NAIST-IS-DD1461213

Doctoral Dissertation

Perspectives on the Making of Multiple Emotion Detection System in Text

Phan Duc Anh

January 20, 2019

Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of SCIENCE

Phan Duc Anh

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Satoshi Nakamura	(Co-supervisor
Associate Professor Masashi Shimbo	(Co-supervisor)
Assistant Professor Hiroyuki Shindo	(Co-supervisor)

Perspectives on the Making of Multiple Emotion Detection System in Text[*]

Phan Duc Anh

Abstract

Emotion detection in text, also known as affective computing in text refers to the use of natural language processing methods to recognize, interpret and simulate human emotions or affects. These emotions maybe the state of the author or the emotional effect intended by the author. Being able to interpret human emotions, the machine adapts itself better and produces appropriate behavior in response to those emotions. On the other hand, being able to simulate human emotions, the machine improves its communication ability and enriches interactivity between human and machine. Emotions in text may be expressed explicitly with emotional words, such as *happy* and *hate* or implicitly through the contexts. There has not been a method of emotion detection in text that can interpret with high accuracy and simultaneously many emotions without being heavily domain-dependent.

This dissertation studies emotion detection in text in two successive parts. The first part investigates both the linguistics and psychology theories behind the expression of emotions in text. The differences between emotion detection and sentiment analysis, opinion mining are discussed. We also investigate the properties of emotional text: subjectivity and objectivity, explicit and implicit expressions of emotions, affects - direct emotions of the author and intended emotional communication - emotional effect intended by the author. Lastly, we study various psychology theory about emotions and discuss what theory we shall use in our study.

[*] Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1461213, January 20, 2019.

The second part approaches emotion detection in text from an application perspective by taking advantages of the investigated theories, using natural language processing tools and machine learning techniques to produce emotion lexicon and model. The lexicon and model predict the multiple emotions that a piece of text may hold implicitly or explicitly. The result are evaluated against several baselines and methods.

This dissertation contributes to the field of Computation Linguistics by improving the state-of-the-art in Multiple Emotion Detection in text, discussing the psychology and linguistics theories behind emotion detection in text, annotating a semi-supervised corpus, proposing a framework for the task, building an emotional lexicon and a predicting model, and deepening our understanding of topic. We believe systems that try to interpret human emotions and adjust their behaviors accordingly will greatly benefits from our work.

Keywords:

Emotion Detection, Psychology, Linguistics annotation, semi-supervised corpus, machine learning, natural language processing, framework, affects, intended emotional communication

Contents

List of Figures

List of Tables

1. Introduction

This dissertation studies the making of emotion detection system in text. Particularly, it investigates what are emotions and the theories about them, how they are presented and how emotions are different from sentiments or opinions. We then make an emotion detection system to demonstrate our approach and verify the effectiveness of the system.

This chapter explains the basics of the study which includes: the motivation (Subsection 1.1), the definition of emotion detection in text (Section 1.2) and the methodology of the study (Section 1.3). We outlines our research goals and questions in Section 1.4 and the organization of the dissertation in Section 1.5.

1.1 Motivation

Emotion detection in text is relatively a more difficult task than in speech or video. Unlike the two other forms of media where sequences of acoustic or facial expression features are highly accurate indicators, emotions are usually expressed explicitly in text by some particular words and phrases, or implicitly and we have to deduct the emotions via the context. We consider the following examples:

Ex.1.1 I am angry now, stay away from me!

Ex.1.2 I could have wrung her neck.

Ex.1.3 My husband comes home late everyday. I have to do all the housework and take care of the children too. Does he think that I am happy to do all of those by myself?

While one may easily notice the emotion in **Ex.1.1** because of the word: *angry*, **Ex.1.2** is a little bit harder because we have to make the effort to relate the phrase *wring someone's neck* to the emotion. In this case, instead of directly using the word for the emotion, the speaker describes hypothetically the consequence action resulted by *Anger*. A simple emotion detection system that parse word-by-word may find the first example very simple and can overcome the second simply by having a dictionary of expressions. However, for **Ex.1.3**, most system will fail. None of the words in the example express or are related to the emotion *angry*, one can only speculate it by considering the whole context and paying attention to some grammatical clues: *have to* usually express unwillingness of the action and the rhetorical question in the last sentence.

According to Collier [1], the expression of emotion depends on not only the words being used, but also on the grammar structure and syntactic variables such as negations, embedded sentence, and the type of sentence - question, exclamation, command or statement. And lastly, as any other tasks in natural language processing, context also plays an important part in expressing implicitly the emotions. Identifying emotions in text is not a simple work that relies on picking up specific words that express emotions. While these words may play an important role, the whole piece of text may not hold any emotion because it is written in narrative style and therefore is objective. Words are also so proved unreliable when dealing with sarcasms, which evolves several negative emotions [2]. Therefore, in this paper, we would like to thoroughly investigate the way emotions are expressed in textual content and build an emotion detection system that takes into account all the factors that have been listed above.

With the tremendous progress in Natural language processing in recent years, machines have been able to mimic human conversation to some extent. Siri has always been a reliable assistant on Apple's devices. Many corporations and companies have implemented chatbots on Facebook Messenger to take charge of promoting campaign or suggesting products on tasks such as: British Airways' Emojibot, Ebay Shopbot or Pernod Ricard's Cocktail Coach. We believe that our research would greatly benefit human-machine interactions in such system as well as advancing our knowledge in Natural Language Processing in general and emotion detection in particular.

1.2 Definition of emotion detection in Text

This section is dedicated to giving the precise definition of emotion detection in text which will be referred to throughout the dissertation. It is of remarkable importance since there is often misunderstanding about the distinction between emotion detection and sentiment analysis or opinion mining. There are also debates on whether we should consider intended emotional communication as a target of emotion detection.

Ex.2.1 This movie is bad. I don't like it at all.

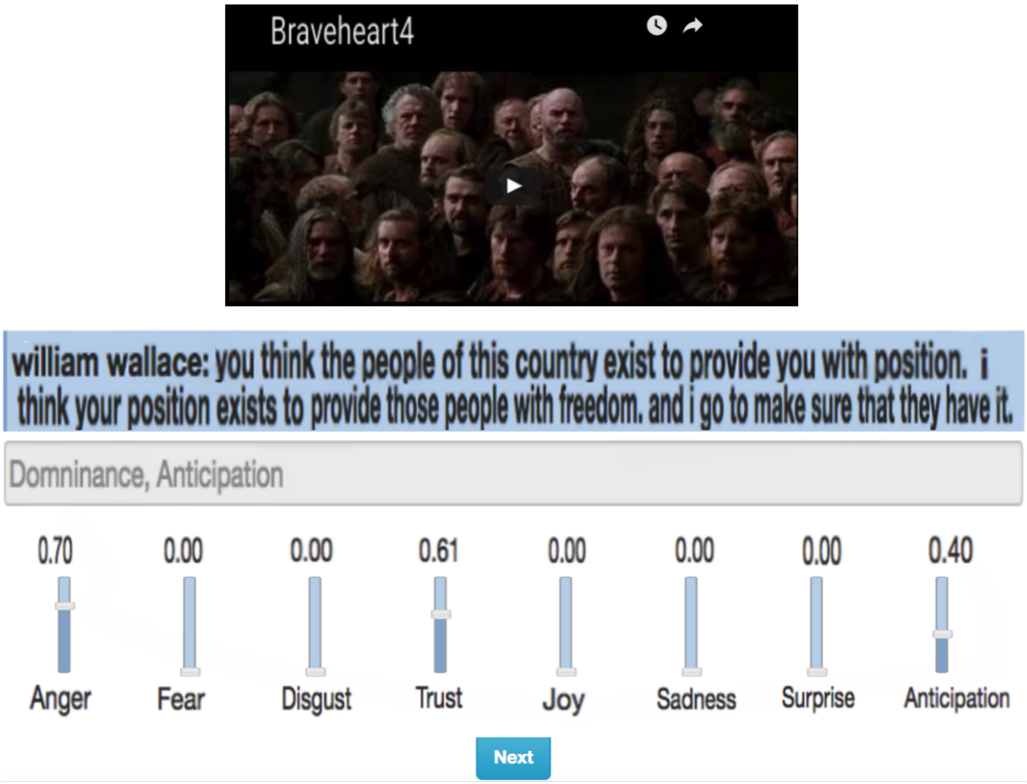
Ex.2.2 A girl was eating her lunch when two girl's jumped her. One girl took her by the hair and smashed her face into the table. Her nose at this point is pretty messed up. The second girl is behind the first girl and waits for the victim to fall down and stomps on victims face. All three girls were suspended for a week. Victim didn't even lay a finger on anyone. Zero tolerance rule in high school and colleges are messed up.

Let us examine examples by feeding them to each system and see the the result. In the first example **Ex.2.1**, the text is written at first person's perspective and the writter expresses his/her subjective opinion about a recent watched movie. Sentiment analysis (SA) systems will try to classify the text with negativity label. Opinion mining (OM) is essentially sentiment analysis which has a target entity [3], so the result of an OM system is *negative feeling towards entity* ***movie***. For an emotion detection (EA) system, the result would be *disappoint, dislike or unhappy* depends on what set of emotions the system adopts.

Moving the second example **Ex.2.2**, we can see that the story is told at a 3rd person's perspective. In most SA, OM or EA system, the text would not be considered for classification because of its objective manner. However, while the details in the text are mere facts, we see the intention of the writer when reporting them. The writer conveys the view on the matter and want the target audience to pick up on all the details to feel the same way as he does. This is called intended emotional communication. While most research would not deal with this type of text, we argue that it is an important part of an EA system and should be studied in a comprehensive manner.

In this dissertation, we define emotion detection to be *the task of using of natural language processing tools to identify and quantify emotional information of both direct emotions of the writer (affectives) and intended emotional communication (emotional effects on target audience by the writer)*. Therefore, in our research, unlike other works, we do not filter out examples that are objective. Instead, we consider all of the example and assume each one would hold emotions with some level of intensity.

To collect the data for our research, we have built an annotation website which have short clips and dialog transcripts from movies. In figure 1, we show the web



william wallace: you think the people of this country exist to provide you with position. i think your position exists to provide those people with freedom. and i go to make sure that they have it.

Domninance, Anticipation

Anger	Fear	Disgust	Trust	Joy	Sadness	Surprise	Anticipation
0.70	0.00	0.00	0.61	0.00	0.00	0.00	0.40

Figure 1. Annotation web interface: Annotators can choose the intensity of each basic emotion

interface of our annotation website. Annotators can choose the intensity of each basic emotions: Anger, Fear, Disgust, Trust, Joy, Sadness, Surprise, Anticipation. In the next chapter, we would go into details of the basic emotions and explore various theories about them.

1.3 Methodology

We examine the making of an emotional analysis system in text from both theoretical and application approach. The theoretical approach studies the nature of human emotions from a psychology perspective and the expression of emotion in textual data from a linguistics perspective. The application approach analyzes the hypotheses proposed by the theoretical approach and verify them by the mean of building model to predict emotions from textual data. This section outlines the methodology used in both approaches.

1.3.1 Theoretical Approach for emotion detection

Our work supports the idea that emotion detection in text is not merely a text classification problem that can easily be solved using natural language processing tools. We argue that traditional approaches would be limited of theirs use and would never cover the full scope of complex human emotions. Our proposal is that we approach the problem from psychology viewpoint, study the origin of human emotions and examine theories that can help us quantify emotions. Plutchik's theory of emotions [4, 5] is used throughout our work and it provides the scalability to our system to adapt to more domains and adopt more complex emotions.

Recent works by other researchers also employ Plutchik's theory rather than long used emotion theory by Paul Ekman [6] which was originated from facial emotion expression [7, 8, 9, 10, 11, 12, 13]. However, most of the previously mentioned approaches do not consider the existing of emotions in objective text as we see in example **Ex.2.2**. They would have their corpus to go through a subjective/ objective classifier to filter out objective text. This method may works for SA or OM task because the target of such tasks is the polarity. Meanwhile, the target of EA is the emotion and there are many ways and forms to express emotions. Even when some text does appear to be objective, It may carry an intention of the writer to have the audiences experience a specific feeling.

The expression of emotions in text is also studied. We would explore the way emotions are express through text via examples. At the end, we would summary the important clues that help us identify emotions in text.

1.3.2 Application Approach for emotion detection

An emotion lexicon and a predicting model deepen our knowledge and understanding in emotion detection and Natural language processing in general. In our work, we use existing lexicon [14, 15] and then build our own lexicon both bootstrapping method and word-embedding method. Experiments are carried to verify the effectiveness of the lexicon via Bag-of-Words approach.

We also annotate the Imdb quotes dataset [1] with emotions using multi-label scheme. One utterance (one turn in conversation) may hold more than one emotion. This is a feature that is more close to real life settings and adheres to the nature of human emotions. Our model would then be used on the annotated corpus in supervised and semi-supervised manner. We build two separate neural networks: one with careful feature selection and bootstrapped lexicon and the other with word-embedding lexicon. The results are compared to existing baselines and methods [16, 7, 9] to prove the effectiveness of our method. The application approach gives directions on how to analyze emotions in text with the help of machine learning techniques.

1.3.3 Procedures of the method

In this dissertation, we propose procedures for the detection of emotions in conversation:

1. Approaching emotion detection from a psychology perspective, we propose a fitting theory for quantification of emotions.

2. Approaching emotion detection from a linguistics view, we highlight important expression clues of emotion in text documents.

3. Building appropriate emotion corpus which is practical for real life application.

4. Building Emotion Lexicon using bootstrapping method on general domain and word-embedding method on target domain.

5. Creating models for Emotion Detection: manual feature selection model, word-embedding model. Comparing the performance supervised method and semi-supervised method of word-embedding model.

6. Verifying the performance of the system on different settings and reporting the result.

[1] `ftp://ftp.fu-berlin.de/pub/misc/movies/database/`

1.4 Goal of the research

1.4.1 The goal system

In this paper, our goal is to propose a comprehensive approach by exploiting Plutchik's theory which covers the full spectrum of human emotions to work on challenging multi-label conversation corpus. Our system is different from previous methods in four main ways:

- Plutchik's idea of basic emotion dyads and intensity are intergrated in our models, providing scalability to address emotions at fine-grained level in the future.
- We build a conversation corpus for emotion detection, which is really close to real-life settings. Thus, proving the practicality of our method.
- We experiment on both bootstrapping method and word-embedding method to automatically produce an emotion lexicon which is essential for automatic feature extraction of the raw input data.
- We compare predicting performance of Feature Selection model and Word-embedding model.
- We verify the superior performance of semi-supervised method over supervised one in word-embedding model. This allows us to take advantages of both the sacred labeled data and the vast of unlabeled data available on the Internet.
- The output of our system is multi-label, which means that it is capable of addressing multiple emotions simultaneously.

1.4.2 Research question

This dissertation contributes by trying to answer the following research questions:

1. What is the origin of human emotions? How can we quantify them?
2. How emotions are expressed in text documents? How to use these clues for emotion detection?

3. What are the steps of building an emotion lexicon for a better emotion detection system?

4. What is the method to create a model that is scalable to many domains and emotions?

Giving answers to these question is the goal of our research. We will summarize them in the last chapter 7

1.5 Organization of the dissertation

The rest of this dissertation is organized as follows:

Chapters 2 investigates various existing psychology theories on emotions and points out the most suitable theory to quantify emotion in text. Chapter 3 studies the linguistics properties of emotional expression in text documents, i.e. how writers express their feelings and their method to transmit those feelings to the audience.

Chapter 4 introduces the previous corpora on the same topic of emotion detection. We will explain the need for a new corpus and its annotation scheme. Chapters 5 and 6 investigate the marking of EA system from the applicational perspective. Chapter 5 introduces the previous works on emotion lexicon and the making of emotion lexicon by bootstrapping method and word-embedding method. The performance of each lexicon will also be evaluated against existing lexicon. Chapter 6 describes the related works in making EA system and our steps to build predicting models for EA.

Chapter 7 summarizes the contributions of the dissertation and discusses directions for future work.

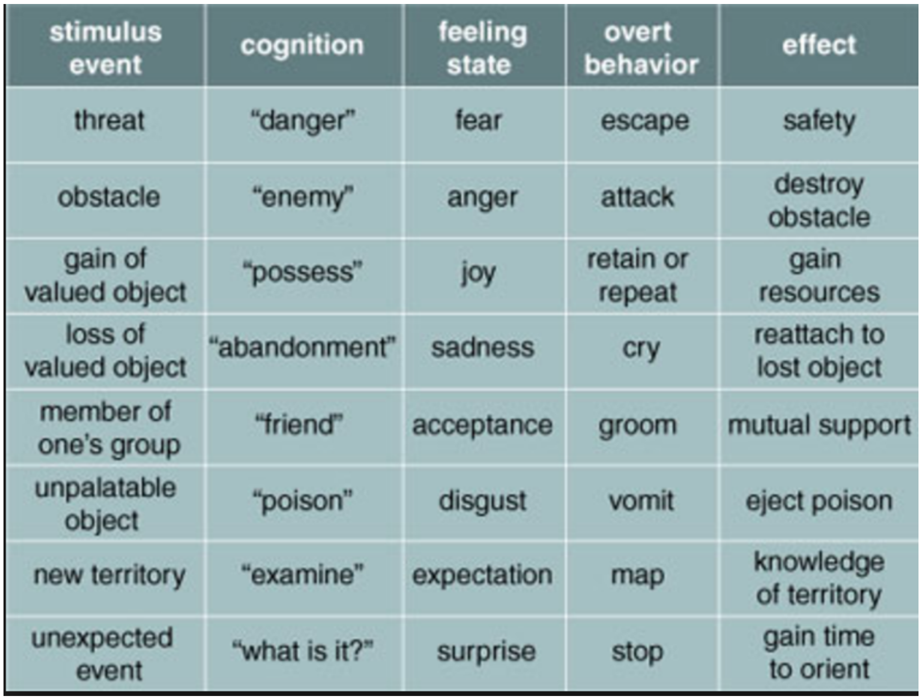
2. Psychology theories on emotions

In psychology, emotion is often defined as a complex state of feeling that results in physical and psychological changes that influence thought and behavior. Emotionality is associated with a range of psychological phenomena, including temperament, personality, mood, and motivation. According to author David G. Meyers [17], human emotion involves "...physiological arousal, expressive behaviors, and conscious experience."

It was naturalist Charles Darwin who proposed that emotions evolved because they were adaptive and allowed humans and animals to survive and reproduce. Charles Darwin's 1872 book The Expression of the Emotions in Man and Animals [18]. Darwin argued that emotions actually served a purpose for humans, in communication and also in aiding their survival. Darwin, therefore, argued that emotions evolved via natural selection and therefore have universal cross-cultural counterparts. Feelings of love and affection lead people to seek mates and reproduce. Feelings of fear compel people to either fight or flee the source of danger. According to the evolutionary theory of emotion, our emotions exist because they serve an adaptive role. Emotions motivate people to respond quickly to stimuli in the environment, which helps improve the chances of success and survival. Figure 2 explain in more details the relation between emotional states and other factors such as: stimulus events, cognition, overt behaviours and effect of the reaction.

More contemporary views along the evolutionary psychology spectrum posit that both basic emotions and social emotions evolved to motivate (social) behaviors that were adaptive in the ancestral environment [19]. Current research suggests that emotion is an essential part of any human decision-making and planning, and the famous distinction made between reason and emotion is not as clear as it seems. It is claimed that emotion competes with even more instinctive responses, on one hand, and the more abstract reasoning, on the other hand.

The classification of emotions has mainly been researched from two fundamental viewpoints. The first viewpoint is that emotions are discrete and fundamentally different constructs while the second viewpoint asserts that emotions can be characterized on a dimensional basis in groupings. While both Paul Ekman and Plutchik agree on evolutionary theory of emotions, they shares different viewpoint about the matter.



stimulus event	cognition	feeling state	overt behavior	effect
threat	"danger"	fear	escape	safety
obstacle	"enemy"	anger	attack	destroy obstacle
gain of valued object	"possess"	joy	retain or repeat	gain resources
loss of valued object	"abandonment"	sadness	cry	reattach to lost object
member of one's group	"friend"	acceptance	groom	mutual support
unpalatable object	"poison"	disgust	vomit	eject poison
new territory	"examine"	expectation	map	knowledge of territory
unexpected event	"what is it?"	surprise	stop	gain time to orient

Figure 2. Although emotional substrates cannot always be discerned in the behavior of nonhuman animals, many stimuli are experienced by people and animals alike and result in prototypical behavior followed by, generally, the reestablishment of an equilibruim state that might not have been achieved without the impulse precipitated by the inner state. In human experience it is common to use the term "emotion" to describe the feeling state, but in fact emotion is considerably more complex.

2.1 Paul Ekman's basic emotion

Paul Ekman has supported the view that emotions are discrete, measurable, and physiologically distinct [20]. Ekman's most influential work revolved around the finding that certain emotions appeared to be universally recognized, even in cultures that were preliterate and could not have learned associations for facial expressions through media. Another classic study found that when participants contorted their facial muscles into distinct facial expressions (for example, disgust), they reported subjective and physiological experiences that matched the distinct facial expressions. His research findings led him to classify six emotions as basic: anger, disgust, fear, happiness, sadness and surprise (figure 3).

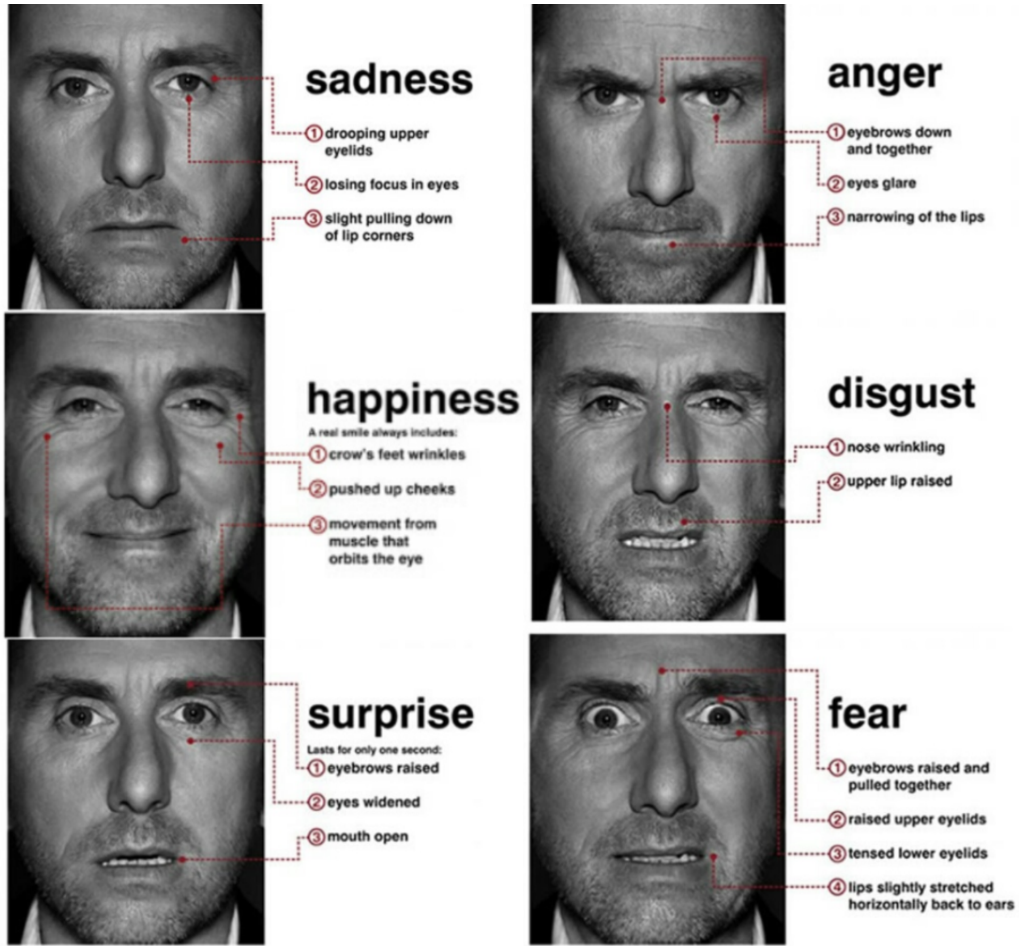


Figure 3. Illustration of basic emotions by Paul Ekman's research on facial expression traits - Photo from `https://www.paulekman.com/micro-expressions/`

However, we should note that Ekman's theory is based on his observation of facial expression of emotions. While the facial expression may be limited to what our face can do, the expressive power of words and languages is limitless. We argue that Ekman's theory is not suitable for our task of emotion detection.

2.2 Dimensional analysis

Through the use of multidimensional scaling, psychologists can map out similar emotional experiences, which allows a visual depiction of the ”emotional distance” between experiences. A further step can be taken by looking at the map's dimensions of the emotional experiences. The emotional experiences are divided into two dimensions known as valence (how negative or positive the experience feels) and arousal (how energized or enervated the experience feels). These two dimensions can be depicted on a 2D coordinate map [21, 22, 23]. This two-dimensional map was theorized to capture one important component of emotion called core affect. Figure 4 show the two dimensions of emotion according to the theory.

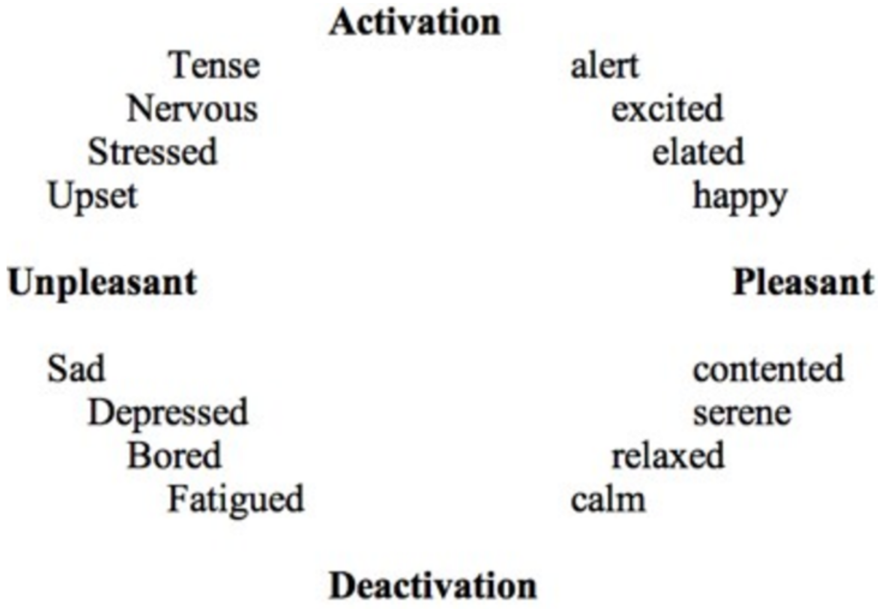


Figure 4. The emotional experiences are divided into two dimensions known as valence (how unpleasant or pleasant the experience feels) and arousal (how energized or enervated the experience feels)

The idea that core affect is but one component of the emotion led to a theory called *psychological construction.* According to this theory, an emotional episode consists of a set of components, each of which is an ongoing process and none of which is necessary or sufficient for the emotion to be instantiated. The set of components is not fixed, either by human evolutionary history or by social norms and roles. Instead, the emotional episode is assembled at the moment of its occurrence to suit its specific circumstances. One implication is that all cases of, for example, fear are not identical but instead bear a family resemblance to one another. The idea is a key to answer our problem: we would like a way to quantify all emotions, find the basic units and dimensions.

2.3 Plutchik's wheel of emotion

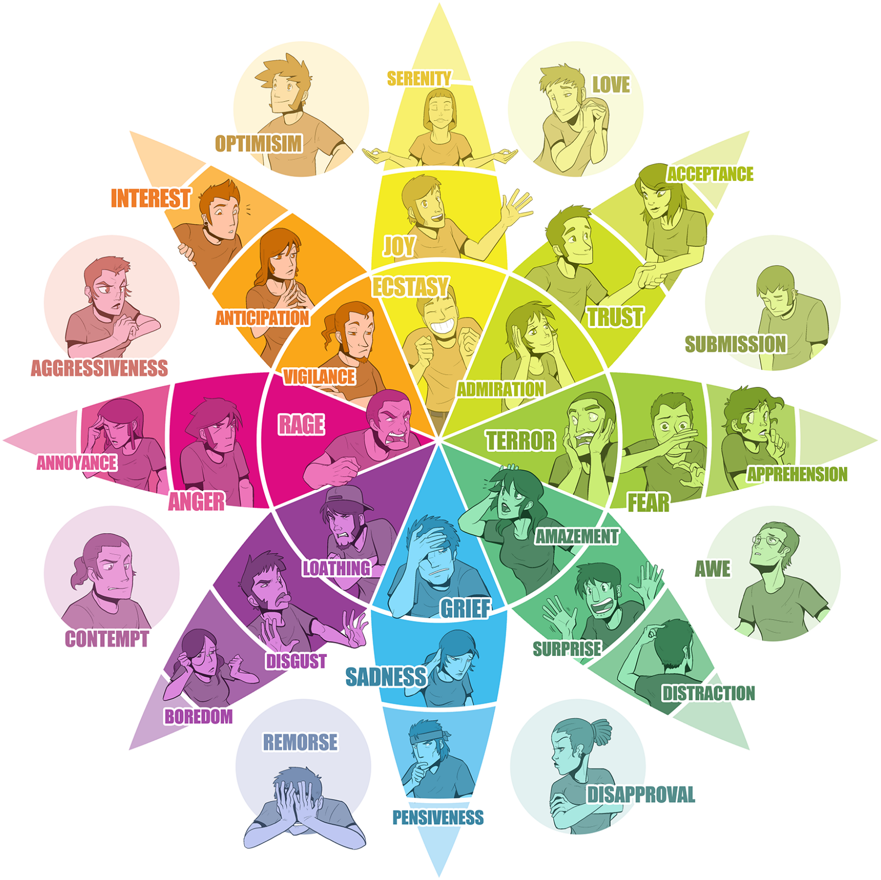
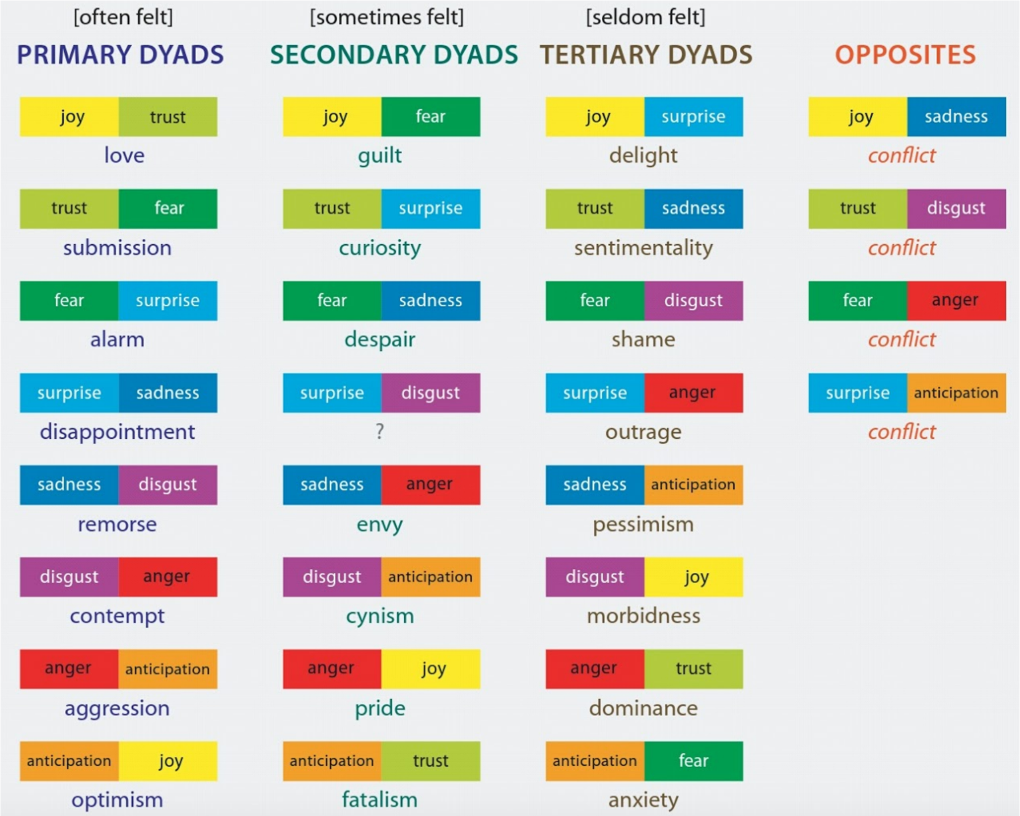


Figure 5. Plutchik's wheel of emotions - image taken from http://twinklet8.blogspot.jp

Robert Plutchik agreed with Ekman's biologically driven perspective but developed the "wheel of emotions", suggesting eight primary emotions grouped on a positive or negative basis: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation [5]. Some basic emotions can be modified to form complex emotions. The complex emotions could arise from cultural conditioning or association combined with the basic emotions. Alternatively, similar to the way primary colors combine, primary emotions could blend to form the full spectrum of human emotional experience. For example, interpersonal anger and disgust could blend to form contempt. Relationships exist between basic emotions, resulting in positive or negative influences. Plutchik's theory is a wonderful combination of dimension analysis theory and evolutionary theory.



[often felt]	[sometimes felt]	[seldom felt]	
PRIMARY DYADS	SECONDARY DYADS	TERTIARY DYADS	OPPOSITES
joy trust love	joy fear guilt	joy surprise delight	joy sadness *conflict*
trust fear submission	trust surprise curiosity	trust sadness sentimentality	trust disgust *conflict*
fear surprise alarm	fear sadness despair	fear disgust shame	fear anger *conflict*
surprise sadness disappointment	surprise disgust ?	surprise anger outrage	surprise anticipation *conflict*
sadness disgust remorse	sadness anger envy	sadness anticipation pessimism	
disgust anger contempt	disgust anticipation cynism	disgust joy morbidness	
anger anticipation aggression	anger joy pride	anger trust dominance	
anticipation joy optimism	anticipation trust fatalism	anticipation fear anxiety	

Figure 6. Dyads: the combination of two emotions

According to Plutchik, there have been many proposal on the primary emotions: all of them include fear, anger and sadness, most include joy, love and surprise. He follows the psychology evolutionary theory and assumes eight basic emotions dimensions arrange in four pairs: sadness, joy, surprise, anticipation, anger, fear, trust and disgust. one called dyad (figure 6) the same way as we mix color. By that way, we can cover the full scope of emotion words in English. Plutchik bases his idea on William McDougall notation on the parallel between emotions and colors in 1921 and extend it to include a third dimension of emotional intensity (figure 5).

Plutchik's notion reasonably explains the connection between emotions. Some emotions are similar but of different intensity. Some emotions will not occur at the same time since they are on the opposite side of the axis. Complex emotions can also be viewed as combinations of primary ones. The idea enables us to approach emotion detection in a more comprehensive manner. [24]

3. Linguistic properties of the emotional expression in text

In this chapter, we discuss the various properties of emotional expression in text. First, the importance of using words and phrases in expressing emotion will be explained. Secondly, we will differentiate the explicit expression of emotion and the implicit expression of emotion. After that, we will visit several linguistics properties that are related to the expression of emotion such as: sarcasm, subjectivity/ objectivity, opinion expression, context information and domain dependency.

3.1 Words and phrases - the obvious expression of emotion

Most research agree that emotion words and phrases are the most obvious clue to identify emotions [14, 25, 16]. Human have developed language to fit their needs of expressing ideas and feelings. Therefore, when describing our emotion, we tend to use some specific words or phrases. For example, *sorrow, downcast, gloomy, blue, low-spirited, cry one's eyes out, one's heart sinks* all describe the feeling of sadness. By picking up on these words and phrases, we have a general idea about the emotional direction of the examined text.

Besides words directly referring to emotional states (e.g., “fear”, “cheerful”) and for which an appropriate lexicon would help, there are words that act only as an indirect reference to emotions depending on the context (e.g. “monster”, “ghost”). Strapparava et al., 2006 called the former direct affective words and the latter indirect affective words [26].

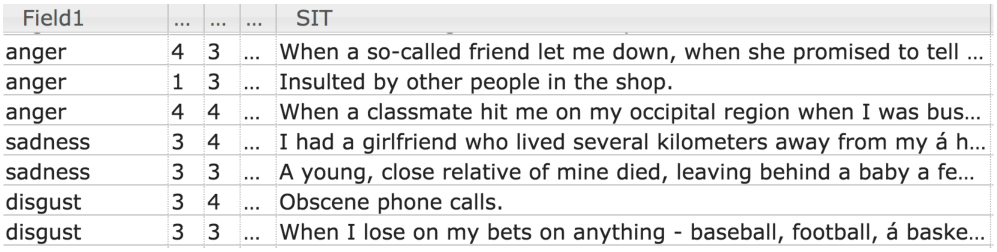
However, solely relying on text would cause problem for emotion detection system. We consider the following example: *I am looking for a good health insurance for my family.* While there are the occurrences of words such as *good or health insurance* may make us think that this sentence expresses some positive emotions, the sentence itself is somewhat objective. It only states a fact about the author's desire and intention.

We should also consider the effect of negation words and phrases. Simply by putting a negation word, we reverse the emotion state of the text. The sentence: *You are not bad at all!* indicate a strong feeling of approval instead of the usual negative feelings from the word *bad.* The but - clauses in English generally means contrary. However, in our case of emotion detection, it may indicates complex mixture of emotions: *I was surprised by the news but feels happy*. The similar phrases such as *with the exception of, except for* suggests a lesser and not complete feeling: *I am happy for all the students except for John*.

We can conclude that not only emotion words but also grammatical phrases and clauses plays crucial part in the expression of emotion in text. By carefully handling these words and phrases, we get more clues about the expressed emotions.

3.2 Implicit expression of emotion

Emotions may be expressed implicitly by the mean of objective sentences. We consider the following paragraph: *I purchased a tablet via ebay. I waited for it for a whole month. It finally came yesterday. However, it was the wrong model. I returned it immediately.* Each sentence is objective and hold no emotion but by reading the whole paragraph, we may sense the disappointment of the customer. It is not because of any specific words but the whole content of the paragraph that suggest the feeling.



Field1	...	...	...	SIT
anger	4	3	...	When a so-called friend let me down, when she promised to tell ...
anger	1	3	...	Insulted by other people in the shop.
anger	4	4	...	When a classmate hit me on my occipital region when I was bus...
sadness	3	4	...	I had a girlfriend who lived several kilometers away from my á h...
sadness	3	3	...	A young, close relative of mine died, leaving behind a baby a fe...
disgust	3	4	...	Obscene phone calls.
disgust	3	3	...	When I lose on my bets on anything - baseball, football, á baske...

Figure 7. ISEAR dataset reports on situations which respondents experience certain emotion

The ISEAR dataset [2] collect the response from students, both psychologists and non-psychologists, who were asked to report situations in which they had experienced 7 major emotions 7. The majority of the reports are objective, however, they hold clues about situations where certain feeling is experienced. We

[2] http://www.affective-sciences.org/researchmaterial

believe that the dataset gives us example about the implicit expression of emotion in text. In our research, we extracted words from this dataset and form a collocation list which connects the words to certain emotion. The use of this collocation list in our work is closely similar to emotional expression clues in [13].

3.3 Sarcasm

Sarcasm is a very challenge case in any text classification problem. According to `Dictionary.com`, sarcasm is harshly used ridicule or mockery , often crudely and contemptuously, for destructive purposes. It may be used in an indirect manner, and have the form of irony, as in "What a fine musician you turned out to be!," "It's like you're a whole different person now...," and "Oh... Well then thanks for all the first aid over the years!" or it may be used in the form of a direct statement, "You couldn't play one piece correctly if you had two assistants."

In sarcasm, people often say things that is opposite of what they really think so any system would fail when dealing with sarcasm. However, hypothetically, if we have an effective mechanism to identify sarcasm, it will be a very powerful clue for emotion detection. Psychologist Clifford N. Lazarus [27] have suggested that sarcasm tends to be a mal-adaptive coping mechanism for those with unresolved anger or frustrations and *hostility disguised as humor*. In our work, we have no mechanism to identify sarcasm so we cannot exploit the possibility, nevertheless, we would like to further investigate the relation between sarcasm and emotion expression in our future work.

3.4 Subjectivity/ Objectivity expression of emotion

Bing Liu [28] differentiate objective sentence and subjective sentence as follow: *An objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs*. We consider the following example

Ex.3.1 This past Saturday, I bought a Nokia phone and my girlfriend bought a Motorola phone.

Ex.3.2 The voice on my phone was not so clear, worse than my previous phone.

Ex.3.3 My girlfriend was quite happy with her phone.

Ex.3.4 I wanted a phone with good voice quality.

Ex.3.5 So my purchase was a real disappointment.

Sentences **Ex.3.1**,**Ex.3.2** are objective because they simply state actual facts. The rest of the story are subjective sentences because they expresses personal feelings. Let us go back to previous example **Ex.3.2** and **Ex.3.4**. While example **Ex.3.2** is objective and simply state a fact about the purchased phone. It implies a negative feeling about the product, specifically: disapproval. Example **Ex.3.4** is subjective but hold little emotional information in it, even though it expresses the desire of the author. Subjective sentences are not necessary contain emotions, and many objective sentences can also imply emotion.

On the other hand, the information about the subjectivity of a sentence can be an important clue about emotions. Obviously, a subjective sentence like **Ex.3.5** clearly give away the emotion of the author. Objective sentence may not explicitly express the emotions but imply them through factual information. In fact, objective sentences that imply positive or negative emotions often state the reasons for the emotions [28]. For example, in the sentence *The voice quality of this phone is amazing*, the author expresses implicitly his surprise and happiness by stating a fact about his phone.

Therefore, in our work, we consider both subjective and objective sentences for the classification task.

3.5 Opinion expression and emotion

Sentiment Analysis and opinion mining is closely related to emotion detection. Bing Liu [28] consider emotions and sentiments are very similar in concepts but not equivalent. In fact, they have very large intersection. Many emotions can be considered positive feelings such as: happiness, trust, amazement. Many others are negative: sadness, disgust, disappointment. Surprise and Anticipation can be viewed as negative, positive or neutral depending on the situation. Because of the similarity, emotion detection can benefit greatly from the methods of opinion mining.

According to Bing Liu [29], an opinion expresses the opinion holder h positive, negative or neutral view, attitude, emotion or appraisal on a feature f of object o. The opinion maybe direct on the object or comparative via another object. In other words, the key difference between opinion expression and emotion expression is the *object*. An opinion must have an object while it is not a requirement in an emotion expression.

It comes as no surprise that the most important indicators of opinion expression are sentiment words, also called opinion words. These are words that are commonly used to express positive or negative sentiments. For example, good, wonderful, and amazing are positive sentiment words, and bad, poor, and terrible are negative sentiment words. Apart from individual words, there are also phrases and idioms, e.g., cost someone an arm and a leg. Sentiment words and phrases are instrumental to sentiment analysis for obvious reasons. A list of such words and phrases is called a sentiment lexicon (or opinion lexicon). Although sentiment words and phrases are important for sentiment analysis, only using them is far from sufficient. The problem is much more complex. In other words, we can say that sentiment lexicon is necessary but not sufficient for sentiment analysis.

The same conclusion can be applied emotion detection. Therefore, following [14, 15], we take advantages of emotional words the same way as sentiments and build an emotion lexicon. More details about the building of emotion lexicon will be discussed in chapter 5.

3.6 Context information

The expression of emotion heavily depends on the context. Especially in the case of implicit emotion expression and sarcasm. Without context information, we would view each individual sentence in the paragraph in subsection 3.2 as objective. We would also fail to recognize sarcasms and treat them at face value.

Context information is also proved crucial in comparative expression. For example: *Cola is better than Pepsi.* Without context information, we do not know the object of the emotion in this sentence. Maybe the speaker is simply making a general statement and the sentence hold no emotion. On the other hand, if the speaker is having a Cola, this statement may hold satisfactory and if he is having a pepsi then it maybe disappointment.

In most of the research, short text such as tweets or news headlines are considered and context information is often ignored. However, if we want to apply emotion detection on paragraphs or conversations, context information is an must feature of the system.

3.7 Domain dependency

As we have mention above, the most obvious clues for emotion expression are words. However, words usually have different meanings and sense in different domain. A tiny size maybe a desired feature for a phone but is unacceptable for a house or a car. In our research, we investigate the performance of a general domain lexicon, a domain-adapted lexicon and a lexicon generate from the specific domain.

4. Corpus and annotation

In this chapter, we explain previous emotion corpora and explain why we need to annotate new corpora. We will go into detail over each of our corpora and their annotation scheme in the later part of this chapter.

4.1 Previous emotion corpora

4.1.1 Affective text corpus

One of the most well studied corpus is Strapparava's Affective Text that is used in SemEval Task 14 competition [16]. The *Affective Text* task was intended as an exploration of the connection between lexical semantics and emotions. Their corpus was collected by extracting news headlines from Google News and New York Times, CNN, and BBC News and provided with six emotion labels (i.e., Anger, Disgust, Fear, Joy, Sadness, Surprise). The goal of this dataset is to conduct sentence-level annotations of emotions. There are training data set consisting of 250 annotated headlines, and a test data set with 1,000 annotated headlines. The test data set was independently labeled by six annotators. The annotators were instructed to select the appropriate emotions for each headline based on the presence of words or phrases with emotional content, as well as the overall feeling invoked by the headline. Annotation examples were also provided, including examples of headlines bearing two or more emotion

The significant of the dataset was fine-grain annotated unlike previous annotations of sentiment or subjectivity (Wiebe et al., 2005 [30]; Pang and Lee, 2004 [31]), which typically relied on binary zero and one annotations. Table 1 shows the inter-annotators agreement score calculated by Pearson correlation measure (Pcm). We can see that the evaluation of emotions from text is quite hard for even human annotators and there are a lot of disagreement among them.

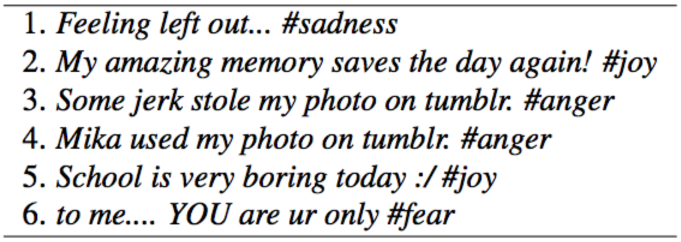
Despite the Affective Text corpus was pioneer in an fine-grained annotation scheme, their target - news headlines was of little use. The news were discrete and there was no context information about them. Using the corpus, one may only develop a word-based emotion detection system which is not enough for real-life tasks.

Emotions	Pcm - agreement
Anger	49.55
Disgust	44.51
Fear	63.81
Joy	59.91
Sadness	68.19
Surprise	36.07

Table 1. Inter-annotators Agreement score of Affective Text corpus

4.1.2 Twitter emotion corpus

Creating a large emotion corpus is often expensive and time consuming because it involves annotators. Therefore, Mohammad [32] proposed Twitter Emotion Corpus which crawled tweets with emotion hashtags. They conduct experiments to show that their hashtags annotations are consistent and comparable to annotations of trained judges. After filtering unsuitable tweets, Twitter Emotion Corpus were left with 21,000 instances, each of which was annotated with a single emotion from Paul Ekman's basic six emotions. Figure 8 shows some examples from the corpus.



1. *Feeling left out... #sadness*
2. *My amazing memory saves the day again! #joy*
3. *Some jerk stole my photo on tumblr. #anger*
4. *Mika used my photo on tumblr. #anger*
5. *School is very boring today :/ #joy*
6. *to me.... YOU are ur only #fear*

Figure 8. Examples from Twitter emotion corpus

One can easily figure the emotions from example 1 to 3. However, the emotions are implicitly expressed in the rest examples. Basing only on the text, we cannot figure the emotion of the speaker. This is the major shortcoming of the corpus. Because they collected data from micro-blogging platform like Twitter, their data is not related to each other and is too short to provide context information.

The corpus was also only labeled with multi-class scheme which only allows one example to hold only one emotion at most. This contradicts the nature of human emotion that there are many cases where multiple emotions are expressed at the same time.

4.2 Our corpora

The limitation with those previous corpora is that they only consist of short, independent pieces of text and undoubtedly not close to real-life conversation. As a matter of fact, modeling emotions in a conversation is indeed a difficult but rewarding task with a wide range of applications. A good system should consider every word in the conversation, the grammatical structure and syntactic variables such as negations, embedded sentences, and type of sentence (question, exclamation, command, or statement), the general context of the conversation, each and every utterances in the conversation - especially when what is said in the previous utterance can have an impact on the emotions of the later one (Collier, 2014 [1]). As pointed out in many psychology research (Plutchik, 2001 [5]; Russell, 2003 [23]), emotions are not mutually exclusive. In fact, in many cases, people may experience a mixture of various emotions at the same time (Choe et al., 2013 [33]). Therefore, the corpus for any emotion analysis task should be multilabel. Limiting the number of emotion labels may narrow down the problem but can cause troubles for the annotators to provide correct judgement when the emotion

In addition, text only corpus is very hard for annotators to interpret emotions. Research by Kruger et.al., [34] have proved that the lack of social and emotional cues over virtual communication platforms can result in increased instances of misinterpreting emotion and intentions. Therefore, the above listed corpora are not suitable for real-life application where we would focus on human-human or human-machine conversations. The machine should not only interpret the human emotions directly from each exchange in the conversation but also the implicit emotions by considering the context of the whole conversation as well.

We propose to construct a new corpus that satisfies the following criteria:

1. The corpus should be transcripts of real-life conversations or at least close to them.

2. The corpus should be in emotionally rich domain.
3. The inter-annotators agreement score of the corpus should be acceptable. Otherwise, the reliability of the research will be questioned.
4. The corpus should be annotated using multi-label scheme.

As explained previously, we do not agree with the construction of existing corpora. Therefore, we decided to build and annotate an emotional rich corpus ourselves following the listed criteria. Our first attempt is the Cornell movie dialog corpus where the annotator will only rely on text information to give their judgement. However, this corpus suffered from low agreement score. As a result, we then created EMTC corpus where the video information are added in the annotation process. We will explain the building of these corpora and discuss the improvement of the agreement score in the following sections. Both of the corpora follow the annotation scheme shown below:

- One utterance may hold zero, one or more emotions at the same time. The list of emotions to assign includes Plutchik's basic emotions and dyads. The system will treat the dyads as combination of basic emotions. In case an utterance holds no emotion, it should be annotated with "None". The intensity of emotions is also considered in the labeling phrase as in subfigure 9a.
- The annotators need to assign the whole utterance which may have two or more sentences with a set of all emotions expressed inside it. There may be cases where conflict emotions according to Plutchik's notion appear simultaneously in the same utterance as in the last example of the subfigure 9b.

4.2.1 Cornell movie dialog corpus

Cornell movie dialog corpus is our first attempts to annotate and create a reliable corpus for the task of emotion detection. [24]. The Cornell Movie Dialog dataset was originally used for understanding the coordination of linguistic style in dialogs (Danescu-Niculescu-Mizil and Lee, 2011 [35]). It includes in total 304,713 utterances (turns in conversation) out of 220,579 conversational exchanges between 10,292 pairs of 9,035 movie characters from 617 movies. The annotating scheme is as mention above.

The followings are some statistics of the annotated corpus: total of 11,610 utterances, 10,008 of which are in the training data , 1,602 others are in the testing data, the average number of label per utterance is 1.29. We separated the training data which was annotated by only one annotator and the testing data which was annotated by all three annotators.

Emotion class	**Accuracy**
Anger	0.504
Fear	0.535
Disgust	0.184
Trust	0.146
Joy	0.142
Sadness	0.171
Surprise	0.790
Anticipation	0.202
Average accuracy (by class)	0.334
Average accuracy (by annotator)	0.31
Average F1 (by annotator)	0.297
Total No. utterances	1,602

Table 2. Agreement score with gold standard data as ground-truth in Cornell movie dialog corpus.

We define the gold-standard data as the data agreed by at least 2/3 annotators. We measure the Accuracy and F1 score for each emotion class in comparison with the gold-standard data and then average them as shown in table 2.

One of the most common **Inter-Annotator Agreement** measurement - the Kappa statistics [36]. However, it is not applicable to multi-label dataset because its way of computing causes hypothetical probability of chance agreement P_e to be greater than 1 since there are cases where two or more labels are annotated to a given instance. Therefore, we consider the gold standard data as ground-truth data and measure the average accuracy of each of the emotions and F1-score of the annotators in Table 2.

The survey by Artstein and Poesio (2008) [37] suggested that low agreement scores are often observed in multi-label annotating tasks even when the annotators do not make much use of the ability to assign multiple tags. Some strong emotions: ”Anger”, ”Fear”, ”Surprise” have better agreement scores as they have indicators such as question marks and excalmation forms. Nevertheless, they are easier for human to identify because they are the basic emotions that we - human inherits from animals. They are the emotions that triggers “fight or flight” and “stop and examine” response. (Plutchik, 1980) [4]. The following table 3 includes an example of conversation where the utterances do not receive the agreement among the annotators:

In this corpus, because the annotators only worked with the text data, it was very difficult for them to visualize the situations and make correct judgments. Therefore we decided to use the movie clips to add the annotation process. However, because this corpus is movie scripts (and not the movies' subtitles), to truly understand the situations in each dialogs, annotators need to watch the whole movies; which is highly impractical. Therefore, we moved to the IMDB movie quotes dataset.

4.2.2 EMTC: Emotion movie transcript corpus

In order to mimic real-life conversation settings, we resort to the newly published and frequently updated IMDB quotes corpus [3]. It includes in total 2,107,863 utterances (turns in conversation) out of 117,425 movies. To our assumption, movies conversation should be close to real-life conversation and emotionally rich. We can also easily ensure the acceptable agreement score between annotators by providing them the clips from the movies in addition to the transcripts. We also believe that misinterpretation of the emotions from text only can be reduced dramatically then.

While we have confidence values for each annotation, we decide not to use the values in this work yet. In future work, we hope to be able to predict the intensity of the emotion as well.

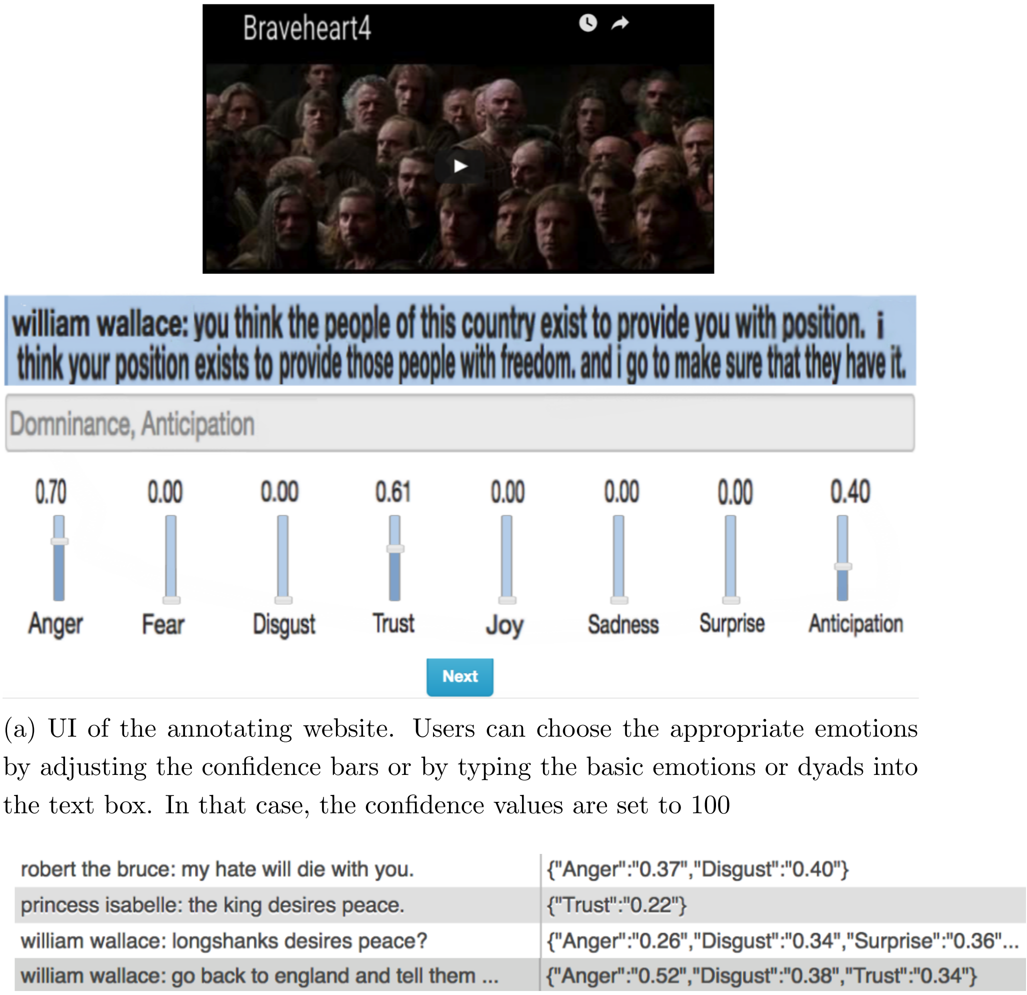
[3] The datasets are available from http://www.imdb.com/interfaces

	Utterances	**Anno1**	**Anno2**	**Anno3**	**Agreement**
1	PINZON: You lied! You cheated! We're way past 750 leagues	Angry	Angry	Angry	Angry
2	COLUMBUS: Six days ago, yes.	Anticipation	None	Sadness	
3	PINZON: You must be mad...!	Angry	Surprise	Fear	
4	COLUMBUS: We have to keep the hopes of these men alive!	Angry	Trust	Sadness	
5	PINZON: We're on the verge of a mutiny, Colon!	Angry	Fear, Sadness	Fear	Fear
6	COLUMBUS: You think I don't know that?	Angry	Disgust	Anticipation	
7	PINZON: We're lost!	Angry	Fear	Fear	Fear
8	COLUMBUS: The land is there. I know it!	Angry, Trust	Trust, Anticipation	Trust	Trust

Table 3. Example of disagreement among annotator in Cornell movie dialog corpus.

No. Annotators	**No. utterances**
Agreed by 2 annotators	996
Agreed by 3 annotators **(gold-standard)**	896
Agreed by 4 annotators	443
Agreed by 5 annotators	253
Total No. utterances	1,000

Table 4. Number of utterances received agreements by annotators in the testing data



william wallace: you think the people of this country exist to provide you with position. i think your position exists to provide those people with freedom. and i go to make sure that they have it.

Domninance, Anticipation

Anger	Fear	Disgust	Trust	Joy	Sadness	Surprise	Anticipation
0.70	0.00	0.00	0.61	0.00	0.00	0.00	0.40

(a) UI of the annotating website. Users can choose the appropriate emotions by adjusting the confidence bars or by typing the basic emotions or dyads into the text box. In that case, the confidence values are set to 100

robert the bruce: my hate will die with you.	{"Anger":"0.37","Disgust":"0.40"}
princess isabelle: the king desires peace.	{"Trust":"0.22"}
william wallace: longshanks desires peace?	{"Anger":"0.26","Disgust":"0.34","Surprise":"0.36"...
william wallace: go back to england and tell them ...	{"Anger":"0.52","Disgust":"0.38","Trust":"0.34"}

(b) Examples of annotated transcripts from movie: Brave Heart (1995) - In the last example, opposite emotions of Trust and Disgust are both annotated

Figure 9. Annotating scheme of the testing data. Each utterance is annotated with basic emotions

We separated the *training data* which was annotated by only one annotator and the *testing data* which was annotated by 5 annotators. The *gold standard data* is generated by applying the majority rule: any emotion annotated by at least 3 annotators is considered a valid label for the utterance. If an utterance has no valid label, it is considered to be objective and have no emotion. We report the statistics of the testing data in Table 4. As we can observe, there are 104 utterances which have no label according to the *golden standard*.

The followings are some statistics of the corpus: a total of 2,107,863 utterances of over 26 millions words; 10,000 of which are in the labeled training data and annotated by only one annotator; 1,000 others are in the golden testing data; the rest are used as unsupervised data. The average number of label per utterance in the labeled dataset is 1.41.

Emotion class	**Accuracy**
Anger	0.72
Fear	0.673
Disgust	0.624
Trust	0.65
Joy	0.606
Sadness	0.584
Surprise	0.575
Anticipation	0.491
Average accuracy (by class)	0.615
Average accuracy (by annotator)	0.43
Average F1 (by annotator)	0.626
Total No. utterances	1,000

Table 5. Agreement score with gold standard data as ground-truth in EMTC.

We can see that, our corpus, even when being annotated using multi-label scheme, yields better agreement score to the existing multi-class Affective Text, Twitter Emotion Corpus (Average F1-score is 43.7). In addition, EMTC in comparison with our own multilable Cornell movie dialog corpus 2, a clear improvement in agreement scores can be observed. The accuracy for each and every emotion classes as well as the overall F1 score increases. We understand that because the two corpus are not identical, this observation might be not technically sound. However, we did not have the opportunity to run again the annotation without video clips on EMTC. Therefore, this indirect comparison have been made. We argue that the improvement have been achieved with the help of video clips during the annotation process.

5. Building emotion lexicon for emotion detection

Using Lexicon is proven to provide significant improvement in identifying the emotion conveyed by a word [14]. Therefore, in our case, new lexicon is built, each lexical item of which displays not only its association with Plutchik's basic emotions but also how strong the association is. In our work, we experiment on two method of building lexicon, one by bootstrapping via Wordnet domain and another by word-embedding.

5.1 Previous work on emotion lexicon

An emotion lexicon, in its simplest form, is a list of words and associated emotions and sentiments. For example, the word excruciating may be associated with the emotions of sadness and fear. Note that such lexicons are at best indicators of probable emotions, and that in any given sentence, the full context may suggest that a completely different emotion is being expressed. Therefore, it is unclear how useful such word-level emotion lexicons are for detecting emotions and meanings expressed in sentences, especially since supervised systems relying on tens of thousands of unigrams and bigrams can produce results that are hard to surpass. For example, it is possible that classifiers can learn from unigram features alone that excruciating is associated with sadness and fear.

The WordNet Affect Lexicon (Strapparava and Valitutti, 2004 [15]) has a few thousand words annotated for associations with a number of affect categories. This includes 1536 words annotated for associations 587 with six emotions considered to be the most basic - joy, sadness, fear, disgust, anger, and surprise (Ekman, 1992 [20]). It was created by manually identifying the emotions of a few seed words and then labeling all their WordNet synonyms with the same emotion. Affective Norms for English Words has pleasure (happy–unhappy), arousal (excited–calm), and dominance (controlled–in control) ratings for 1034 words.

Mohammad and Turney (2010; 2012 [38], [14], [39]) compiled manual annotations for eight emotions (the six of Ekman, plus trust and anticipation) as well as for positive and negative sentiment.3 The lexicon was created by crowdsourcing to Mechanical Turk. This lexicon, referred to as the NRC word-emotion lexicon

(NRC-10) version 0.91, has annotations for about 14,000 words.

5.2 Bootstrapped lexicon

We define the primary emotions and dyads in Plutchik's theories as the seeds of our lexicon. Throughout Wordnet, we search for *synonyms, hypernyms, hyponyms* of the seeds. A reverse lemmatisation is necessary to retrieve related verbs, adjectives and adverbs and their derived forms (verb forms and comparative, superlative adjectives) of the seeds. We keep tracks of the original nouns and the seeds where the new words were derived from (Table 6). Note that sometimes a word was derived from different nouns and seeds, which suggests mixed emotional states.

Words	Original Nouns - **Seeds**
joy	(primary)- **joy**
sadness	(primary)- **sadness**
fear	(primary)- **fear**
love	(dyad)- **love**
benevolent	benevolence- **love**
worship	worship-**fear** , worship-**love**

Table 6. Wordnet expansion.

Each lexical item in the lexicon has a vector of values on each axis of the basic emotions: Joy - Sadness, Fear - Anger, Trust - Disgust, Surprise - Anticipatation . We manually assign the primary emotions with a value vector of 1, 0 or -1 and the dyads with 0.5, 0 or -0.5, depending on the axes they belong. For example, "joy" came from the axis of Joy-Sadness, thus, its vector is [1,0,0,0] while the vector for "sadness" is [-1,0,0,0] (Table 7). The dyad "love" came from primary emotions "joy" and "trust", hence its vector is [0.5,0.5,0,0]. It is to be noted that the minus sign only indicates that the emotion is on the other side of the axis. It is not a suggestion of negative emotion in any case.

In addition, we calculate the *wup* similarity [40] between a new word and the seed where it came, based on the depth of the two senses in the Wordnet taxonomy and that of their Least Common Subsumer.

$$wup(word, seed) = \frac{2 * dep(lcs)}{dep(word) + dep(seed)} \tag{1}$$

We assumed that the higher the similarity, the closer emotional state of the word to the seed. Thus, the value vector of a word is the sum of the products of each seed vector and the similarity between the word and such seed.

$$\begin{aligned} vector(word) = \textstyle\sum_{k=1}^{n} & vector(seed_k) \\ & \times wup(word, seed_k) \end{aligned}. \tag{2}$$

For example, in the case of the word ”worship”, we first calculate the *wup* scores between the word and its two seeds: fear and love (Table 6). Next, they are multiplied by the vectors of the seeds fear-[0,0,1,0] and love-[0.5,0.5,0,0], and then summed up to get the result (Table 7).

Words	**J-S**	**T-D**	**F-A**	**S-An**
joy	1	0	0	0
sadness	-1	0	0	0
fear	0	0	1	0
love	0.5	0.5	0	0
benevolent	0.47	0.47	0	0
worship	0.14	0.14	0.29	0

Table 7. Value vector of some example words. (*J-S: Joy-Sadness, T-D: Trust-Disgust, F-A: Fear-Anger, S-A: Surprise-Anticipation*)

5.2.1 Adapting the lexicon to new domain

We understand that a lexicon bootstrapped from a general domain resource such as Wordnet has its limited effectiveness when it is applied on a specified domain. In order to partly solve this problem, we built a simple multi-task neural network with an input layer, a hidden layer and an output softmax layer - figure 10.

The input to the network is the Bag-of-Words features of the training data. We then steps by steps, try to do binary classification on the basic emotion e_j.

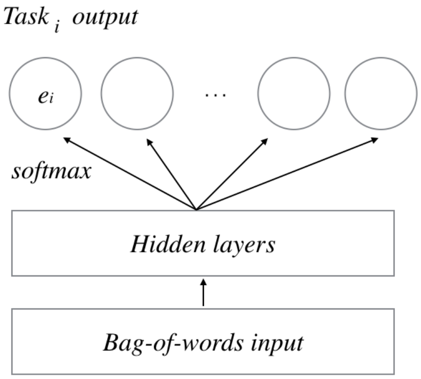


Figure 10. Adapting lexicon for emotional state e_j

We would like to minimize the combination loss of the classification of all 8 basic emotions. We use the log-likelihood as the cost function for the network input: $C = -\ln(a_y^L)$ where a^L is the output of the final layer and y is the desired output. The combination loss function is the average of all 8 tasks:

$$loss = \sum_{i=1}^{N=8} loss_i \tag{3}$$

In the end, we updated the lexicon with weights from the input layer of the network.

5.3 Word-embedding lexicon

We combine word2vec features and calculated emotion features to form a hybrid vector representation of a lexicon item as follow:

5.3.1 Word2vec features

Using word2vec, we generate embedding of all words available in the corpus. With the embedding, the cosine similarity between each word and the primary emotion words is calculated - equation 4. In our work, the embedding of a word is 96-dimensional vector.

$$sim(word, e_i) = \frac{vec(word) \cdot vec(e_i)}{\|vec(word)\|_2 \|vec(e_i)\|_2} \tag{4}$$

5.3.2 Emotion features

On the other hand, we define the primary emotions and dyads in Plutchik's theories as the emotional vectors of our lexicon and give them initial values. Different levels of intensity of emotional words are also considered. Each lexical item in the lexicon has a vector of values on each axis of the basic emotions: Joy - Sadness, Fear - Anger, Trust - Disgust, Surprise - Anticipation . We manually assign the primary emotions with a value vector of 1, 0 or -1 and others with 1.5, 0.5 or -0.5,-1.5, depending on the intensity of the emotion according to Plutchik's theory. For example, "joy" came from the axis of Joy-Sadness, thus, its vector is [1,0,0,0] while the vector for "sadness" is [-1,0,0,0]. The word "ecstasy" is of higher intensity than "joy", hence its vector is [1.5,0,0,0]. It is to be noted that the minus sign only indicates that the emotion is on the other side of the axis. It is not a suggestion of negative emotion in any case. Figure 11 explains the intensity and polarity of basic emotions according to the theory.

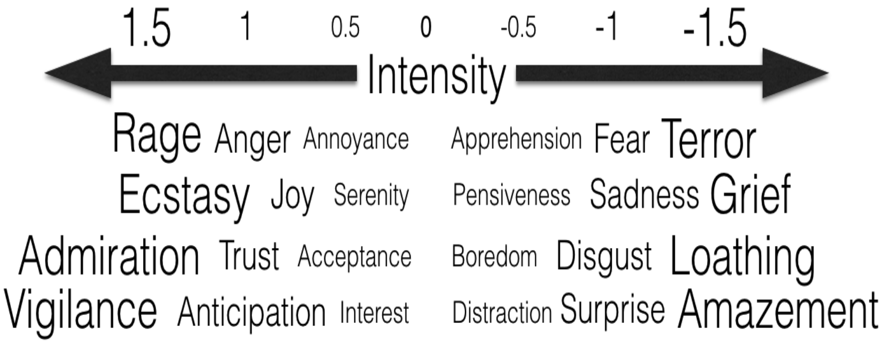

1.5
1
0.5
0
-0.5
-1
-1.5
Intensity
Rage Anger Annoyance Apprehension Fear Terror
Ecstasy Joy Serenity Pensiveness Sadness Grief
Admiration Trust Acceptance Boredom Disgust Loathing
Vigilance Anticipation Interest Distraction Surprise Amazement

Figure 11. Intensity and Polarity of basic emotions.

For the dyads and other words, we calculate the similarity between these words and all the primary emotions e_i. We assume that the higher the similarity, the closer emotional state of the words to the primary emotions. The emotional vector of one word is the averaged result of all primary emotion vectors multiplied by the similarity weights - equation 5. Because there are 4 axes of basic emotions in Plutchik's theory, the result of this step is a 4-dimensional vector of emotion features.

$$vec(word) = \frac{\sum_1^n sim(word, e_i) \times vec(e_i)}{n} \tag{5}$$

The final embedding is the concatenation result of the word2vec generated vectors and the newly calculate emotional vectors. In this research, our embedding is 100 dimensional vectors, 96 of which are generated with word2vec and other 4 are generated using the above steps.

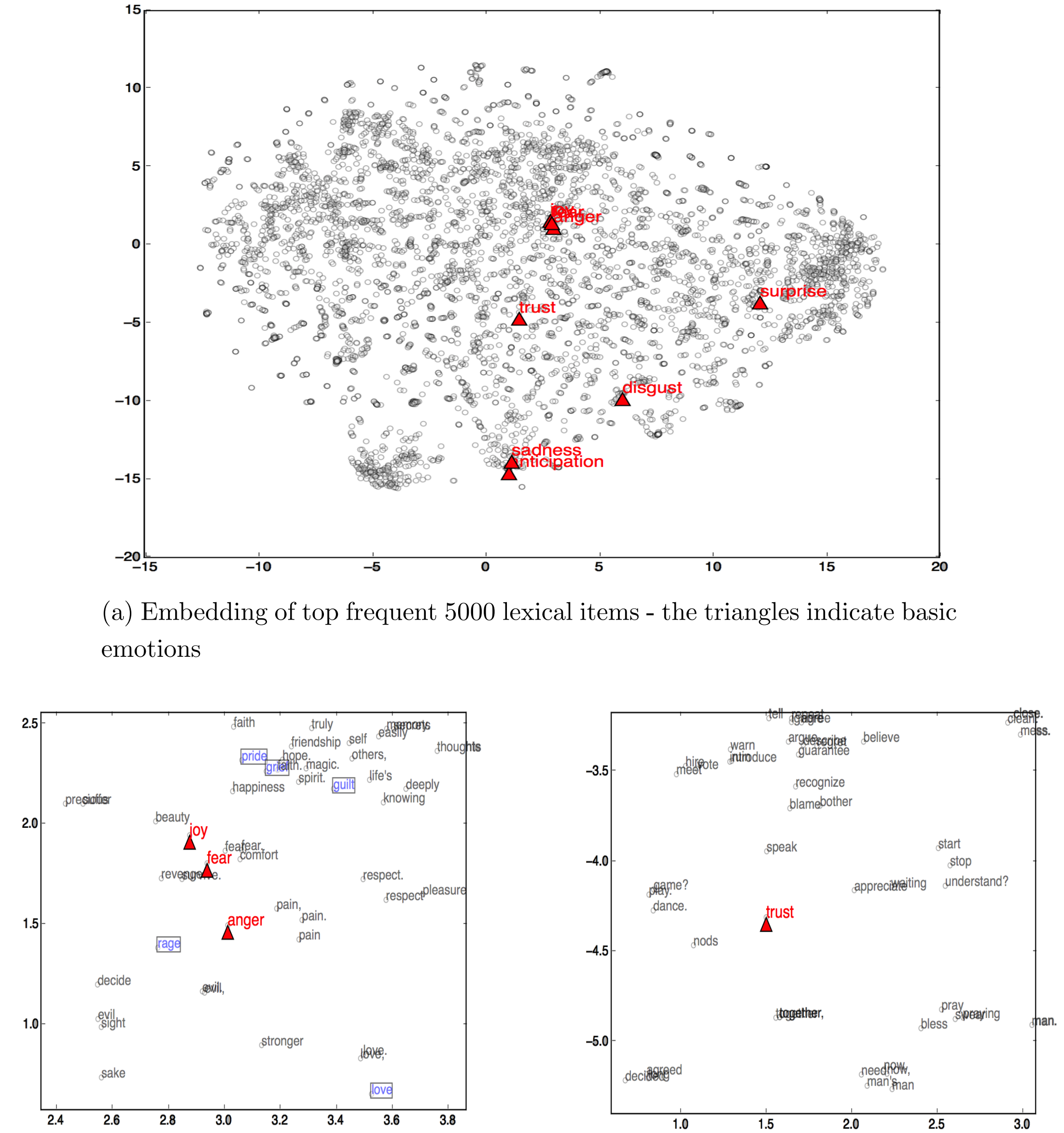
5.3.3 Visualization of the lexicon

Our lexicon consists of 181,276 lexical words which is much larger than most of previous lexicon by other researchers. NRC Emotion Lexicon [14], Wordnet-Affect [15] contains 25,000 and 2,876 synsets respectively. Figure 12a shows the visualization of the top 5,000 popular lexical items and some of the basic emotions. Despite the fact that the visualization is done by reducing the number dimensions of each lexical item from 100 to only 2 dimensions, some interesting results are found in figure 12.

From the figure, we can observe that except for the pair Fear-Anger, other opposite basic emotions located quite far from each other (sub-figure 12a) which is the desirable outcome of the lexicon. Interestingly, in a small cluster of sub-figure 12b, we observe three basic emotion: Joy, Fear and Anger. The surrounding lexical item while appear to be random at first sight, somehow make sense: words like: *pain, rage, evil, and curse* are close to **Anger**; *pride, happiness, and beauty* are close to **Joy**. The dyad *guilt*, which according to Plutchik's theory is the combination of Joy and Fear (sub-figure 6), is also present in this small cluster. In the cluster of **Trust** (sub-figure 12c), we see lexical items which suggest agreement such as: *nods, agreed, and appreciate.* These results indicate that the lexicon obtained is quite the way as it is expected.

5.4 Experiment

We conducts experiment on our three lexicon: wordnet bootstrapped lexicon, domain adapted bootstrapped lexicon and the word-embedding lexicon and the baselines lexicon: NRC word emotion lexicon [14] and Wordnet Affect [26]. Using Bag-of-Words approach, we translate the input in the the sum vectors of individual words and then feed them into a simple neural network with 2 hidden layers on supervised training data. The result is show in figure 13.



(a) Embedding of top frequent 5000 lexical items - the triangles indicate basic emotions

(b) Items with most similarity to basic emotions: Joy, Fear and Anger

(c) Items with most similarity to basic emotion: Trust

Figure 12. Visualization of the lexicon in 2 dimensions: The opposite emotions are often far from each other while lexical items with similar meaning are close.

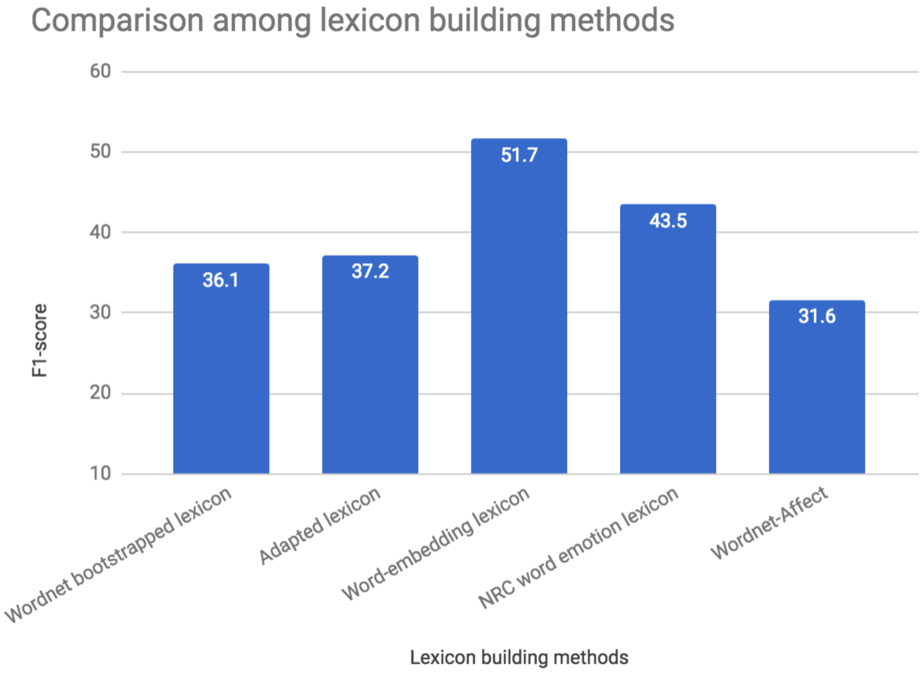


Figure 13. Performance of each lexicon building method

5.5 Conclusion

We can see from the figure 13 the effectiveness of the word-embedding method. It stays at the top with the F1-score of 51.7. We believe that fact that the lexicon is built from the very corpus that we use for the experiment favors the method. The lexicon is also highly dimensional which allows it to fit and adjust better in the neural network.

On the other hand, while the bootstrapped lexicon and the domain adapted lexicon perform better than Wordnet-affect, they are both far behind NRC word emotion lexicon. Their dimensions are too low and there is not much room for adjustment.

6. Models for predicting emotion in text documents

In this chapter, we will review the existing works in emotion detection field. By discussing the advantage and disadvantage of the existing approach, we explain important points that we will have to pay attention to in order to build good emotion detection systems.

In the second part of this chapter, we present our two approaches of making emotion detection systems: the first approach of manual choosing the features for the model and the second approach of using automatic word-embedding. The performance of the two methods and several baselines will be evaluated in the conclusion subsection

6.1 Related works

According to Tao et al.,[41], emotion detection or Affective computing or the task of assigning computers the human-like capabilities of observation, interpretation and generation of affect features, is an important topic for the harmonious human-computer interaction, by increasing the quality of human-computer communication and improving the intelligence of the computer.

Most of the works in the field focus on: 1) brain signals, video of facial expression, audio recording,etc. 2) multiclass classification of emotions. Little research has been done on the detection of multiple emotions simultaneously in textual data.

Nowadays, along with the popularity of social networks, the Internet itself contains more than ever an enormous amount of unlabeled data, most of which is textual. By mining and applying semi-supervised Emotion Detection techniques on such data, we open ourself to a wide range of useful applications such as: measuring citizen happiness, improving customer services, social mental health care, early screening of possible suicides or crimes, etc. While there has been such research on news headlines, tweets or paragraphs; the most useful type of text, conversational text such as chat logs, comments on social media, is often ignored. Text of this type is of great practicality to applications that use Emotion Detection as it gives us information about the initial emotional states, how they change during the conversation, what causes those changes, what are the final states, and what kind of actions are resulted from those final states. The survey on ”Trend Analysis in Social Networking using opinion mining” [42] also predicts the need for Emotion Detection in streaming data and live chat.

Conversational text is also more of a challenge than other types of text. For short text such as news headlines [16] or tweets [43] , the expression of emotion generally depends on the words being used. Meanwhile, for longer text, the grammar structure and syntactic variables such as negations, embedded sentence, and the type of sentence - question, exclamation, command or statement all play a part in expressing emotions [1]. Identifying emotions from conversational text is also very different from paragraphs because there are often more than one party in a conversation. In a conversation, each party takes turns, which are called utterances, to express different ideas and emotions and make impacts on the other party's emotions. Therefore, to detect emotions in conversational text, one must monitor not only the current utterance but also the previous utterance as well as the context of the whole conversation [24].

Automatically identifying emotions expressed in text has a number of applications, including tracking customer satisfaction [44], determining popularity of politicians and government policies [32], depression detection, affect-based search (Mohammad, 2011 [38]), and improving human-computer interaction. Supervised methods for classifying emotions expressed in a sentence tend to perform better than unsupervised ones. They use features such as unigrams and bigrams (Alm et al., 2005 [45]; Aman and Szpakowicz, 2007, [10]).

Using Plutchik's basic emotions, [7] proposed a simple bag-of-words approach and fine-tuned RAkEL for multi-label classification of movie reviews. We delve further and work on the conversation data, where the exchange between the characters and the context of the entire dialogue is of great importance. The closest example of our work is [46] on paragraphs and documents, which tried to improve the sentence level prediction of some special emotions which, owing to data sparseness and inherent multi-label classification, were very difficult to predict. These researchers incorporated label dependency between labels and context dependency into the graph model to achieve the goal. However, their work is for paragraphs in Chinese. In our case, we take advantage of text2vec and deep networks to capture the abstract representation of context information.

6.2 Features selection model

This is our first attempts for emotion detection. We use a set of manually constructed features instead of the direct word-embedding from the bootstrapped lexicon to input into the multilabel neural network.

6.2.1 Features extraction

The process of feature selection for the network is an heuristic one. We initially used a lot of features and then through logistic regression, unimportant features such as the genre of the movie or n-grams features were filtered out.

The core part of the extraction process is to take advantages of the lexicon to transform an utterance to a vector of values expressing the tendency towards each emotion state. This task is done in a rule-based manner (Figure 14). Each word in the utterance is mapped to the lexicon to retrieve the value vector. The representation vector of an utterance is the sum vector of all the word inside it. The negation and word dependency are also taken into account when we calculate the sum with the help of NLTK [47] dependency parsing.

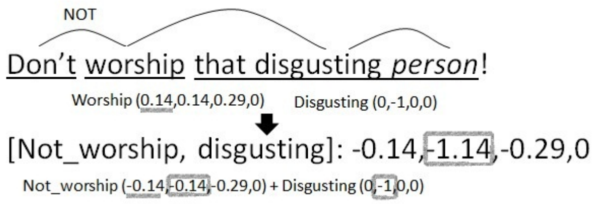

NOT
Don't worship that disgusting person!
Worship (0.14,0.14,0.29,0)
Disgusting (0,-1,0,0)
[Not_worship, disgusting]: -0.14,-1.14,-0.29,0
Not_worship (-0.14,-0.14,-0.29,0) + Disgusting (0,-1,0,0)

Figure 14. Extracting tendency features

Each utterance in the dataset is presented by the following compact set of 23 features:

1. The sum vector of the current utterance which suggest the *local tendency.*

2. The sum vector of all the utterances in the lexicon that appear in the conversation which provides the *context of the conversation.*

3. The sum vector of the previous utterance in the conversation which also provides the *context of previous exchange* (of what triggered the current emotion).

4. The *polarity* (negative/ positive) score of the sentence.

5. *Features* such as: length, is_it_a_question, is_it_an_exclamatory_sentence, is_there_negation_word.

6. *Collocation features* which indicate the number of appearances of words inside the ISEAR collocations list.

The reason for us to use extracted features is that it is very hard to capture the context of both the conversation and previous exchange using direct word-embedding. While using a recurrent neural network can solve the latter, it is a challenge to address the first. Each conversation has different number of utterances, it may hurt the performance of the system and result in network architecture complexity if we use a non-fixed size window to monitor all the utterances in a same conversation.

6.2.2 Building the deep network

The structure of the network is built as shown in Figure 15. The raw input is generalized to produce a small set of features. These features are fed to the network as input layer. We have 2 fully connected hidden layers and an output layer. Since the task is a multi-label classification problem where softmax cannot be used, the output layer is changed into sigmoid. We add a set of threshold values (one for each basic emotions). Only the labels, whose output values greater than the threshold are considered valid. The thresholds are randomly initialized and then updated after each epochs the same way we updated the biases and weights. In our implementation of the network, Theano [48] was used to take advantages of GPU computing power.

The global cost function, similar to [49], is defined to reward the system for right predictions and severely punish for wrong ones in equation 6.

$$E = \sum_{i}^{m} = \frac{1}{|Y_i||\bar{Y_i}|} \sum_{(k,l)\in Y_i \times \bar{Y_i}} \exp\left(-(c_k^i - c_l^i)\right) \tag{6}$$

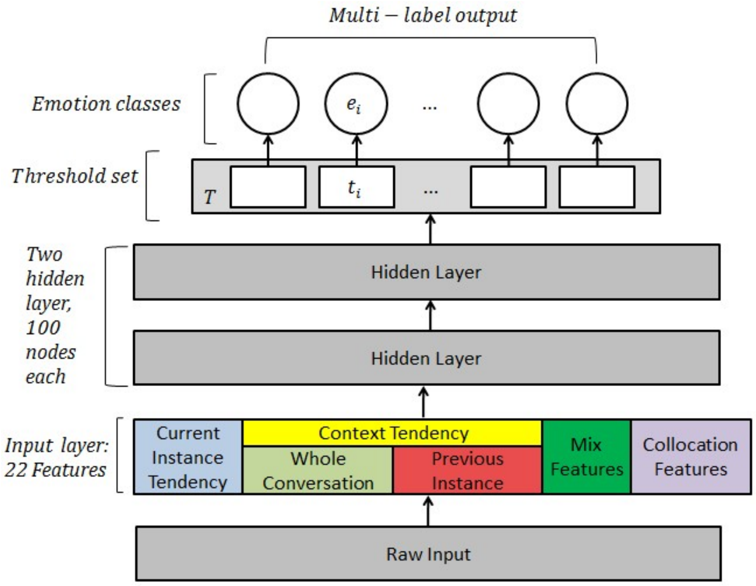


Figure 15. Structure of the Deep Network

Let X be the set of all m instances. Let $Y = \{1, 2, .., Q\}$ be the set of all possible labels, Y_i is the set of true labels for ith instance x_i and $\bar{Y_i}$ is the set of the labels not belong to x_i. Obviously, $Y_i \cup \bar{Y_i} = Y$. We define E as the global cost function of the network. c^i is the set of actual outputs of the model for input x_i, each label has its own output. c^i_k is the output of label k belongs to the set of true labels, $k \in Y_i$. Meanwhile, c^i_l is the output of label l for $l \in \bar{Y_i}$. The difference $c^i_k - c^i_l$ measures the output of the system between the labels, which an instance belongs to and which it doesn't. Naturally, we want this difference to be as big as possible.

6.3 Word-embedding model

In this second attempt, we use the word-embedding lexicon and employ a text2vec mechanism to vectorize the input. It is then fed to an auto-encoder to learn the representation of the unsupervised data and fine-tuned using supervised data.

6.3.1 Texts to vectors

We consider a bag-of-lexical-items approach to transform the raw input text into vectors form. Therefore, for a piece of text, its representation is the sum vector of all lexical items inside. Because our goal is to predict the emotional labels for each utterance in a conversation, we also have to vectorize the previous utterance and the whole conversation to capture the context information (Figure 16). The vectorization of these two components of the context information produces 100 dimensional vector each. Totally, the vector representation of an utterance is a 300 dimensional concatenated vector of the utterance itself and the above mentioned context information. This representation is then fed to the input layer of the Neural Networks in the below sections.

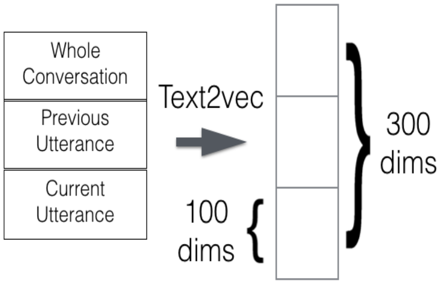


Figure 16. Vector representation of an utterance

Similar to [50], the goal of our auto-encoder is to learn the representation of the input data. We hope that in the process of encoding and reconstructing the input data, the underlying structure is learned and the model after retraining on labeled data, provides better result than using the labeled data solely. Figure 17 displays the components of our network. During unsupervised training phase, we use the encoder and decoder network and during supervised training phase, we use the encoder and the classifier. In our implementation, Tensorflow [51] was used for its GPU computing power. The whole system use Sigmoid function and Gradient Descent Optimizer with Learning rate of 0.001. To avoids overfitting, we use dropout method with a keep rate of 0.75 at all the hidden layers.

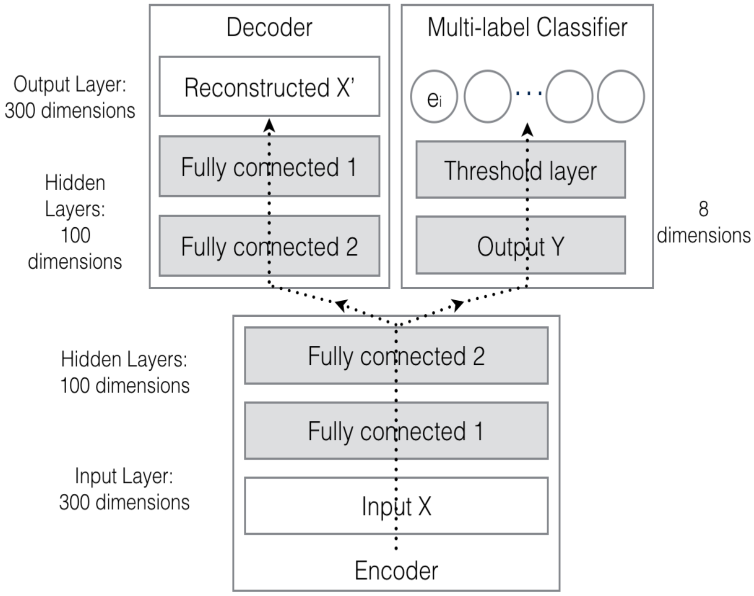


Figure 17. Structure of the Auto-encoder Deep Network

6.3.2 Unsupervised training

The encoder is a straight forward Neural Network with a 300 dimensional input layer and two 100 dimensional fully connected hidden layers. Logically, the decoder is the mirror image of the encoder with the same settings but in reverse order. Unlike the auto-encoder in Image Classifying task, we do not add noise to the network. The least square error loss function for reconstructing the input is as follow:

$$L = (X - X')^2 \tag{7}$$

Where X is input vector and X' is the reconstructed vector. For the auto-encoder to fully learn the underlying structure, we set the mini-batch size to 128 and the number of training epochs to 200.

6.3.3 Supervised retraining

After learning the representation of the input, the model is trained with our 10,000 labelled utterances. We use the bias and weights of the encoder to initialize the network. Output layer and Threshold layer are added to monitor the multi-label prediction (Figure 15. The output of our system is a set of predicted label e_i for the 8 basic emotions. The Threshold layer is a simple set of all t_i for each e_i. Let o_i be an output node of the Output layer, we have the following equation :

$$e_i = \begin{cases} 1, & \text{if } o_i \geq t_i \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The thresholds are randomly initialized and then updated after each epoch the same way we updated the biases and weights. Only the labels, whose output values greater than the corresponding threshold are considered valid. We initially decide a fixed threshold for all the emotion but soon realize that a flexible set of thresholds, one for each emotion is more effective and reasonable.

With Y is the true labels set and Y' is the set of labels predicted by our model, we define the cross entropy loss function as follow:

$$CrossEntropy = -[Y \ln Y' + (1 - Y) \ln(1 - Y')] \tag{9}$$

The global cost function is regularized by l2 regularization using the weights w_j of all l layers, the lambda λ is fixed to 0.01 in this work

$$Loss = CrossEntropy + \lambda \sum_{1}^{l} \frac{|w_j^2|}{2} \tag{10}$$

As we use only a small part of the corpus as training data, l2 regularization helps us avoiding overfitting problem.

6.4 Experiment on the proposed models

6.4.1 Evaluation metrics

In our study, We use the common evaluation metric F1-score which have been popularly used in multi-label classification problems [52, 46] are employed to measure the performance of our system. Let Y_i be set of true labels for a given instance i , and Y_i' is the set the labels predicted by a system. Let N be the total number of instances.

1. ***F*1-measure:** the harmonic mean of Precision and Recall. In our study, we gave equal importance to Precision and Recall.

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{11}$$

Precision: the fraction of correctly predicted labels over all the predicted labels in the set.

$$Precision = \frac{1}{N} \sum_{i}^{N} \frac{|Y_i \cap Y_i'|}{|Y_i'|} \tag{12}$$

Recall: the fraction of correctly predicted labels over all the true labels in the set.

$$Recall = \frac{1}{N} \sum_{i}^{N} \frac{|Y_i \cap Y_i'|}{|Y_i|} \tag{13}$$

6.4.2 Result

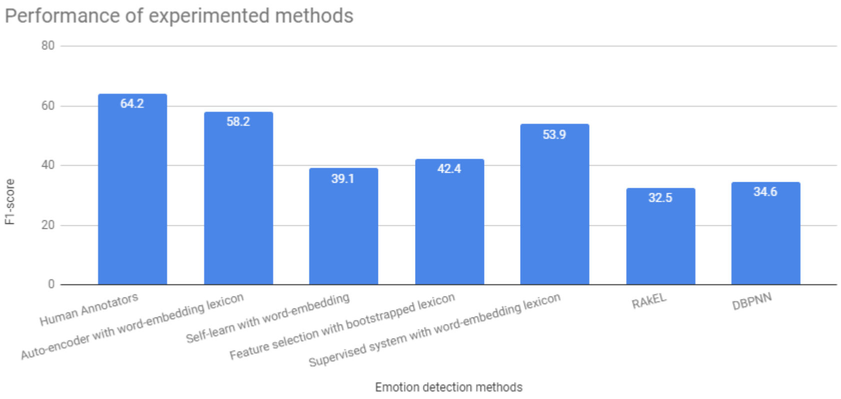


Figure 18. Evaluation of the system: 1) Human Annotators, 2) Our Auto-encoder system using Word-embedding lexicon, 3) Word-embedding lexicon with Self-learn, 4) Our system using Feature selection and bootstrapped lexicon 5) Our supervised system using Word-embedding lexicon 6)RAkEL 7)DBPNN

To evaluate our system, comparison has to be made to other systems. We replicated the works by others and applied them on our new corpus. A similar work is [7] which used the same Plutchik's theory of basic emotions and worked on multi-label data. We used similar Meka's [4] RAkEL method and Bag-of-Words

[4] `http://meka.sourceforge.net/\#about`

approach as in their work for the first baseline. We understand that while [7]'s system is fine-tuned for their corpus of user-generated movie reviews, it is a little unfair to apply it to our corpus and make comparison. In their work, they do not have to consider neither the emotions of each and every sentence nor the context information. Instead, they only work on the whole review and predict the general emotions. Therefore, the second baseline is Meka's DBPNN which is reported as generally having better performance than RAkEL [53].

We believe that the most important baseline is the human annotation. To obtain this baseline, we calculated the evaluation metrics based on the average agreement score between each annotator and the golden standard. We report the performance of our own system using different settings: word-embedding lexicon with auto-encoder and self-learn method, feature selection and domain adapted bootstrapped lexicon, supervised learning using labeled data only. Figure 18 compares the performance of our system to the baselines.

vs. Baselines: Our system, in any different settings performed remarkably better than the simple approaches using Meka's DBPNN and RAkEL. The supervised method use the same dataset as those two methods and all of them do not take advantages of the unsupervised data. Yet, it outperformed the two methods remarkably by 19 and 21 in F1-score respectively. We argue that the context features and our emotion lexicon played an important factor here.

Our system: Word-embedding Lexicon vs Bootstrapped Lexicon We can clearly see the higher performance of the word-embedding method to bootstrapped lexicon. Despite being supported by manual feature selection. The initial dimension of the bootstrapped lexicon is too low.

Our system: Semi-supervised vs Supervised The performance of the Auto-encoder is much better than that of the supervised method. Similar to an infant's learning process in its early years, our system learns and tries to make sense of the enormous unsupervised examples. In the retraining process, it uses what it has learned and makes comparison between its predictions and the true results. Therefore, it has better understanding of the data and make better predictions than using only the supervised examples. However, the performance of the Self-learn method is worse than the supervised one. We will go into details about this in the next section.

vs. Human Annotator: This is the most important baseline, which explains how well our system performs in comparison with a human. Please note that this evaluation of the human annotators is the average agreement between each annotator and the gold standard data (decided by majority rule). Our system's performance is slightly worse than that of Human Annotator *F1-measure* by 4.3. However, we should also take into consideration that the input for human annotators are movies with full video, sound signals and transcript texts; while the input for our system is only the transcripts. There is also no guarantee that the human annotators will continue to perform well together when dealing with the rest of the corpus.

6.4.3 Correlations among emotions

We are also interested in observing the correlations among emotions to verify our hypotheses about Plutchik's primary emotions: the opposite emotions and to see if our system is able capture the correlations of emotions the same way as the human does. Figure 19 illustrates the Pearson correlations among emotions. Certainly, the map is symmetrical across the diagonal line. The Pearson correlations is calculated using the following formula:

$$\tau = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{14}$$

Where x_i, y_i is the pair of binary values of two emotions in sample i, and $\bar{x}, \bar{y}$ are the average values of such emotion pairs across the dataset. In subfigure 19a, the binary values are calculated based on the golden annotations, while in 19b, the binary values are calcuated based on the label predicted by our best system. Certainly, the closer the Pearson values τ to the boundaries (-1,1) are, the stronger the correlations of the emotion become. A positive τ means that the two emotions will often appear together and a negative τ means that if one emotion appears, rarely does the other also appear. If $\tau = 0$ then the two emotions are hardly related.

By the similarity in subfigure 19a and 19b, we can safely conclude that the system do learn the correlations among emotions. This is a desirable purpose of using multi-label classification. The only significant different from the two map is the *Surprise*'s line. The system think that Surprise and other emotions are somehow related, hence the warmer and colder colors in this line.

Moreover, subfigure 19a illustrates the correlation among annotated emotions, from which we can find out about the relations among emotions and see how it corresponds to Plutchik's theory . From the correlation map, we can confirm that opposite pairs (Anger-Fear, Trust-Disgust, Joy-Sadness, Surprise-Anticipation) rarely appear together and often have the lower correlation values in their row. The exception is the pair Surprise-Anticipation where τ is very high at 0.74. This suggested that the two emotions are often mistakenly annotated together by the annotators. The system is also affected by this mistake as a result.

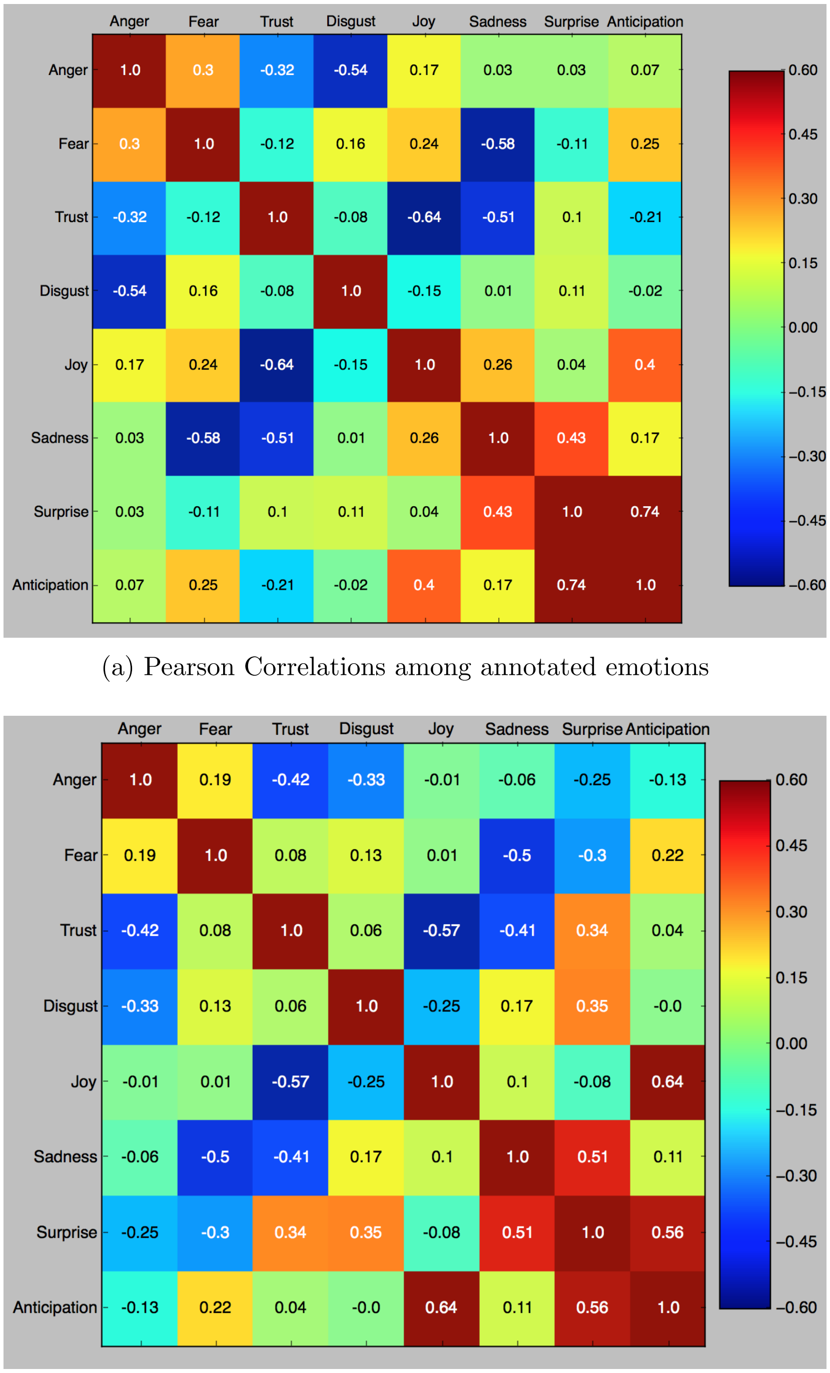
We can also observe some highly negatively related pairs such as: Disgust-Fear (-0.54), Fear-Sadness (-0.58), Joy-Trust (-0.64). Interestingly, according to Molho et.al. [54], these pairs are often felt interchangeably by us depending on the subject of impact. In other words, if one action impacts us directly, it may cause us to feel *fear*. On other hand, if it impacts other people, we may feel *sadness*. The finding suggests that there might be more dimensions than those proposed by Plutchik in emotion space.

From the figure, it can be understood that being given the ability to annotate multiple labels, the annotators are reluctant to do so. The fact that correlation map from subfigure 19a has less warm and cold colors and the low average labels per example in table 4 verifies this claim.

6.4.4 Experiment on the unsupervised data size

From the experiment, we see that the performance of the word-embedding model is better than the features selection model and the features selection model is better than the baselines. While they basically use the same approach of vectorizing the input , the Feature selection method benefits from its features such as: context, collocations list, and the set of binary features about grammar traits (subsection 6.2.1). Therefore, we need to examine if these features would improve the performance of the word-embedding model. We would also want to examine the effect of unsupervised data on our model.

In figure 20, because the word-embedding model have already included context features in its embeddings, the only difference between 2 models is that the hybrid



(a) Pearson Correlations among annotated emotions

(b) Pearson Correlations among predicted emotions (by our best system)

Figure 19. Comparison between Pearson Correlations of annotated and predicted emotions

model include collocation and binary features as mention above in subsection 6.2.1. We report the performance of the original model and hybrid model on different unsupervised data size in comparison to the training data size: equal - 10,000 utterances; ten times - 100,000 utterances; one hundred times - 1,000,000; two hundred times - 2,000,000 (all the unsupervised data). In the graph, we also add the Self-learn methods and the supervised method to highlight the reason we choose the word-embedding model using auto-encoder.

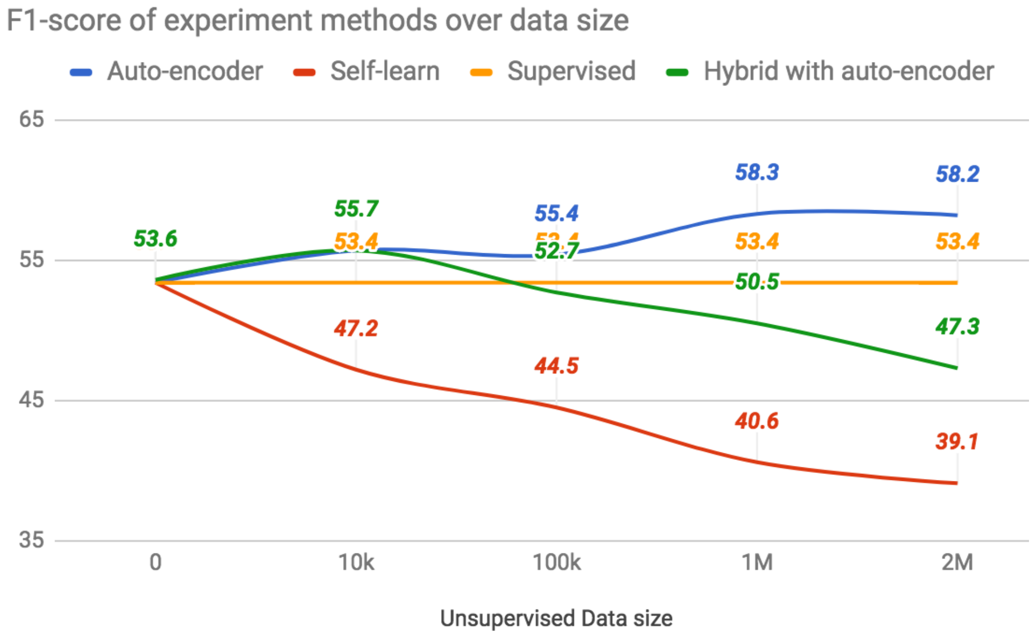


Figure 20. F1-score of auto-encoder model, hybrid auto-encoder model, supervised model and self-learn model

Self-learn vs Auto-encoder: The red line indicates the self-learn model's performance. It declines rapidly with any unsupervised data size. This suggests that the methods is naive and does not fare well with the increase of unsupervised data.

Hybrid auto-encoder model vs original auto-encoder model: At unsupervised data size 0 (No unsupervised data), original model starts at 53.4 which is the performance of supervised method. The hybrid achieve slightly higher score of 54.6. It continues to perform better than the original model at unsupervised data size of 10,000 utterances. However, from that point, its performance declines rapidly and end up at 47.3 when the data size is 2,000,000. From our point of view, the added features of the hybrid models improve its performance at first. However, when the unseen data add up, these features are ill-adapted to the new data, there are more cases when these features (clues) do not adhere to. Therefore, they become hindrances and lower the performance. We also understand that our proposed features for hybrid model are somewhat overfitting to the training data. The hybrid method benefits from the added features at first but soon loses its performance when the unsupervised data size is increased. This phenomenon urges us to go deep into the unsupervised data to find out which features (clues) are essential to identify emotions.

6.4.5 Experiment on the effectiveness of context features.

We carry on the experiment to verify the effectiveness of the context features by using two of our best systems: Autoencoder and Supervised system. In both of our systems, we integrate the context features by including the previous utterances and the whole conversation and apply text2vec. In this experiment, we nullify these two context features by removing them from the text2vec process.

Systems	**Autoencoder**	**Supervised**
Original	58.2	**53.9**
Without whole conversation feature	**58.3**	53.4
Without previous utterances feature	40.7	31.2
Without both features	40.9	31.9

Table 8. F1-score of our best systems when removing context features

From the table 8, we can safely say that the previous utterance input improve the F1-score of both of our systems while the effectiveness of whole conversation input is negligible. We think that the normalization mechanism from text2vec for the whole conversation have “over-normalized” important clues from the context: to avoid imbalance vector values for short and long utterances, we normalized the sum vector of one utterance by the length (number of words) in that utterance. This has an opposite effect when there are many emotions in a long conversation and no emotion dominates the conversation. The normalization will produce an ordinary vector.

We double confirmed this observation by examining the weights of the networks assigned to *whole-conversation* features and acknowledge that the values are insignificant comparing to the weights of *current utterances* or *previous utterances* features.This conclusion suggests that we should consider other methods of incorporate whole conversation information to the network such as LTSM architecture.

6.4.6 Examples output of the autoencoder

In this section, we demonstrate some of the prediction results of our best model. We will revisit examples **Ex.1.1**, **Ex.1.2**, **Ex.1.3** as well as some annotation example of EMTC in chapter 4. The prediction result will be discussed in the light of our model. For **Ex.1.3**, we consider each sentence to be one utterance in the conversation and feed them to the model.

From the table 9, we can see that our model is quite good with the examples **Ex.1.1**, **Ex.1.2**, **Ex.1.3** where the emotions can be inferred directly from the words. Moreover, in **Ex.1.3**, all sentences basically carry the same emotions which will stacks through the text2vec process and make it easier for the model to make the predictions.

However, our system expose its vulnerability when it encounters utterances which emotions cannot be inferred from the text. In examples EMTC1 and EMTC2, just base on the text, we can hardly make the annotation. Therefore, errors are to be expected from the model. The second vulnerability of the system is long utterances such as example EMTC3. As we have observed in the previous section, long utterances result in over-normalization in the input. At this point, the only clue for our model is the previous utterance which is falsely classified as 'Surprise' and 'Joy'. The errors accumulate and cause difficulties in the prediction of the next utterance.

	Utterances	**Annotation**	**Prediction**
Ex.1.1	I am angry now, stay away from me!	Anger	Anger
Ex.1.2	I could have wrung her neck.	Anger	Anger Disgust
Ex.1.3	My husband comes home late everyday.	Disgust	Disgust Sadness
Ex.1.3	My husband comes home late everyday. I have to do all the housework and take care of the children too.	Anger Disgust	Disgust Sadness
Ex.1.3	My husband comes home late everyday. I have to do all the housework and take care of the children too. Does he think that I am happy to do all of those by myself?	Anger Disgust	Disgust
EMTC1	Princess Isabelle: the king desires peace	Trust 0.22	Anticipation
EMTC2	William Wallace: Longshank desires peace?	Anger 0.26 Disgust 0.34 Surprise 0.36 Joy 0.22	Surprise Joy
EMTC3	William Wallace: slaves are made in such ways. the last time Longshanks spoke of peace I was a boy. and many Scottish nobles, who would not be slaves, were lured by him under a flag of truce to a barn, where he had them hanged. I was very young, but I remember Longshanks' notion of peace.	Anger 0.43 Disgust 0.15	Joy

Table 9. Example of predictions made by autoencoder

6.5 Conclusion

In short, Our best method involves building an emotion lexicon using the word2vec word-embedding technique, extracting a vectorized representation of the input, and classifying the emotions in a semi-supervised manner with the help of an autoencoder that exploits both the unlabeled and labeled data. While relying only on textual data, our system performs slightly worse than human annotators whose input is full movies' footage. The experiments show that our method's power to detect emotion is comparable to that of a human annotator despite receiving only text as input. We also understand the vulnerabilities of our model for long utterances and the accumulation of errors from previous utterances. In the future, we should consider other network architectures to avoid the over-normalization of current models for long utterances.

7. Conclusion

This dissertation investigates the marking of emotion detection system from two major perspectives. The first perspective is from the theoretical viewpoint of psychology and linguistics. We discussed the definition of emotion as well as important linguistics clues to identify emotion in text. The second part of the dissertation approach the marking of emotion detection system in an applicational manner. We studied different methods of building emotion lexicon and verify their effectiveness. The lexicon is then used for building a predicting model. We understand that a semi-supervised method using word-embedding lexicon and autoencoder take the most advantages in the task of multi-label emotion detection.

We summary the answers proposed in section 1.4 in the following section.

7.1 Summary

7.1.1 What is the origin of human emotions? How can we quantify them?

Chapter 2 presents evolutionary theory on emotions. According to the theories, emotion is a trait we inherits from animals to adapt and adjust ourselves better to the change of environment. We coordinate our perception of stimulus event and feelings with our overt behaviors to have better chance of survive and reproduce.

We follow Plutchik's wheel of emotions and proposed 8 basic emotions. These emotions may mix together and form more complex emotions called dyads. There are also different level of intensity for each basic emotions. Using the theory, we covered the full spectrum of words that express emotion in English.

7.1.2 How emotions are expressed in text documents? How to use these clues for emotion detection?

Chapter 3 mentions about the way emotions are expressed in text documents. The clues to identify them is the words and phrases that is used, the subjectivity of the documents, the contextual information and the domain of the documents.

While we discussed about sarcasm being a strong indicator of anger and frustration, we do not have a reliable sarcasm classifier for the task. We would want to revisit the problem in the future work.

7.1.3 What are the steps of building an emotion lexicon for better emotion detection system?

In chapter 4, we build our lexicon via two methods: bootstrapping and word-embedding. We argue that rather than building the lexicon from general domain such as Wordnet, we should use unsupervised method to extract the word-embedding lexicon from the corpus. The experiments results confirmed the better performance of our method.

7.1.4 What is the method to create a model that is scalable to many domains and emotions?

Throughout the dissertation, we have emphasized on the use of Plutchik's wheel of emotions. This theory covers the full spectrum of emotional space. Therefore it is scalable to the extension of emotions for classification.

For the domain problem, emotion detection suffers heavily from domain dependence problem. In our work, we have tried domain-adaptation but the improvement is subtle. To achieve business level performance, we would advice to use unsupervised methods to build both the lexicon and the model on the target domain.

7.2 Our models

Through experiments in chapter 6, we confirm that the Auto-encoder models using word-embedding method performs the best in relation to the increase of unsupervised data size. It basically meets the requirements that we have proposed in subsection 1.4: the model is able to cover the full spectrum of emotions, predicting multi-label results; it works on movies conversation domain which is very close to real-life setting; the vast unsupervised data is taken advantages of, providing a boost in the performance of the system.

7.3 Applications of the research

This research investigate and implement several approaches for emotion detection. As a result, it produces 2 multilabel emotion corpora, 2 lexicon, and 4 classifiers. This resource can be used directly or as a reference for other researchers to continue using the same approach. For our best system: the autoencoder network, we have the multilabel output of 8 basic emotions. For some specific applications in emotion detection, not all of 8 basic emotions are necessary. We hope that our work can be used as a base model or a feature extractor in those applications, where others can apply their own top layers to produce their desired output.

7.4 Future work

By building our system on movie conversation domain, we try to produce a closest imitation of real conversations. We hope that our system will perform well in real-life settings such as SMS texting, messenger applications and play a supporting role in identify emotions in speech processing. We are planing to apply our work on others domains and evaluate the results. The model and lexicon described in this paper are soon to be published on the authors' GitHub repository.

We plan to investigate the unsupervised data of the corpus more thoroughly to find out more linguistics clues that would help in the identification of emotions in text. We would also verify the existing clues to see how they contribute to the performance of the system.

We would also want to examine the relation between sarcasm and emotions and use it to improve the performance of both sarcasm detection and emotion detection. At the time of the submission, we are working on building a chat bot for the task of collection more conversation data from messengers application.

7.5 Closing remark

Emotion detection is a challenging problem in Natural Language Processing. Various approaches have been proposed but none of them are proved to be reliable for real-life application. Emotion detection will play a very important part in systems that depends on human-machine interactions. A robot that can interpret and mimic human emotions making the communication smoother and more reliable. I hope that in the future, I can continue to work on this task or some similar task that apply psychology into Natural Language Processing.

References

[1] Gary Collier. *Emotional expression.* Psychology Press, 2014.

[2] William Brant. *Critique of sarcastic reason: the epistemology of the cognitive neurological ability called "theory-of-mind" and deceptive reasoning.* [Germany]: Südwestdeutscher Verlag für Hochschulschriften, 2012.

[3] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[4] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31, 1980.

[5] Robert Plutchik. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[6] Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[7] Lars Buitinck, Jesse Van Amerongen, Ed Tan, and Maarten de Rijke. Multi-emotion detection in user-generated reviews. In *Advances in Information Retrieval*, pages 43–48. Springer, 2015.

[8] Rafael A Calvo and Sunghwan Mac Kim. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543, 2013.

[9] Taner Danisman and Adil Alpkocak. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 53, 2008.

[10] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *Text, speech and dialogue*, pages 196–205. Springer, 2007.

[11] Futoshi Sugimoto and Masahide Yoneyama. A method for classifying emotion of text based on emotional dictionaries for emotional reading.

[12] Changqin Quan and Fuji Ren. Sentence emotion analysis and recognition based on emotion words using ren-cecps.

[13] Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. Predicting and eliciting addressee's emotion in online dialogue. In *ACL (1)*, pages 964–972, 2013.

[14] Saif Mohammad. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591. Association for Computational Linguistics, 2012.

[15] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. Citeseer, 2004.

[16] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics, 2007.

[17] D.G. Myers. *Psychology, Seventh Edition, in Modules.* Worth Publishers, 2004.

[18] C. Darwin, P. Ekman, and P. Prodger. *The Expression of the Emotions in Man and Animals.* Oxford University Press, 1998.

[19] Louise Barrett, Robin Dunbar, and John Lycett. *Human evolutionary psychology.* Princeton University Press, 2002.

[20] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[21] D.L. Schacter, D. T. Gilbert, and D. M. Wegner. *Psychology (2nd Edition).* Worth, New York, 2011.

[22] James A Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999.

[23] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[24] Duc-Anh Phan, Hiroyuki Shindo, and Yuji Matsumoto. Multiple emotions detection in conversation transcripts. *PACLIC 30*, page 85, 2016.

[25] Zhongqing Wang, Sophia Lee, Shoushan Li, and Guodong Zhou. Emotion detection in code-switching texts via bilingual and sentimental information. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 763–768, Beijing, China, July 2015. Association for Computational Linguistics.

[26] Carlo Strapparava, Alessandro Valitutti, and Oliviero Stock. The affective weight of lexicon.

[27] Clifford N Lazarus. Think sarcasm is funny? think again: Sarcasm is really just hostility disguised as humor, 2012.

[28] Bing Liu. Sentiment analysis and subjectivity.

[29] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[30] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

[31] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of ACL*, pages 271–278, 2004.

[32] Saif M Mohammad. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings*

of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 246–255. Association for Computational Linguistics, 2012.

[33] Wonhee Choe, Hyo-Sun Chun, Junhyug Noh, Seong-Derok Lee, and Byoung-Tak Zhang. Estimating multiple evoked emotions from videos. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.

[34] Justin Kruger, Nicholas Epley, Jason Parker, and Zhi-Wen Ng. Egocentrism over e-mail: Can we communicate as well as we think? *Journal of personality and social psychology*, 89(6):925, 2005.

[35] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[36] J Cohen. Kappa: Coefficient of concordance. *Educ. Psych. Measurement*, 20:37, 1960.

[37] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[38] 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 the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[39] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[40] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[41] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.

[42] Saurin Dave and Hiteishi Diwanji. Trend analysis in social networking using opinion mining a survey. 2015.

[43] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. 2011.

[44] Roger Bougie, Rik Pieters, and Marcel Zeelenberg. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the Academy of Marketing Science*, 31(4):377–393, 2003.

[45] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.

[46] Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. Sentence-level emotion classification with label and context dependence. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1045–1053, Beijing, China, July 2015. Association for Computational Linguistics.

[47] S Bird, E Klein, and E Loper. Nltk book, 2009.

[48] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[49] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *Knowledge and Data Engineering, IEEE Transactions on*, 18(10):1338–1351, 2006.

[50] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–161. Association for Computational Linguistics, 2011.

[51] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[52] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*, pages 22–30. Springer, 2004.

[53] Pablo Fernandez-Gonzalez, Concha Bielza, and Pedro Larranaga. Multi-dimensional classifiers for neuroanatomical data. In *ICML Workshop on Statistics, Machine Learning and Neuroscience (Stamlins 2015)*, 2015.

[54] Catherine Molho, Joshua M Tybur, Ezgi Güler, Daniel Balliet, and Wilhelm Hofmann. Disgust and anger relate to different aggressive responses to moral violations. *Psychological science*, page 0956797617692000.

List of publications

Name: Phan Duc Anh

Peer review journal paper (Author[underline your own name], Thesis title, Journal title, Volume, Number, Pages, Date, The chapter or section of your dissertation which this paper relates)

1. Duc-Anh Phan, Hiroyuki Shindo, Yuji Matsumoto, Autoencoder for Semisupervised Multiple Emotion Detection of Conversation Transcripts, IEEE Transactions on Affective Computing, *Volume, Issues and Pages are not yet determined (Early Access)*, DOI 10.1109/TAFFC.2018.2885304, Dec 10th 2018, Chapter IV, V, VI
2.

Peer review journal paper (letters, technical reports etc. The format is the same as above.)

1.
2.

Peer review international conference (Author[underline your own name],Thesis title, Journal title (or conference name), (Volume, Number, Pages), Date, The chapter or section of your dissertation which this paper relates)

1. Phan Duc-Anh and Yuji Matsumoto, EMTC: Multi-label Corpus in Movie Domain for Emotion Analysis in Conversational Text, In Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018): p2641-2646, May 10th 2018, Chapter IV, V, VI
2. Duc-Anh Phan, Hiroyuki Shindo, Yuji Matsumoto, Multiple Emotions Detection in Conversation Transcripts, In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30): p 85-94, Oct 28th 2016, Chapter IV, V, VI

※If you can not fit everything on one page, include your most important work, and indicate the number of papers that you do not include.