. Semantic Role Labeling of Emotions in Tweets. In *Proc. 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, 2014.

Saif M. Mohammad and Peter D. Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, 2010.

Alena Neviarouskaya and Masaki Aono. Extracting Causes of Emotions from Text. In *International Joint Conference on Natural Language Processing*, pages 932–936, 2013.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. AM: Textual Attitude Analysis Model. In *Proc. NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 80–88, 2010.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Affect Analysis Model: Novel Rule-based Approach to Affect Sensing from Text. *Natural Language Engineering*, 17(1):95–135, 2011.

Lenora M. Olson, Stephanie Wahab, Cheryl W. Thompson, and Lynne Durrant. Suicide Notes Among Native Americans, Hispanics, and Anglos. *Qualitative Health Research*, 21 (11):1484–1494, 2011.

Jimena Olveres, Mark Billinghurst, Jesus Savage, and Alistair Holden. Intelligent, Expressive Avatars. In *In Proc. of the WECC'98*, pages 47–55, 1998.

Andrew Ortony, Allan Collins, and Gerald L. Clore. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

Bo Pang, Lillian Lee, and Shivakumar 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, pages 79–86, 2002.

W. Gerrod Parrot. *Emotions in Social Psychology*. Psychology Press, 2001.

James W. Pennebaker and Martha E. Francis. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum, 2001.

John P. Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, Kevin B. Cohen, John Hurdle, and Christopher Brew. Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, 5:3–16, 01 2012.

Rosalind W. Picard. *Affective Computing*. The MIT Press, 1997.

John C. Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, 1998.

Robert Plutchik. *The Emotions*. New York: Random House, 1962.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

Ellen Riloff. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of 2003 Conf. on Empirical Methods in Natural Language Processing*, pages 105–112, 2003.

Stephanie S. Rude, Eva-Maria Gortnera, and James Pennebaker. Language Use of Depressed and Depression-vulnerable College Students. *Cognition and Emotion*, 18(8):1121–1133, 2004.

James A. Russell. Pancultural Aspects of the Human Conceptual Organization of Emotions. *Journal of Personality and Social Psychology*, 45(6):1281–1288, 1983.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. EMOCause: An Easy-adaptable Approach to Emotion Cause Contexts. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 153–160, Stroudsburg, PA, USA, 2011.

Klaus R. Scherer. *Criteria for Emotion-antecedent Appraisal: A Review*, pages 89–126. Kluwer, 1988.

Klaus R. Scherer. What Are Emotions? And How Can They be Measured? *Social Science Information*, 44:695–729, 2005.

Statistics Canada. Mortality, Summary List of Causes (2008), 2011. URL `http://www.statcan.gc.ca/pub/84f0209x/84f0209x2008000-eng.pdf`.

Statistics Canada. Suicide Rates: An Overview, 2012. URL `http://www.statcan.gc.ca/pub/82-624-x/2012001/article/11696-eng.pdf`.

Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. Multi-perspective Question Answering Using the OpQA Corpus. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 923–930, 2005.

Carlo Strapparava and Rada Mihalcea. SemEval-2007 Task 14: Affective Text. In *Proceedings of Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007.

Carlo Strapparava and Rada Mihalcea. Learning to Identify Emotions in Text. In *Proceedings of 2008 ACM symposium on Applied computing*, SAC '08, pages 1556–1560, 2008.

Carlo Strapparava and Alessandro Valitutti. WordNet-Affect: an Affective Extension of WordNet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, 2004.

Marc Sumner, Eibe Frank, and Mark A. Hall. Speeding Up Logistic Model Tree Induction. In *Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 675–683, 2005.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. Emotion Classification Using Massive Examples Extracted from the Web. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 881–888, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL*, pages 252–259, 2003.

Peter 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, pages 417–424, 2002.

Orizu Udochukwu and Yulan He. A Rule-Based Approach to Implicit Emotion Detection in

Text. In *Natural Language Processing and Information Systems*, volume 9103 of *Lecture Notes in Computer Science*, pages 197–203. Springer International Publishing, 2015.

Hanna M. Wallach. Conditional Random Fields: An Introduction. Technical report, University of Pennsylvania, 2004.

H. G. Wallbott and K. R. Scherer. How Universal and Specific is Emotional Experience? *Social Science Information*, 24:763–795, 1986.

Lisa Watson and Mark T. Spence. Causes and Consequences of Emotions on Consumer Behaviour: A Review and Integrative Cognitive Appraisal Theory. *European Journal of Marketing*, 41:487 – 511, 2007.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning Subjective Language. *Computational Linguistics*, 30(3):277–308, 2004.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *In Proceedings of HLT-EMNLP*, pages 347–354, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, 35(3):399–433, 2009.

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(2):165–183, 2006.

Yiming Yang and Xin Liu. A Re-examination of Text Categorization Methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 42–49, 1999.

Hong Yu and Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 129–136, 2003.