



Universitat d'Alacant
Universidad de Alicante

Tackling the Challenge of
Emotion Annotation in Text

Lea Canales Zaragoza



Tesis

Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE



Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

Tackling the Challenge of Emotion Annotation in Text

Lea Canales Zaragoza

Tesis presentada para aspirar al grado de

DOCTOR POR LA UNIVERSIDAD DE ALICANTE

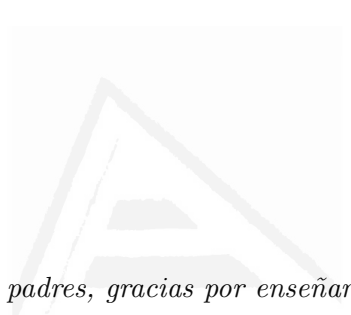
MENCIÓN DE DOCTOR INTERNACIONAL

DOCTORADO EN INFORMÁTICA

Dirigida por

Dr. Patricio Martínez Barco

Esta tesis ha sido financiada por el Ministerio de Economía y Competitividad a través del programa de ayudas para la Formación de Personal Investigador (Ref. BES-2013-065950).



*A mis padres, gracias por enseñarme a ser una luchadora
A mi hermana y mis hermanos, por vuestro apoyo incondicional
A Dani, por hacer que esta locura no acabara conmigo*

Universitat d'Alacant
Universidad de Alicante

Acknowledgements

El desarrollo de una tesis doctoral no es un camino de rosas y tengo claro que no hubiera llegado a este punto sin la ayuda de mucha gente a la que quiero agradecer enormemente su apoyo.

En primer lugar, quiero dar las gracias a mis directores de tesis, Patricio Martínez Barco y Ester Boldrini. Patricio, muchas gracias por darme la oportunidad de desarrollar mi tesis doctoral, por tu apoyo y por animarme a hacer una estancia predoctoral en otro país que me ha permitido conocer nuevos horizontes. Ester, aunque oficialmente no apareces como directora, por cuestiones burocráticas, para mí lo has sido. Gracias por tus millones de correcciones, por sacar tiempo de tu ajetreado trabajo para leer nuestros artículos, por tus consejos y por tus ánimos cuando mi cara no mostraba excesiva felicidad. *Grazie mille!*

Por supuesto, también quiero agradecer a todos los miembros del grupo de investigación de Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI) de la Universidad de Alicante. Especialmente a Paloma Moreda, por animarme a descubrir este mundo loco y a la vez apasionante de la investigación.

Debo hacer una mención especial a mis chicas, Isabel Moreno y Marta Vicente. ¡Millones de gracias por todo! Sin duda, habéis sido un apoyo fundamental en este largo camino. Gracias por nuestros momentos de "terapia", por vuestra visita a Amberes, por vuestros millones de mensajes de apoyo y ánimo durante todo este tiempo.

Otra persona muy importante en momentos clave ha sido Miguel Ángel Guerrero, nuestro arquitecto infiltrado en un grupo de informáticos. Muchísimas gracias por tus consejos, por tu ayuda para que mantuviese la cordura cuando no era capaz de pensar con claridad y enseñarme a ver los problemas

Acknowledgements

con diferente perspectiva. ¡No sabes cuánto te echamos de menos en el laboratorio!

No me olvido de la gente con la que he compartido momentos en el laboratorio: José Manuel Gómez, Elena Lloret, Javier Fernández, Fernando Peregrino, Jorge Cruañes, Yoan Gutiérrez, Antonio Guillén, Cristina Barros, Paco Agulló, Jorge Torregrosa, Ulises Serrano, Bea Botella, María de los Ángeles Herrero, Saray Zafra, Alejandro Reyes, Luke Blanes y a mis dos cubanos favoritos, Suilan Estévez y Alejandro Piad. Suilan, millones de gracias por quedarte aquella noche conmigo para que terminara la primera versión de la tesis, sin ti no lo hubiera conseguido.

Furthermore, I am greatly thankful to everybody in the HLT-NLP group in Fondazione Bruno Kester (FBK) and Computational Linguistics & Psycholinguistics research group (CLiPS) in the University of Antwerp for having received me and having made my stays so interesting. Especially, I would like to thank Carlo Strapparava and Walter Daelemans for having invited me to be part of that family and for having shared their time and knowledge with me.

De estas estancias, además de haber sido grandes experiencias, me he llevado una gran amiga y compañera, mi *bella* Anna Feltracco. No sé cómo agradecerte toda tu ayuda durante mis estancias en Italia y después de estas. Sin duda, tú hiciste que mi tiempo allí mereciera la pena. Muchas gracias por todo tu apoyo durante estos años. ¡Te espero por las playas de Alicante!

Tampoco me olvido de otros compañeros de la Universidad de Alicante. Javier Sober, mi compañero desde el primer momento que empecé la carrera, gracias por ayudarme y apoyarme a pesar de no tener mucha idea de informática cuando entré en la universidad. A mis compañeros de doctorado, Daniel Torregrosa y José Javier Valero. José Javier, *moltíssimes gràcies pel teu suport en la recta final de la tesi, cadascun dels teus missatges em donaren molta força per afrontar la escriptura de la tesi.*

También quiero agradecer a Irene Molina y María Leal, mis anotadoras de emociones que decidieron participar en uno de mis experimentos de forma totalmente desinteresada y ayudarme para que pudiera llevar a cabo una de mis ideas. Chicas, muchísimas gracias porque me ayudasteis a seguir adelante.

Además, he de agradecer a mi familia y amigos que son pilares fundamen-

tales en mi día a día. Gracias a mis amigas María Mollá, Lidia Fernández, Laura Molina, Beatriu Pla y Ester Onteniente, por vuestra ayuda, comprensión y por vuestros mensajes de ánimo durante toda la tesis, tanto cuando he estado por tierras alicantinas como cuando he estado perdida por el mundo.

A mis padres, Joaquín y Toñi, que necesitaría millones de folios para agradecerles todo lo que han hecho por mí. Estoy donde estoy gracias a ellos. *Papà, moltíssimes gràcies per les xarretes sobre la tesi i per sempre oferir-me el teu suport per a tot, per transmetre'm calma i serenitat quan estic nerviosa i per veure l'aspecte positiu de tot en la vida. Mamà, moltíssimes gràcies per la teua paciència, per escoltar-me tots els dies, per animar-me a seguir, per fer-me creure que sóc la millor del món i perquè sempre tens un somriure per a mi.*

A mi hermana, María, mi compañera de batallas, una luchadora nata y un ejemplo a seguir. Muchísimas gracias por estar siempre a mi lado, tanto en lo bueno como en lo malo. Siempre te has preocupado por mí, me has mandado mensajes de apoyo, has venido a verme cuando he estado fuera y has estado al otro lado del teléfono para todo. Me siento la persona más afortunada.

A mi hermano pequeño, Marcos, por tu apoyo y ánimo en todo momento, por protegerme y preocuparte por mí como un hermano mayor. Eres una persona muy importante en mi vida y te necesito a mi lado. Muchas gracias por todo.

A mi hermano mayor, Jose, por preocuparte en mis viajes y por apoyarme "a tu manera" durante este tedioso camino.

A mi cuñado, Daniel Jerez, por preocuparte por mí como otro hermano, por venir a verme al extranjero y por tu gran apoyo y ayuda en este arduo camino.

Y por último, pero no menos importante, a Dani. Mi compañero de vida, mi pareja, mi confidente, mi amigo y mi gran apoyo. Eres el coautor de esta tesis y, sin duda, uno de los más perjudicados porque has tenido que aguantarme en los malos momentos y te he robado mucho tiempo para dedicárselo al trabajo. Millones de gracias por tu ayuda, paciencia y comprensión que han hecho que esta aventura no acabara conmigo. Prometo compensarte en esta nueva etapa. Te quiero.



Universitat d'Alacant
Universidad de Alicante

The work in this dissertation has been carried out at the Language Processing and Information Systems Group of the University of Alicante (Spain) between February 2014 and May 2018, with research stays from April 2015 to June 2015 and April 2016 to June 2016 at the Fondazione Bruno Kessler of Trento (Italy), as well as March 2017 to June 2017 at Computational Linguistics & Psycholinguistics research center of the University of Antwerp (Belgium). This research has been supported by the FPI grant (BES-2013-065950) and the research stay grants (EEBB-I-15-10108, EEBB-I-16-11174, and EEBB-I-17-12578) from the Spanish Ministry of Science and Innovation.

Contents

List of Figures	viii
List of Tables	xi
Acronyms	xiv
1 Introduction	1
1.1 Motivation	3
1.2 Problem Definition and Scope	4
1.3 Thesis Structure	5
2 Background in Emotion Resources	7
2.1 Emotion Theories	8
2.1.1 Categorical Emotion Models	9
2.1.2 Dimensional Emotion Models	11
2.1.3 Categorical vs Dimensional	14
2.2 Emotion Lexicons	15
2.2.1 Manual Creation Lexicons	15
2.2.2 Semi-Automatic/Automatic Creation Lexicons	18
2.2.3 Discussion	20
2.3 Emotion Corpora	22
2.3.1 Manual Corpora Annotation	23
2.3.2 Semi-Automatic/Automatic Corpora Annotation	32

2.3.3	Discussion	38
2.4	Conclusion	40
3	Background in Annotation Techniques	43
3.1	Bootstrapping Technique for Intensional Learning	44
3.2	Pre-annotation Process	47
3.3	Conclusion	51
4	Intensional Learning for Emotion Annotation	53
4.1	Intensional Learning Process	54
4.1.1	Step 1.1: Selecting of Seed Sentences	55
4.1.2	Step 1.2: Seed Extension via Semantic Similarity	61
4.1.3	Step 2: Training supervised classifiers	64
4.2	Evaluation	65
4.2.1	Data Description	65
4.2.2	Methodology	67
4.2.3	Results	68
4.2.4	Analysis	72
4.3	Conclusion	74
5	EmoLabel: Semi-Automatic Methodology for Emotion An- notation	77
5.1	Phase 1: Pre-annotation Process	78
5.1.1	<i>Unsupervised</i> Pre-annotation	79
5.1.2	<i>Supervised</i> Pre-annotation	85
5.2	Phase 2: Manual Refinement	88
5.3	Evaluation	91
5.3.1	Data Description	91
5.3.2	Intrinsic Evaluation	93
5.3.3	Extrinsic Evaluation	98

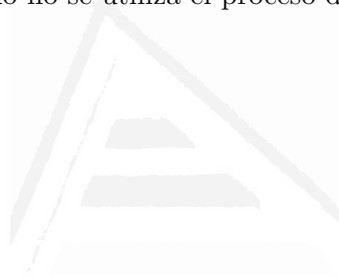
5.4	Conclusions	100
6	Conclusions and future perspectives	103
6.1	Contributions	105
6.1.1	Publications	107
6.2	Future work	109
A	Annotation Guidelines	113
B	Resumen	119
B.1	Introducción	119
B.2	Contribuciones	124
B.2.1	<i>Intensional Learning</i> para la anotación de emociones .	124
B.2.2	<i>EmoLabel</i> : metodología semi-automática para la anotación de emociones	128
B.3	Conclusiones y trabajo futuro	135
	References	141

Universitat d'Alacant
Universidad de Alicante

List of Figures

2.1	Plutchik’s Wheel of Emotions.	10
2.2	The Circumplex Model of Affect.	12
4.1	Overview of the Intensional Learning (IL) process.	55
4.2	Workflow of the step 1.1 in Intensional Learning (IL) process.	57
4.3	Examples of the process of selecting seed sentences (Step 1.1).	59
4.4	Process of the extension of NRC Emotion Lexicon (EmoLex) by WordNet (WN) and Oxford synonyms.	60
4.5	Workflow of the step 1.2 in Intensional Learning (IL) process.	63
4.6	Workflow of the step 2 in Intensional Learning (IL) process.	65
5.1	Overview of EmoLabel methodology.	78
5.2	Overview of pre-annotation process (Phase 1).	79
5.3	Overview of <i>unsupervised</i> pre-annotation process.	80
5.4	Graphical representation of the pre-processing step (1.1).	82
5.5	Overview of <i>supervised</i> pre-annotation process.	87
5.6	Overview of manual refinement (Phase 2).	88
5.7	An example of how sentences are shown to the human annotators when the pre-annotation process is employed.	89
5.8	An example of how sentences are shown to the human annotators when the pre-annotation process is no employed.	89

A.1	An example of how the sentences are shown to the human annotators when the pre-annotation process is employed. . . .	116
A.2	An example of how the sentences are shown to the human annotators when the pre-annotation process is no employed. . .	116
B.1	Descripción general del proceso bootstrapping basado en Intensional Learning (IL).	126
B.2	Descripción general de la metodología <i>EmoLabel</i>	129
B.3	Un ejemplo de como se muestran las frases a los anotadores humanos cuando se utiliza el proceso de pre-anotación. . . .	132
B.4	Un ejemplo de como se muestran las frases a los anotadores humanos cuando no se utiliza el proceso de pre-anotación. . .	132



Universitat d'Alacant
Universidad de Alicante

List of Tables

2.1	Emotion Categories identified by researchers.	11
2.2	Emotion Dimensions identified by researchers.	13
2.3	Emotion Lexicons	21
2.4	Categorical Emotion Corpora (manual annotation)	29
2.5	Dimensional Emotion Corpora (manual annotation)	31
2.6	Categorical Emotion Corpora (semi/automatic annotation)	37
2.7	Dimensional Emotion Corpora (semi/automatic annotation)	38
4.1	Distribution of the emotion words in the reduced version of EmoLex	57
4.2	Distribution of the emotion words in the enriched version of EmoLex with WN and Oxford synonyms.	61
4.3	Distributional Semantic Models (DSMs) parameters for IL approaches	64
4.4	Distribution of the sentences per emotion on Aman corpus, a corpus of blog posts annotated with Ekman’s basic emotions.	66
4.5	Distribution of the sentences per emotion on Affective Text corpus, a corpus of headlines annotated with Ekman’s basic emotions.	66
4.6	Results for the Sequential Minimal Optimization (SMO) multi-classifier on the gold standard of Aman corpus and the six SMO binary-classifiers on the gold standard of Affective Text corpus. Precision, recall, F1-score per class and their macro-average scores.	68

4.7	Results for the SMO multi-classifier trained on the corpus developed applying Latent Semantic Analysis (LSA) and ukWak Word2Vec (W2V) (Continuous Bag-Of-Words (CBOW)) models on Aman corpus. Precision, recall, F1-score per class and their macro-average scores.	69
4.8	Results for the SMO multi-classifier on the corpus developed applying Gigaword W2V (CBOW & Skip-gram (SKIP)) models on Aman corpus. Precision, recall, F1-score per class and their macro-average scores.	69
4.9	Results for the SMO six binary-classifiers on the corpus developed applying LSA and ukWak W2V (CBOW) models on Affective Text Corpus. Precision, recall, F1-score per class and their macro-average scores.	70
4.10	Results for the SMO six binary-classifiers on the corpus developed applying Gigaword W2V (CBOW & SKIP) models on Affective Text Corpus. Precision, recall, F1-score per class and their macro-average scores.	70
4.11	Inter-Annotator Agreement (IAA) in terms of Cohen’s kappa on the comparison of the annotation of the <i>Original</i> and <i>Enriched</i> approaches to the gold standard of Aman corpus. .	71
4.12	IAA in terms of Cohen’s kappa on the comparison of the annotation of the <i>Original</i> and <i>Enriched</i> approaches to the gold standard of Affective Text Corpus.	71
5.1	Distribution of the emotion words annotated with only one emotion in the resultant lexicon (<i>EmoSenticNet</i> + <i>EmoLex</i>) .	82
5.2	DSMs features for EmoLabel approaches	84
5.3	Examples of <i>unsupervised</i> pre-annotation process. The <i>1st ranking</i> column shows the order proposed by the system before employing the polarity and subjective information. The <i>Emotion proposed</i> column shows the pre-annotated emotions by the system after re-ordering the first ranking.	86
5.4	Cross-validation setup	90

5.5	IAA in terms of Fleiss' kappa between the three annotators in each training task.	90
5.6	IAA in terms of Fleiss' kappa between each annotator and the Aman corpus' gold standard in each training task.	91
5.7	Distribution of the sentences per emotion on EmoTweet-5, a reduced version of EmoTweet-28 that contains tweets annotated with Ekman's basic emotions.	92
5.8	Results for the <i>unsupervised</i> pre-annotation using different distributional representations on Aman corpus. Precision, recall, F1-score per class and their macro-average scores. . . .	94
5.9	Results for the <i>unsupervised</i> pre-annotation using different distributional representations on EmoTweet-5 corpus. Precision, recall, F1-score per class and their macro-average scores. . . .	95
5.10	Results for the <i>supervised</i> pre-annotation using different set of features on Aman corpus. Precision, recall, F1-score per class and their macro-average scores.	97
5.11	Results the <i>supervised</i> pre-annotation using different set of features on EmoTweet-5 corpus. Precision, recall, F1-score per class and their macro-average scores.	98
5.12	Distribution of the number of sentences per emotion annotated in each manual task.	98
5.13	IAA in terms of Fleiss' kappa between each annotator and the Aman corpus' gold standard.	99
5.14	Annotation time of each annotator in all manual annotation tasks.	100
B.1	Configuración de la <i>Validación cruzada</i>	133

Acronyms

AC	Affective Computing
ADEs	Adverse Drug Events
AI	Artificial Intelligence
AL	Active Learning
ANET	Affective Norms for English Text
ANEW	Affective Norms for English Words
ANPST	Affective Norms for Polish Short Text
AMT	Amazon Mechanical Turk
ASRL	Automatic Semantic Role Labeling
BOW	Bag-Of-Words
BNC	British National Corpus
CANEW	Categorical Affective Norms for English Words
CBOW	Continuous Bag-Of-Words
CDSMs	Compositional Distributional Semantic Models
CF	CrowdFlower
CL	Computational Linguistics
CMC	Computer-Mediated Communication
CVAT	Chine Valence-Arousal Text

CS	Computer Science
DAL	Dictionary of Affect in Language
DL	Deep Learning
DSMs	Distributional Semantic Models
EmoLex	NRC Emotion Lexicon
ESN	EmoSenticNet
EARL	Emotion Annotation and Representation Language
E-ANEW	Extended Affective Norms for English Words
ER	Emotion Recognition
EL	Extensional Learning
EM	Expectation-Maximization
F8	Figure Eight
GEW	Geneva Emotion Wheel
GALC	Geneva Affect Label Coder
HCI	Human-Computer Interaction
HEC	Hashtag Emotion Corpus
IAPS	International Affective Picture System
IADS	International Affective Digitized Sounds
IAA	Inter-Annotator Agreement
IL	Intensional Learning
ISEAR	International Survey on Emotion Antecedents and Reactions
LSA	Latent Semantic Analysis
LIWC	Linguistic Inquiry and Word Count
MASC	Manually Annotated Sub-Corpus of the American National Corpus

ML	Machine Learning
MAE	Mean Absolute Error
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
OP	Opinion Mining
PAD	Pleasure - Arousal - Dominance
PANA	Positive Activation - Negative Activation
PANAS-X	Positive and Negative Affect Schedule-Expanded
POS	Part-Of-Speech
RMSE	Root Mean Square Error
SA	Sentiment Analysis
SAM	Self-Assessment Manikin
SKIP	Skip-gram
SMO	Sequential Minimal Optimization
SREC	Sport-related Emotion Corpus
SRL	Semantic Role Labeling
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TC	Text Categorization
TEC	Twitter Emotion Corpus
UI	User Interface
VAD	Valence-Arousal-Dominance
VSM	Vector Space Model

WN	WordNet
WNA	WordNet-Affect
WSD	Word Sense Disambiguation
W2V	Word2Vec



Universitat d'Alacant
Universidad de Alicante

Introduction

*“Your intellect may be confused, but
your emotions will never lie to you.”*

ROGER EBERT

As an essential part of social interactions between humans, emotion analysis has been a captivating topic in disciplines such as neuroscience, cognitive studies, psychology or behavioral science. This interest has also attracted the attention of Artificial Intelligence (AI) since emotions are crucial to improve user experience in Computer-Mediated Communication (CMC) and Human-Computer Interaction (HCI) (Cowie et al., 2001). In this iteration, language plays an important role.

Language is a medium of human communication, both spoken or written, to express our ideas, our thoughts and most importantly, our emotions. Based on the communicative function model of the language defined by Jakobson (1960), it may be observed the importance of the interrelation of language and emotion. He identifies emotive function as one of the six roles of the language. It is, therefore, a powerful tool to communicate and convey emotion information.

In HCI, emotion analysis has been analyzed through various sensor channels on User Interface (UI) such as facial expression, speech, and text (Kim, 2011). The importance of text as a medium of communication has increased with the use of computers and, notably, with the appearance of

the Web 2.0 or Social Web. Unlike Web 1.0 where users were limited to the passive viewing of content, Web 2.0 websites allow communicating and information sharing on the Internet using computers, mobile phones or any internet connected device. There are many social media platforms such as Facebook¹, Instagram² or YouTube³ where people exchange different types of content (messages, photos, videos, etc.); Blogs platforms as Blogger⁴ or WordPress⁵ where people post chronological publications of personal thoughts or other information; or Microblogging services as Twitter⁶ that are blogs where users share small elements of content (sentences, individual images, or video links) (Kaplan & Haenlein, 2011).

As shown by statistics, the social media phenomenon has expanded throughout the world and quickly attracted billions of users (Farzindar & Inkpen, 2015). For instance, the last ranking for social network published by Statista⁷, the world's largest statistic portal, in January 2018, placed Facebook at the top of the ranking with 2,167 million of active users, YouTube is in the second place with 1,500 and Instagram has over 800 million active users (7th place). As a consequence, there has been an exponential growth in the amount of subjective information on the Web 2.0 due to this massive use of these social media services by users.

Parallel to the growth of the subjective information, there has been an increasing interest from Natural Language Processing (NLP) researchers to develop methods to automatically extract knowledge from these new sources. NLP research field deals with the interactions between human language and computers, aiming to communicate machines and humans by using natural language. Given the importance of emotions in language, within NLP has emerged a subtask concerned with the identification and extraction of affective states and subjective information in text, called Sentiment Analysis (SA).

The basic goal of SA is to identify sentiments, opinions, and emotions from text. Most of works in this field has typically focused on recognizing the polarity of sentiment (POSITIVE, NEGATIVE, or NEUTRAL), and these works

¹<https://www.facebook.com/>

²<https://www.instagram.com/>

³<https://www.youtube.com/>

⁴<https://www.blogger.com>

⁵<https://www.wordpress.com/>

⁶<https://twitter.com/>

⁷<https://www.statista.com/>

are framed into Opinion Mining (OP) task. However, the recognition of types of emotions such as emotional categories (ANGER, DISGUST, FEAR, etc.) or emotional dimension (*valence, arousal, dominance*, etc.) has recently increased since recognizing emotions conveyed by a text can lead to better understanding of the text's content (Aman, 2007). This analysis is known as Emotion Recognition (ER) and is where this work is framed.

There has recently been an increasing interest in textual ER from the research community mainly due to the appearance of the new genres of Web 2.0 and its potential of bringing substantial benefits to different sectors as suicide prevention (Cherry et al., 2012; Desmet & Hoste, 2013), identification cases of cyberbullying (Dadvar et al., 2013), or contribution towards the improvement of student motivation and performance (Suero Montero & Suhonen, 2014).

1.1 Motivation

Different are the techniques applied by NLP researchers to tackle textual ER task, including the use of machine learning, rule-based methods and lexical approaches. However, the majority of such proposals have been performed with machine learning algorithms mainly due to their scalability, learning capacity and fast development.

Machine Learning (ML) is a scientific discipline that deals with the construction and study of algorithms that can learn using *experience*. This is data to improve the performance or to make accurate predictions (Mohri et al., 2012). The data available for analysis (called training data) should be labeled when *supervised learning* is employed whereas *unsupervised learning* receives unlabeled data. The common scenario in textual ER is the use of *supervised learning* since these algorithms lead to better results than the rest of alternatives.

Focusing on ER in text, *supervised ML* algorithms consist of inferring a function from a set of examples labeled with the correct emotion (labeled corpus or training data). After this, the model is able to predict the emotion of new examples. The success of the predictions made by the model will directly depend on the quality and the size of our training data. Hence, the training dataset employed is crucial to building accurate emotion detection

systems that can generate reliable results.

This requirement of quality and size of training data is even more important in the new discipline called Deep Learning (DL). It is part of a broader family of ML that utilizes a hierarchical level of artificial neural networks to perform the process of ML (Deng & Yu, 2014). One of the most relevant features of these networks is that it does not require task-specific feature engineering. However, this characteristic implies that the training of a DL architecture requires larger amounts of data than a traditional ML algorithm.

However, the creation of a labelled corpus for textual ER is not trivial, since detecting emotion in text can be difficult even for humans because everyone's personal context can influence emotion interpretation. Most relevant research carried out so far has shown difficulties related to this task, such as obtaining a good Inter-Annotator Agreement (IAA) or the time required for its development. As a consequence, data gathering with emotion content has become one of the most challenge tasks in textual ER.

1.2 Problem Definition and Scope

Considering the difficulties of textual ER research and in order to lessen and counteract the challenge of emotion annotation, this research addresses the task of analyzing ways of improving the emotion labelling with semi-automatic techniques. More specifically, two techniques, whose usability and effectiveness have been demonstrated in other NLP tasks, have been investigated: bootstrapping for Intensional Learning (IL) and a pre-annotation process.

These techniques have been assessed with the aim of providing a method able to efficiently annotate a large amount of English data in any genre and with robust standards of reliability. These requirements increase the difficulty of the task since it has been tackled from a general point of view, this is, independent of genres and the set of emotional tags employed.

The emotion annotation task is carried out at sentence level because, in genres such as blogs or tales, a finer-grained level of analysis is beneficial since there is often a progression of emotions in narrative text (Kim, 2011). Moreover, in social networks such as Twitter or Facebook, people share their

opinions or emotions through small elements of content (sentences, images or videos). The possible labels are the six basic emotions proposed by Ekman (1992): ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE because of it has been extensively used in other emotion computational approaches and these emotions have been most widely accepted by the different researchers, as we will see in the next chapter. There are different perspectives from which emotions in text can be analyzed: *written*, *reader* and *text*. *Written* perspective is referred to how someone feels while producing an assertion whereas *reader* perspective is how someone feels after reading this utterance. And in *text* perspective, no actual person is specified as perceiving an emotion and emotion is an intrinsic property of a sentence. Our approaches have been developed considering the *text perspective* since our objective is to analyze the emotional orientation as much as is evident from the written text, without considering the emotional context of writer or reader.

1.3 Thesis Structure

This work has been divided into 6 chapters. Each one has a parallel structure composed by an introduction, the development of the chapter and conclusions.

After the introduction of our research in chapter 1, the following chapters are:

- **Chapter 2: Background in Emotion Resources.** Presents the research underlying the two emotion models employed to represent the affective state and the state of the art in the resources available to tackle textual ER task: emotion lexicons and emotion corpora. These emotion resources are classified according to their creation process (manual or semi/automatic) and their emotion connotation (categorical and dimensional models).
- **Chapter 3: Background in Annotation Techniques.** Provides a review of annotation techniques employed in other NLP disciplines with the aim of simplifying and improving the annotation process and thus reducing time and cost of its development. This study has as main objective exploring alternative annotation techniques to tackle textual emotion labeling. Concretely, two annotation techniques have

been explored: bootstrapping technique for **IL** and a pre-annotation process.

- **Chapter 4: Intensional Learning for Emotion Annotation.** Describes our first proposal to efficiently tackle emotion annotation: bootstrapping technique for **IL**. First, the method is described in detail, then is explained the evaluation carried out, as well as the results and the conclusions drawn from this experiment.
- **Chapter 5: EmoLabel: Semi-Automatic Methodology for Emotion Annotation.** Presents our second proposal in order to improve emotion annotation task: EmoLabel. It is a semi-automatic methodology based on an automatic pre-annotation process. First of all, we will detail the entire process, starting with a complete description of each phase of EmoLabel, following up with the evaluation performed, and finally the conclusions drawn from this experimentation.
- **Chapter 6: Conclusions and future perspectives.** Summarizes the main conclusions of this research work and the main contributions of this thesis. It also addresses some issues that will be faced in the futures. Finally, a list of the relevant publications is also provided.
- **Chapter A: Annotation Guidelines.** Presents the annotation guidelines employed in the second phase of EmoLabel, a manual refinement process where humans annotations determine which is the dominant emotion for each sentence.
- **Chapter B: Resumen.** Provides a summary of the thesis in Spanish. This outline offers a general overview of our work underlying our main contributions and finding, as well as it explains the most relevant experiments carried out and the results obtained.

Background in Emotion Resources

“Data is a precious thing and will last longer than the systems themselves.”

TIM BERNERS-LEE

Emotions have been widely studied in psychology and behavior science, as they are an important element of human nature (Strapparava & Mihalcea, 2014). Even though it is more of an interest to researchers in social and behavioral sciences, emotions have also attracted the attention of researchers in Computer Science (CS) and thus it is becoming a multi-disciplinary research area.

This work is framed within Sentiment Analysis (SA) discipline that studies and treats subjective language. SA is part of the broader area of Affective Computing (AC) which aims to enable computers to recognize and express emotions (Picard, 1997). Within SA, it is possible to differentiate between two tasks: Opinion Mining (OP) and Emotion Recognition (ER). OP can be defined as the task that automatically detects opinion expressed in texts and classifies it depending on its polarity (positive, negative or neutral). While ER is a task more specific than opinion analysis that takes a document (sentence, phrase or word) and classifies it into one of several emotion classes, depending on the underlying emotion theories employed.

Even if it is a relatively recent area, the automatic detection of emotion in text is an active research field where a variety of tools and methods have been developed with the aim of tackling this task. In spite of this, the effective analysis and treatment of subjective data still represent an important challenge to overcome, since textual emotion detection task presents inherent problems. One of the most challenging is the building of emotion resources since emotion detection is a difficult task, even for humans.

Related to this, and given their utmost importance, in this chapter are presented an exhaustive overview above all of the resources (lexicons and corpora) to analyze the emotional content of text. To properly introduce emotion resources, the emotion theories or the emotion frameworks outlines by psychologists for representing the emotions are firstly described in Section 2.1. The other two sections: emotion lexicon (Section 2.2) and emotion corpora (Section 2.3) have as an objective present an extensive review of existing emotion resources considering the emotion model employed for its development, as well as its creation process. The final part of the chapter (Section 2.4) summarizes the most important open issues and aspects to improve in this research framework.

2.1 Emotion Theories

Despite the fact that there is no a general consensus among psychologists on the definition of emotion or how many emotions are there, research in psychology outlines two main approaches to represent the emotions that humans perceive and express: the *categorical* model (the discrete emotions approach) and the *dimensional* one (Scherer, 2005).

This section outlines the most important emotion models for the categorical approach (Section 2.1.1) and the dimensional one (Section 2.1.2), as well as the advantages and disadvantages of each approach. The final part of this section (Section 2.1.3) presents the possibility of using both models in a computational approach and the emotion framework most popular in Computational Linguistics (CL).

2.1.1 Categorical Emotion Models

The categorical model assumes that there are discrete emotional categories, this is, conceptualizes emotion as a set of distinct categories.

The methodology used in this approach consists in asking a respondent to classify a document (text of any length) in one or more emotion categories from the ones previously established. This group of categories can be determined based on a particular theory or by creating ad hoc list of emotion categories that seem relevant in a specific research context.

The words to analyze the human emotional experience have been used by philosophers since the dawn of behavioral science. This is why, this approach has a serious scientific history. The definition of a set of "basic emotions" by Darwin (1998) through identifying observable physiological and expressive symptoms that accompany this set of emotions has led the categorical approach to be accepted for the biological and social science (Scherer, 2005). Following this study, Ekman (1971, 1992) also concluded that some emotions are universal and innate after studying an isolated tribe in Papua New Guinea that could not have been influenced by our culture in any way. According to Ekman, the "basic emotions" are: ANGER, DISGUST, FEAR, HAPPINESS, SADNESS, and SURPRISE, as shown Table 2.1. This collection of emotions is one of the most popular sets between the categorical approaches, and thus one of the most used in CL (Aman, 2007; Vaassen, 2014; Ghazi, 2016).

Nevertheless, this is not the only categorical model to represent emotions. Numerous researchers worked on emotion definition and classification. It is true that many of them defined certain emotion as basic, however, there is not a consensus on which emotion must be considered "basic". For instance, Izard (1971) define ten categories as basic emotions: JOY, SADNESS, FEAR, ANGER, DISGUST, SURPRISE, INTEREST, SHAME, SHYNESS, GUILT. While Plutchik (1962, 1980, 1994) also believed that emotions evolved for the sake of human survival and reproduction as Ekman, he argues that there are eight basic bipolar emotions consisting in a superset of Ekman with two additions: TRUST and ANTICIPATION. Plutchik (1980) creates a wheel of emotions shown in Figure 2.1 with the aim of illustrating how emotions are related. The eight basic emotions are organized into four bipolar sets: JOY vs. SADNESS, ANGER vs. FEAR, TRUST vs. DISGUST and SURPRISE vs. ANTICIPATION. Additionally,

in this wheel, the intensity is also represented by the vertical dimension (higher intensity in the center of the wheel and lower in the periphery). With respect to the emotions with no color, they represent emotions that are a mix of 2 primary emotions. The fact that the emotions are represented by dimensions cause that Plutchick model can be considered a dimensional model from the psychological point of view. However, in CL, is widely used as a categorical model.

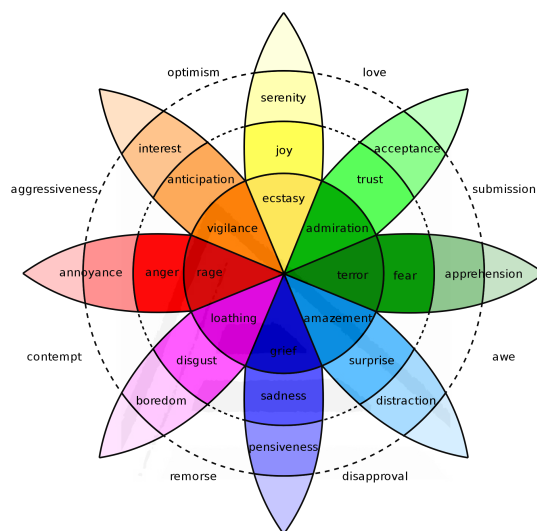


Figure 2.1: Plutchik's Wheel of Emotions.

Several works in this directions have been reported in the literature (Arnold, 1960; Tomkins, 1984; Ortony et al., 1988), as Table 2.1 shows. There is obviously no consensus on the number of basic emotions and not all the theorists agree on the existence of the "basic" emotions. However, all these researchers consider that there are discrete emotional categories. See Ortony and Turner (1990) for a detailed review of many of these models.

The fact that there is no agreement on the number of basic emotion is one of the drawbacks related to these models since there are problems of comparability of results across different studies in which widely different sets of emotion labels (Scherer, 2005). Moreover, a categorical model has the limitations of an identification task in attempting to identify the precise emotional states perceived by people due to the limited number of labels (Kim, 2011).

However, there are several benefits associated with the categorical representation. The great advantage of this framework is that it represents human emotions intuitively with easy to understand emotion labels (Kim, 2011; Vaassen, 2014). As mentioned above, the words have been used by philosophers to analyze the human emotional experience since the dawn of behavioral science. Thus, the use of these models is easier to understand for human’s annotators in a manual label task. Furthermore, textual emotion classification has been traditionally interpreted as a Text Categorization (TC) task, consequently the categorical model is easy to tackle from the computational point of view.

Table 2.1: Emotion Categories identified by researchers.

Researcher	Emotion Categories
Arnold (1960)	ANGER, FEAR, HATE, COURAGE, DEJECTION, DESIRE, DESPAIR, AVERSION, HOPE, LOVE, SADNESS
Tomkins (1984)	ANGER, INTEREST, CONTEMPT, DISGUST, DISTRESS, FEAR, JOY, SHAME, SURPRISE
Izard (1971)	ANGER, CONTEMPT, DISGUST, DISTRESS, FEAR, GUILT, INTEREST, JOY, SHAME, SURPRISE
Plutchik (1980)	JOY, SADNESS, ANGER, FEAR, TRUST, DISGUST, ANTICIPATION, SURPRISE
Ortony et al. (1988)	JOY, SADNESS, FEAR, ANGER, DISGUST, SURPRISE
Ekman (1992)	ANGER, DISGUST, FEAR, HAPPINESS, SADNESS, SURPRISE

2.1.2 Dimensional Emotion Models

The dimensional model represents affects in a dimensional form where each emotion occupies a location in this space.

The method of judging a document used by researchers adopting the dimensional approach is to ask a respondent how positive or negative and how excited or aroused he or she feels. There are two ways of carrying out this method: asking the value for each dimension in two separate steps or asking the respondent to determine the appropriate position on a two-dimensional

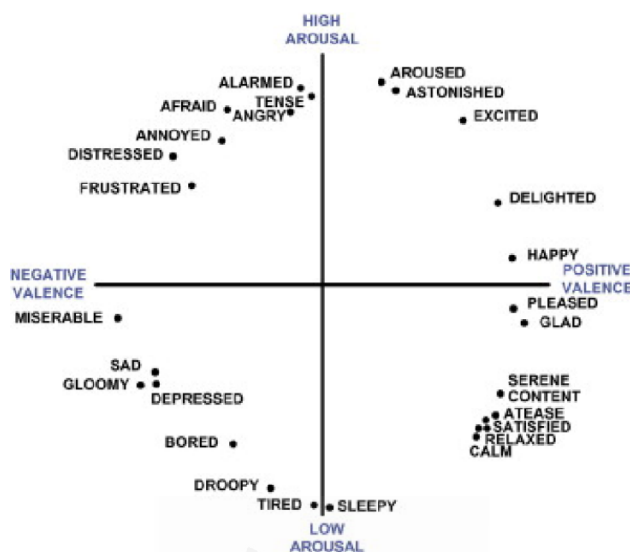


Figure 2.2: The Circumplex Model of Affect.

surface.

Wundt (1905) was one of the pioneers in describing emotion by dimensions with a three-dimensional space. Concretely, he suggested that emotions can be described by the dimensions of *valence* (POSITIVE-NEGATIVE), *arousal* (CALM-EXCITED), and *tension* (TENSE-RELAXED). Following Pleasure - Arousal - Dominance (PAD) representation, Mehrabian (1996) also proposes a three-dimensional model where the *dominance* dimension is used to distinguish whether the subject feels in control of the situation or not and the *pleasure* dimension corresponds to the *valence* of Wuldt model.

Due to the difficulty of identifying the *tension* dimension from *arousal* one, many current proposals are based on two dimensions (Scherer, 2005). One of the most representative two-dimensional models is Russel’s model of affect (Russell, 1980) as know as the Circumplex Model of Affect. Russell suggests a model where emotions are distributed in a two-dimensional circular space: *valence* and *arousal* dimensions, as Figure 2.2 shows. The *valence* dimension indicates how POSITIVE and NEGATIVE is an emotion whereas the *arousal* dimension differentiates EXCITED and CALM states. Thayer (1989) also propose a two dimensional model but with the *energy* and *stress* dimensions. In this space, emotions are distributed according to its value of *energy* and *stress* such as CONTENTMENT is located in low energy/low stress, DEPRESSION

in low energy/high stress, EXUBERANT in high energy/low stress, and ANXIOUS in high energy/high stress. Scherer (2005) propose the Geneva Emotion Wheel (GEW), a two-dimensional model based on the Circumplex Model of Affect (Figure 2.2) where the intensity dimension is mapped as the distance of an emotion category’s position. The number of emotion families is limited to 4 per quadrant, yielding a total of 16.

Another well-known dimensional model is Positive Activation - Negative Activation (PANA) model. It was proposed by Watson and Tellegen (1985) and suggests that positive affect and negative affect are two separate dimensions which they can be measured independently. The vertical axis represents low to high positive affect and the horizontal axis represents low to high negative affect.

Table 2.2: Emotion Dimensions identified by researchers.

Researcher	Emotion Dimensions
Wundt (1905)	<i>(valence, arousal, tension)</i>
Watson and Tellegen (1985)	<i>(positive affect, negative affect)</i>
Mehrabian (1996)	<i>(pleasure, arousal, dominance)</i>
Russell (1980)	<i>(valence, arousal)</i>
Thayer (1989)	<i>(energy, stress)</i>
Scherer (2005)	<i>(valence, arousal)</i>

As happen in categorical models, the dimensional ones present advantages and disadvantages. One of the benefits is that this method of obtaining emotion reports is simple and straightforward since it consists of asking contributors that classify the sentences or words with respect to two or three dimensions. Moreover, it also allows carrying out statistical processing since interval scaling can be used readily.

Nevertheless, the emotional information obtained from these reports are restricted to the degree of positive and negative feeling and the level of excitation in many cases. The two dimensions are not sufficient to individuate the whole spectrum of emotional concepts since there is very little information on the type of event that has produced the emotion (Scherer,

2005; Strapparava & Mihalcea, 2014). In addition, the possibilities of that two individuals who use the same emotion category to express feelings have more similar emotion is higher than those sharing a point in semantic space. As Scherer (2005) mentioned, this can be easily demonstrated by the fact that both FEARFUL and very ANGRY persons would be in a similar region of the two-dimensional space - negatively valenced high arousal. While verbal labels uniquely identify each emotion.

2.1.3 Categorical vs Dimensional

Two emotion models are presented: a categorical model where emotions are conceptualized as a set of distinct categories and a dimensional model representing emotional properties in terms of dimensions. Which of these two approaches is preferable? Both of them have benefits and drawbacks.

With the aim of taking advantage of the benefits of both approaches, there are several works which attempt the mapping between different emotion representation. For instance, Kim (2011) use the dimensional information to classify sentence into four discrete emotion categories (ANGER, FEAR, JOY, and SADNESS) or Hasan et al. (2014), who discretize the space associating a tag to each quadrant of the Circumplex Model of Affect. Most recently, Buechel and Hahn (2017) present a dimensional corpus and they carry out an experiment to automatically map between two emotion format, showing that their best models outperform human agreement in some of the emotions.

In spite of these initiatives, the majority of approaches to automatic emotion detection in text choose one of the models because of both of them are valuable to explore emotion data. However, as we will see in the next sections, the categorical approach is more popular in CL (Aman, 2007; Vaassen, 2014; Ghazi, 2016), especially in emotion recognition in text. Despite this, the election of choosing one, another one or both depends on our research objectives.

The categorical framework is the model selected in our research. Principally, for its intuitive and easily interpretable from human and computational point of view. Moreover, this model has a serious scientific history and a significant part of emotion classification research carried out so far adopts this approach since the emotion detection task is complex and emotion classes it makes the task easier to tackle.

2.2 Emotion Lexicons

After having presented how emotions are represented from a psychological point of view, our section presents emotion lexicons employed in CL for automatic emotion detection in text.

An emotion lexicon is a list of words labeled according to their emotional connotation. The label can be an emotional category (categorical emotion lexicons) or a value of the strength of a given emotion dimension (dimensional emotion lexicons). Its creation can be carried out in manual or semi/automatic way. When a lexicon is manually annotated, it usually is highly precise because the creation process is based on a human process. However, these resources have low coverage since it is hard to manually analyze the emotional connotation of a large set of words. For this reason, the semi/automatic processes are chosen when we need to improve the coverage.

Resources for emotion analysis can be grouped according to different criteria such as level of emotion annotation, language, domain or size. In this section, we decided to classify them depending on their creation process (manual or semi/automatic) with the aim of analyzing the different possibilities to build up an emotion resource. Moreover, within these groups, the emotion lexicons are grouped by its emotion connotation (categorical or dimensional) since these are the two main approaches to represent emotions that humans perceive and express.

2.2.1 Manual Creation Lexicons

Manual text annotation is the process of human annotators associate a text (word, sentence, phrase or document) with a tag or value. Concerning emotion annotation, the association is between a text and an emotion category or a strength value for an emotion dimension.

Manual annotation is a technique widely employed in emotion annotation since allowing encoding emotions from the human point of view. However, when the annotation task is carried out manually has also several drawbacks associated such as the time and cost required for its development or the low coverage of these lexica, as mentioned above.

One of the options to palliate these drawbacks is to annotate via crowdsourcing. The crowdsourcing platforms, Amazon Mechanical Turk ([AMT](#)) or Figure Eight ([F8](#)) (earlier called CrowdFlower ([CF](#))), allow accessing an online workplace to clean, label and enrich data, providing a significantly cheaper and faster method for collecting annotation from non-expert contributors over the Web. The usability and effectiveness of these platforms for emotion recognition has been tested by [Snow et al. \(2008\)](#), obtaining high agreement between the annotation of these platforms and the existing gold standard labels provided by expert labelers.

This section presents the most relevant emotion lexicons manually developed grouped by its emotion connotation (categorical or dimensional). The resources presented below are listed the oldest to the latest.

2.2.1.1 Categorical Emotion Lexicons

Linguistic Inquiry and Word Count (LIWC) dictionary ([Pennebaker et al., 2001](#)) is more a psycholinguistic resource than an emotion resource since it assigns one or more psychological categories to individual words. However, a set of these categories are related to emotions (`POSITIVE EMOTION`, `NEGATIVE EMOTION`, `ANGER`, `DISGUST`, and `ANXIETY`) and thus it is employed in many textual emotion detection systems. The LIWC2007 ([Pennebaker et al., 2007](#)) comprises almost 4,500 words and words stems, but in the most recent version, LIWC2015, the dictionary has been extended and new categories have been added resulting in a dictionary of almost 6,400 words, word stems, and selected emoticons that allowing to better tackle the new genres of the Web 2.0: social networks, blogs, tweets, etc.

Geneva Affect Label Coder (GALC) ([Scherer, 2005](#)) dictionary is an affective lexicon that enumerates for each emotion category the stemmed words that explicitly express the corresponding emotion. The result is a list of 279 stemmed terms associated manually with 36 emotion categories.

Categorical Affective Norms for English Words (CANEW) ([Stevenson et al., 2007](#)) is a characterization of Affective Norms for English Words ([ANEW](#)) by discrete emotional categories. It is a set of 1,034 words manually annotated with a set of cross-culturally universal basic emotions defined by [Levenson \(2003\)](#): `HAPPINESS`, `SADNESS`, `FEAR`, `DISGUST`, and `ANGER`. Each [ANEW](#) word was rated by five independent annotators on the five discrete

emotions on a scale of 1 (low value) - 5 (high value).

Affect database (Neviarouskaya et al., 2011) includes emoticons, abbreviation, adjectives, adverbs, nouns, verbs, injections and modifiers (nearly 3,000 terms) manually labeled with emotion intensity and nine emotion categories: ANGER, DISGUST, FEAR, GUILT, INTEREST, JOY, SADNESS, SHAME and SURPRISE. Emotion intensities describe the intensity of degree of affective states from 0.0 (very weak) to 1.0 (very strong).

NRC Emotion Lexicon (EmoLex) (Mohammad & Turney, 2013) is a dataset of general domain consisting of 14,000 English words associated with the Plutchik's eight basic emotions: ANGER, FEAR, ANTICIPATION, TRUST, SURPRISE, SADNESS, JOY, DISGUST and two sentiments (NEGATIVE and POSITIVE). Emolex was manually collected using AMT platform and each word was labeled by five annotators.

2.2.1.2 Dimensional Emotion Lexicons

Dictionary of Affect in Language (DAL) (Whissel, 1989) consists of 8,742 words manually rated on a three-point scale by 200 people along three dimensions: ACTIVATION (active-in between-passive), EVALUATION (pleasant-in between-unpleasant) and IMAGERY (easy to imagine-in between-hard to imagine).

Affective Norms for English Words (ANEW) (Bradley & Lang, 1999a) is a set 1,034 English words including verbs, nouns, and adjectives with emotional ratings. It is a lexicon based on the Osgood model (dimensional emotion model) where subjects have manually rated the words from 1 (low value) to 9 (high value) in terms of the three dimensions of PLEASURE, AROUSAL, and DOMINANCE using Self-Assessment Manikin (SAM) (Bradley & Lang, 1994). The objective of this resource was to complement the existing International Affective Picture System (IAPS) (Lang et al., 1999) and International Affective Digitized Sounds (IADS) (Bradley & Lang, 1999b), which are collections of picture and sound stimuli with affective rating, respectively.

Extended Affective Norms for English Words (E-ANEW) (Warriner et al., 2013) is a dataset of nearly 14,000 English lemmas manually rated in terms of three dimensions: PLEASURE, AROUSAL, and DOMINANCE like ANEW.

The words included in this resource were compiled from three sources: ANEW database (Bradley & Lang, 1999a), (Van Overschelde et al., 2004) category norms, and the SUBTLEX-US corpus (Brysbaert & New, 2009). To carry out the manual annotation task, AMT platform was employed.

2.2.2 Semi-Automatic/Automatic Creation Lexicons

As we see in the previous section, the manual annotation has drawbacks associated and one way to improve the task is the use of crowdsourcing platforms. However, there are more alternative ways to tackle emotion annotation in an effective way such as automatic or semi-automatic methods.

This section presents the most relevant emotion lexicons created by semi/automatic process classified depending on its emotion connotation (categorical or dimensional). The resources described below are listed the oldest to the latest.

2.2.2.1 Categorical Emotion Lexicons

WordNet-Affect (WNA) (Strapparava & Valitutti, 2004) is an extension of WordNet Domains (Magnini & Cavaglia, 2000), a multilingual extension of WN (Miller, 1995), that includes a subset of 2,874 synsets and 4,787 words suitable to represent affective concepts correlated with affective words. The affective concepts representing emotional states are individuated by synsets marked with the a-label emotion (JOY, LOVE, SADNESS, SURPRISE, APATHY,...). There are also other a-labels for those concepts representing moods, situations eliciting emotions or emotional responses. WordNet-Affect was developed in two stages: 1) the development of the core where a lexical database of affective words was manually realized (named AFFECT), and 2) the extension of affective core exploiting the WN relations with a semi-automatic process where each relation is examined checking if it preserves the effective meaning (the relation generates synsets that still represent affective concepts) or, on the contrary, the relation generates synsets not included in the affective core which are manually checked.

SentiSense (de Albornoz et al., 2012) is a concept-based affective lexicon that attaches emotional meanings to concepts from the WordNet (Miller, 1995) lexical database, instead of terms. It consists of 5,496 words and 2,190

synsets labeled with a set of 14 emotion categories result from combining three categorical emotion models: Arnold (1960), Plutchik (1980), and Parrott (2001). As WNA, this resource was created with a semi-automatic process consisting of two steps: 1) a seed of synsets was manually annotated by two annotators, and 2) these synsets were automatically expanded using several relations in WN and checking if the relation generates synsets that preserve the same emotional meaning. They concluded that only *derived-from-adjective*, *pertains-to-noun*, and *participle-of-verb* relations typically maintain the emotional meaning.

EmoSenticNet (ESN) (Poria et al., 2013) is a lexical resource of 13,189 words that assigns qualitative emotions label and quantitative polarity scores to SenticNet concepts (Cambria et al., 2014). It was automatically built by the assign WNA emotion labels to SenticNet’s concepts using a supervised machine-learning approach. The intuition behind this approach is that words that have similar features are similar in their use and meaning and thus, are related to similar emotions. Ekman’s emotions: ANGER, FEAR, DISGUST, SADNESS, SURPRISE, or JOY) is the set of emotions employed for labelling the concepts.

Synsketch Word Lexicon (Krcadinac et al., 2013) is a dataset consisting of 3,725 words annotated associated with the Ekman’s basic emotions: HAPPINESS, SADNESS, ANGER, FEAR, DISGUST, and SURPRISE. It was built by a semi-automatic process where first they conducted a 20-person study where they asked to list at least five words for each emotion and those words mentioned three or more time were considered good indicators. In the second step, each word was associated with the WN synsets and the lexicon is created by analyzing the semantic relationships of word and synsets.

DepecheMood (Staiano & Guerini, 2014) is a lexicon of 37,000 terms annotated with emotions scores of eight emotion categories: AFRAID, AMUSED, ANGRY, ANNOYED, DONT_CARE, HAPPY, INSPIRED, and SAD. It is built in an automatic way exploiting the affective annotation implicitly provided by readers of news articles from `rappler.com`. This web-page contains news articles where offers the readers the opportunity to click on the emotion that a given Rappler story made them feel (though the Rappler’s *Mood Meter*). Then, Staiano and Guerini (2014) download the documents along with the mood information and they built a word-by-emotion matrix using

an approach based on compositional semantics.

NRC Hashtag Emotion Lexicon (Mohammad & Kiritchenko, 2015) is a dataset of 16,000 unigrams built by automatically from tweets with emotion-word hashtags posted in 2012 (Twitter Emotion Corpus (TEC)). To achieve that they create n -gram-emotion association lexicons from emotion-labeled sentences in the TEC. Concretely, they compute the *strength of association* (SoA) between an n -gram w and an emotion e to be as follows:

$$SoA(w, e) = PMI(w, e) - PMI(w, \neg e) \quad (2.1)$$

where PMI is the pointwise mutual information. If an n -gram has a stronger tendency to occur in a sentence with a particular emotion label than in a sentence that does not have that label, then that n -gram-emotion pair will have a SoA score that is greater than zero. Hence, this resource provides real-values scores that indicate the degree of word-emotion association (high scores imply higher association) for the Plutchik's eight basic emotions: ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE, TRUST, and ANTICIPATION.

2.2.2.2 Dimensional Emotion Lexicons

The number of dimensional emotion lexicons used in CL is really low. This might be the reason why all of them have been developed manually (Section 2.2.1.2) and there is not dimensional emotion lexicon developed with a semi/automatic methodology. However, the tendency of using semi/automatic techniques for improving emotion resources could produce the emergence of dimensional emotion lexicons in the near future.

2.2.3 Discussion

This section presents the most relevant emotion lexicons used in automatic emotion detection in text. The resources have been grouped by their creation process and moreover, they have been classified by their emotion connotation (categorical or dimensional).

With respect to the classification by emotion connotation, it is possible to observe, as previously mentioned in Section 2.1, the popularity of categorical models in CL since the number of resources annotated with emotion categories is significantly higher than dimensional lexicons.

Table 2.3: Emotion Lexicons

Lexicon	Size	Emotion Model	Emotion Categories / Dimensions	Creation Process
LIWC (Pennebaker et al., 2001)	6,400 words	Categorical	POSITIVE EMOTION, NEGATIVE EMOTION, ANGER, DISGUST, ANXIETY	Manual
GALC (Scherer, 2005)	279 stemmed words	Categorical	36 categories GEW (Scherer, 2005)	Manual
CANEW (Stevenson et al., 2007)	1,034 words	Categorical	Levenson (2003)'s emotions	Manual
Affect database (Neviarouskaya et al., 2011)	3,000 words, emoticons and abbreviation	Categorical	ANGER, DISGUST, FEAR, GUILT, INTEREST, JOY, SADNESS, SHAME, SURPRISE	Manual
EmoLex (Mohammad & Turney, 2013)	14,000 words	Categorical	Plutchik (1980)'s emotions	Manual
DAL (Whissel, 1989)	8,742 words	Dimensional	Activation, Evaluation, Imagery	Manual
ANEW (Bradley & Lang, 1999a)	8,742 words	Dimensional	Activation, Evaluation, Imagery	Manual
E-ANEW (Warriner et al., 2013)	14,000 words	Dimensional	Pleasure, Arousal, Dominance	Manual
WNA (Strapparava & Valitutti, 2004)	4,787 words	Categorical	JOY, LOVE, SADNESS, SURPRISE, APATHY,...	Semi-Automatic
SentiSense (de Albornoz et al., 2012)	5,496 words	Categorical	Combination of Arnold (1960); Plutchik (1980); Parrott (2001)'s emotions	Semi-Automatic
ESN (Poria et al., 2013)	13,189 words	Categorical	Ekman (1992)'s emotions	Automatic
Synsketch Word Lexicon (Krcadinac et al., 2013)	3,725 words	Categorical	Ekman (1992)'s emotions	Semi-Automatic
DepecheMood (Staiano & Guerini, 2014)	37,000 words	Categorical	AFRAID, AMUSED, ANGRY, ANNOYED, DONT_CARE, HAPPY, INSPIRED, SAD	Automatic
NRC Hashtag Emotion Lexicon (Mohammad & Kiritchenko, 2015)	16,000 words	Categorical	Plutchik (1980)'s emotions	Automatic

The classification by its creation process allows showing the variety of methodologies to build up emotion lexicons, from manual processes, going

through crowdsourcing platforms to semi/automatic methodologies.

Regarding manual creation lexicons, its coverage, as mentioned in the introductory of this section, are low since the size of these lexicons vary from 279 words in [GALC](#) to 8,742 words [DAL](#). However, the resources annotated via crowdsourcing overcome the 10,000 words, reaching 14,000 words in [EmoLex](#) and [E-ANEW](#). Thus, it could be noted the effectiveness of crowdsourcing platforms to improve the manual text annotation.

Despite the benefits of crowdsourcing platforms, the analysis of the emotion lexicons created by semi/automatic processes allows us to notice that the majority of resources recently developed has opted for these methodologies. This is because the efficiency of semi/automatic methods in terms of time and cost is higher than manual processes or with crowdsourcing platforms. Consequently, the semi/automatic methodologies are becoming an interesting option to avoid, or at least considerably reducing, the need for manual annotations.

2.3 Emotion Corpora

After having reviewed emotion lexicons, our next section presents emotion corpora for automatic emotion detection in text.

An emotion corpus is a large and structured set of sentences where each sentence is tagged with one or more emotional tags. Corpora are a fundamental part of supervised learning approaches, one of the approaches most applied for automatic detection of emotion. Supervised learning algorithms first infer a function from a set of examples labeled with the correct sentiment (training data or labelled corpus). After this, the model is able to predict the emotion of new examples. Hence, the training dataset employed in supervised machine learning algorithms is crucial to build accurate emotion detection systems that can generate reliable results.

The creation of a labelled corpus for [ER](#) is not trivial, since detecting emotion in text can be difficult even for humans, because everyone's personal context can influence emotion interpretation. Most relevant research carried out so far has shown difficulties related to this task, such as obtaining a good Inter-Annotator Agreement ([IAA](#)) or the time and the cost required for its development. As a consequence, data gathering with emotion content has

become one of the most challenge tasks in [ER](#).

Emotion annotation task has been majority approached with a manual process since, in this way, machine learning systems learn from human annotations that are generally more accurate. However, over the past few years, the exponential growth in the amount of subjective information and the appearance of new textual genres on the Web 2.0 such as blogs, microblogging or social networks, there has been an increasing interest from researchers to develop new methods to label the emotion content of the information available in these new sources. Accordingly, several emotion resources have recently been developed applying semi/automatic methodologies with the aim of overcoming the cost and time-consuming shortcoming of manual annotation.

Consequently, the first criteria by which we present the emotion corpora is its creation process: manual annotation (Section [2.3.1](#)) and semi/automatic annotation (Section [2.3.2](#)) with the aim of analyzing the emotion corpora from this point of view. In each subsection, the emotion corpora have been classified by a second criterion according to their emotion connotation, means the emotion model employed for its annotation: the categorical model or the dimensional one.

2.3.1 Manual Corpora Annotation

Manual corpus annotation is widely used in Natural Language Processing ([NLP](#)) and consists of associating labels to words, phrases, sentences, paragraphs or documents depending on the level of annotation. Specifically, in [ER](#), a label could be a discrete category if the categorical emotion model is employed or a strength value for each dimension when the emotion model applied is based on a dimensional model. Respect to the level of annotation, the majority of corpora are annotated at sentence level since document level assume that each document expresses a single sentiment ([Ghazi, 2016](#)).

As previously mentioned, emotion annotation has been generally tackled by manual annotation due to mainly the fact automatic system learn from human annotations. Nevertheless, the annotation scheme, the difficulty of the task or the training of the annotators are factors that turn the task into a difficult assignment. These aspects are even more complex to define in emotion manual annotation task because of its highly subjective.

This section presents the most relevant emotion corpora developed manually grouped by their emotion connotation: the categorical model and the dimensional one. The resources presented below are listed from the oldest to the latest.

2.3.1.1 Categorical Emotion Corpora

The **International Survey on Emotion Antecedents and Reactions (ISEAR)** corpus (Wallbott & Scherer, 1986) is one of the first manual corpus annotated with emotion categories. It is a dataset of 3000 reports (7,667 sentences) in 37 countries on all 5 continents collected over a period of many years during the 1990s. The creation process consisted in asking to students, psychologists and non-psychologists, about situations where they had experienced these seven emotions: JOY, FEAR, ANGER, SADNESS, DISGUST, SHAME, and GUILT. The questions covered the way they had appraised the situation and how they reacted.

Alm corpus (Alm et al., 2005) was developed with the aim of classifying the emotional affinity of sentences in the narrative domain of children's fairy tales. To achieve that, they annotated a corpus of approximately 185 children stories manually annotated where each annotator marked the sentence level with one of eight set of basic emotions: ANGRY, DISGUSTED, FEARFUL, HAPPY, SAD, POSITIVELY SURPRISED, NEGATIVELY SURPRISED and NEUTRAL. In order to make the annotation process more focused, emotion is annotated from the point of view of the text and this is considered the primary emotion, but the sentences are also marked for other affective contents like background mood or the secondary emotions via intensity, feeler, and textual cues. This work is one of the first that shows the complexity of emotion annotation since the IAA obtained in the annotation evaluation ranged from $k = 0.24 - 0.51$, showing fair-moderate agreement according to Landis and Koch (1977).

The **Semeval 2007 task 14 - Affective text** (Strapparava & Mihalcea, 2007) was focused on emotion classification. The task participants had at their disposal a data set consisted of 1,250 manually annotated headlines drawn from major newspapers such as New York Times, CNN, BBC News and the Google News search engine manually annotated for six basic emotions: ANGER, DISGUST, FEAR, JOY, SADNESS and SURPRISE, and for positive and

negative polarity. Unlike previous annotations of automatic recognition of emotion in text (Alm et al., 2005; Aman & Szpakowicz, 2007), they decided to use a finer-grained scale, where each annotator could provide scores between 0 and 100 for each emotion, hence allowing the annotator to select different degrees of emotion load. The Pearson (1956) correlation measure was employed for agreement evaluation, obtaining score ranged from 36.07 and 68.19 for emotions and 78.01 for valence.

Aman corpus (Aman & Szpakowicz, 2007, 2008) is another representative resource in this field. The data presented (4,080 sentences) came from blog posts directly collected from the Web and they are manually annotated with emotion category, emotion intensity and the words/phrases that indicate emotion in text at sentence level. Concerning emotion categories, each sentence was annotated with the Ekman’s six emotions and two more categories: MIXED EMOTION for those sentences that exhibit more than one emotion and NO EMOTION for those sentences that could not be attributed to any basic category. About emotion intensity, all emotion sentences in the corpus, irrespective the emotion category associated (except NO EMOTION category) were assigned emotion intensity: HIGH, MEDIUM, or LOW. Finally, they also annotate emotion indicators, this is spans of text that convey emotional content in a sentence. The average IAA (Cohen (1960)’s kappa) on emotion categories was in the range $k = 0.6 - 0.79$, for emotion intensity ranged from $k = 0.37 - 0.72$, and for emotion indicators 0.66.

Gill et al. (2008) corpus. They presents a study to examine the ability of naive annotators of emotion vs expert raters. To achieve that, a ‘gold standard’ emotion corpus is manually annotated by five human raters who had extensive experience in the text used in the task. Gill corpus consists of 135 texts (the first 200 words of each post) extracted from blog texts previously collected (Nowson et al., 2005). Each annotator rated these text as expressing one of these eight emotions: ANTICIPATION, ACCEPTANCE, SADNESS, DISGUST, ANGER, FEAR, SURPRISE, JOY and NEUTRAL represented by the activation-evaluation wheel (Russell, 1980; Plutchik, 1994). After that, 20 of these 135 texts was annotated by naive raters to examine the differences. The reliability of the corpus was assessed through measuring the agreement between naive and experts, obtaining the range $A_o = 0.1 - 0.8$ for emotion categories in short texts.

The **Neviarouskaya blogs** (Neviarouskaya et al., 2011) is 700 sentences of diary blog posts provided by BuzzMetrics¹. The sentences were manually annotated by three independent annotations with one of nine emotions categories (ANGER, DISGUST, FEAR, GUILT, INTEREST, JOY, SADNESS, SHAME, and SURPRISE) defined by Izard (1971) or NEUTRAL and a corresponding intensity value. Fleiss (1971) ' Kappa coefficient was the metric to evaluate the reliability of the annotators and the level of agreement of 700 sentences was moderate (0.47 in emotion categories and 0.59 in case of polarity annotations).

The **Neviarouskaya stories** (Neviarouskaya et al., 2010) is a corpus composed of 1,000 sentences extracted from personal stories grouped by topics with 13 different categories, such as "Arts and entertainment", "Education", "Health and wellness" and others. The personal stories were obtained from the social networking website Experience Project², where people anonymously published their life experiences. They considered three hierarchical level of attitude labels in the experiments, being ALL level the level that contains emotion categories. Three independent annotations manually labeled the sentences with one of 14 categories from ALL level (INTEREST, JOY, SURPRISE, POSITIVE JUDGMENT, POSITIVE APPRECIATION, ANGER, DISGUST, FEAR, GUILT, SADNESS, SHAME, NEGATIVE JUDGMENT, POSITIVE JUDGMENT or NEUTRAL) and a corresponding intensity value. As in Neviarouskaya blogs corpus, Fleiss' Kappa coefficient was used as a measure of agreement. The reliability of annotators on 1,000 sentences was 0.53 on ALL level.

EmotiBlog corpus (Boldrini & Martínez-Barco, 2012) consists of a collection of blog posts about three subjects of interest: the Kyoto Protocol, the election in Zimbabwe, and the 2008 USA presidential elections. For each topic, 100 texts were collected and manually annotated with three annotation levels: document, sentence, and element. The emotion categories employed are an extended set of Scherer's classification (Scherer, 2005): CRITICISM, HAPPINESS, SUPPORT, IMPORTANCE, GRATITUDE, GUILT, FEAR, SURPRISE, ANGER, ENVY, INDIFFERENCE, PITY, PAIN, SHYNESS and BAD. For agreement evaluation, due to the annotation granularity, they use the following pairwise agreement:

$$agr(a||b) = \frac{|A \text{ matching } B|}{|A|} \quad (2.2)$$

¹<http://buzzmetrics.com/>

²<http://www.experienceproject.com/>

where A and B are two annotators. They obtain 73,6% of IAA for Spanish corpus.

Suicide notes corpus (Pestian et al., 2012) contains suicides notes written by 1,319 people, collected between the years of 1950 and 2011. The corpus was used in the shared task on emotion classification in suicide notes organized from the 2011 i2b2 NLP Challenge. Each note was manually annotated at least three times and annotators were asked to identify the following emotions: ABUSE, ANGER, BLAME, FEAR, GUILT, HOMELESSNESS, SORROW, FORGIVENESS, HAPPINESS, PEACEFULNESS, HOPEFULNESS, LOVE, PRIDE, THANKFULNESS, and two more categories related to the content of the note: INSTRUCTIONS, and INFORMATION. The annotations were collect at token and sentence level, but the gold standard was created at the sentence level. As metric to evaluate the reliability of the annotators, they used Krippendorff (1980)' alpha, obtaining a 0.546 IAA.

EmpaTweet corpus (Roberts et al., 2012) is a collection of 7,000 tweets manually annotated at the tweet-level with Ekman's basic emotions (ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE) and LOVE since they considered that this category would be commonly found in informal texts. The annotation was carried out into three phases: (i) initial teaching phase; (ii) independent annotation phase where 1500 tweets were double-annotated, obtaining a $k = 0.67$ IAA when the disagreement was resolved; (iii) and the vast of the annotation was done individually to maximize the number of annotations.

The **music and lyrics corpus** developed by Mihalcea and Strapparava (2012) is a novel corpus consisting of 100 songs annotated for emotions. They compiled 100 popular songs due to the message and emotion they convey. To annotate the emotions in songs they use the six basic emotions proposed by Ekman: ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE. The annotations are collected at line level, from the writer perspective and with a separate annotation for each of the six emotions. To carry out the manual annotation task, they used the AMT³ service and each song was labeled between two-five annotators. The metric to evaluate the reliability of human's annotators was Pearson (1956)'s correlation, obtaining a 0.73 as an overall correlation between the remaining reliable annotators.

³www.mturk.com

The **Balabantaray corpus** (Balabantaray et al., 2012) consists in a collection of 8,150 tweets manually annotated with the Ekman’s basic emotions: ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE. Five annotators participated in the task and each tweet was subject to two judgments. The annotators did not receive training, though they were given samples of annotated sentences to illustrate the kind of annotations required. About the evaluation of the reliability of the humans’ annotations, Cohen’s kappa was the metric chosen and the average IAA on emotion categories was in the range $k = 0.59 - 0.75$.

Sport-related Emotion Corpus (SREC) (Sintsova et al., 2013) is a collection of 1,987 tweets about the gymnastic competitors of the 2012 Olympic games. The manual annotation task was carried out on AMT and each annotator was asked to read a tweet and associated an emotion label, the emotion strength, and related emotion indicators. They employ the emotion categories from GEW (Scherer, 2005) which present 20 (10 positive/10 negative) emotion categories arranged on the circle following the underlying 2-dimensional space of valence (positive-negative) and control (high-low). Thus, the approach is based on dimensional emotion model. Despite this, the corpus is annotated with emotion categories and not with the values of each dimension, hence, from our point of view, SREC corpus is considered a categorical emotion corpus. The measure of agreement for emotion labels employed is Fleiss Kappa with a 0.24 value which is considered to be fair by Landis and Koch (1977).

EmoTweet-28 corpus (Liew et al., 2016) is a recent collection of 15,553 tweets manually annotated with 28 emotion categories. These categories were identified in a previous phase of annotation task where the annotators were asked to create their own emotion labels to the emotion(s) expressed in the tweets. After refine the set of emotion tags that emerged from data and resolve the disagreement, the corpus was annotated with 28 emotion categories. Moreover, the dataset contains annotations for three more facets of emotion: valence, arousal, and emotion cues. As the previous work, the AMT platform is employed to carry out the manual annotation task where each tweet was annotated by at least three annotators. Krippendorff’s alpha and Fleiss’ Kappa are the metrics of agreement employed in this work and both of them obtain 0.43 value of the IAA for emotion categories.

Table 2.4: Categorical Emotion Corpora (manual annotation)

Corpus	Source	Size	Emotion Categories	IAA
ISEAR (Wallbott & Scherer, 1986)	Reports	7,667 sentences	7 categories	-
Alm et al. (2005) corpus	Children's Tales	185 stories	Ekman (1992)'s emotions	$k = 0.24 - 0.51$
Affective text corpus (Strapparava & Mihalcea, 2007)	News Headlines	1,250 headlines	Ekman (1992)'s emotions	$r = 0.36 - 0.68$
Aman and Szpakowicz (2007, 2008) corpus	Blog Posts	4,080 sentences	Ekman (1992)'s emotions	$k = 0.6 - 0.79$
Gill et al. (2008) corpus	Blog Posts	135 texts	Plutchik (1994)'s emotions	$A_o = 0.1 - 0.8$
Neviarouskaya et al. (2011) Blogs corpus	Blog Posts	700 sentences	Izard (1971)'s emotions	$k = 0.47$
Neviarouskaya et al. (2010) Stories corpus	Personal Stories	1,000 sentences	14 categories (included Izard (1971)' emotions)	$k = 0.53$
EmotiBlog corpus (Boldrini & Martínez-Barco, 2012)	Blog Posts	100 texts	15 categories (included Scherer (2005)' emotions)	$agr(a b) = 0.74$ Equation 2.2
Suicide Notes Corpus (Pestian et al., 2012)	Suicide Notes	1,319 notes	16 categories	$\alpha = 0.55$
EmpaTweet corpus (Roberts et al., 2012)	Twitter Messages	7,000 tweets	Ekman (1992)'s emotions	$k = 0.67$
Music and Lyrics corpus Mihalcea and Strapparava (2012)	Songs	100 songs	Ekman (1992)'s emotions	$r = 0.73$
Balabantaray corpus (Balabantaray et al., 2012)	Twitter Messages	8,150 tweets	Ekman (1992)'s emotions	$k = 0.59 - 0.75$
SREC (Sintsova et al., 2013)	Twitter Messages	1,987 tweets	20 categories (GEW) (Scherer, 2005)	$k = 0.24$
EmoTweet-28 corpus (Liew et al., 2016)	Twitter Messages	15,553 tweets	28 categories	$k = 0.43$ and $\alpha = 0.43$

2.3.1.2 Dimensional Emotion Corpora

Affective Norms for English Text (ANET) (Bradley & Lang, 2007) is a collection of 120 English sentences designed for psychological research rated manually for three dimensions: *pleasure, arousal, dominance*.

Affective Norms for Polish Short Text (ANPST) (Imbir, 2016) is a dataset of 718 emotive sentences collected from literature quotations, movies, newspapers, television, programs, humorous stories, and web pages. Each sentence was rated manually with six dimensions: *valence, arousal, dominance, origin, significance, and source*. For *valence, arousal, and dominance*, the Self-Assessment Manikin (SAM) scale is used. While the remaining three dimensions were created specifically for research concerning the emotion-duality model (Imbir, 2015).

Facebook posts corpus (Preotjiuc-Pietro et al., 2016) is a collection of 2,895 anonymized Facebook post shared by participants as part of the MyPersonality Facebook application (Kosinski et al., 2013). The corpus was manually labeled by two annotators with psychology training. The emotion model employed was the Circumplex Model of Affect (Russell, 1980) and thus, the ratings are made on two independent nine-point scales, where the scales represent valence (sentiment) and arousal (intensity). Regarding the reliability of the annotation, Pearson (1956)'s correlation coefficient was the metric used, obtaining agreement correlation of 0.767 for valence and 0.827 for arousal.

Chine Valence-Arousal Text (CVAT) (Yu et al., 2016) is a corpus consisting of 2,009 sentences extracted from web text from different categories: news articles, political discussion forums, car discussion forums, hotel reviews, books reviews, and laptop reviews. A crowdsourcing annotation platform was employed for its manual annotation using the SAM annotation scheme where the annotations were asked to rate individual sentences from 1 to 9 in terms of valence and arousal (the Circumplex Model of Affect). Each sentence was annotated by at least 10 annotators. The reliability of human's annotations was measure with Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), obtaining agreement correlation values 0.56 (MAE) - 0.72 (RMSE) for valence and 1.02 (MAE) - 1.27 (RMSE) for arousal.

EMOBANK (Buechel & Hahn, 2017) is the most recent corpus developed based on dimensional emotion model. It is a corpus of 10,548 English sentences built from other existing resources: the Manually Annotated Sub-Corpus of the American National Corpus (**MASC**) and the corpus of SemEval-2007 Task 14 Affective Text. Thus, the corpus contains sentences from different genres: news headlines, blogs, essays, fiction, letters, newspapers and travel guides. These sentences were manually annotated with a dimensional model in the Valence-Arousal-Dominance (**VAD**) representation format by five annotators using **CF** platform. The most relevant feature of EMOBANK is that has a bi-perspectival (each sentence was rated according to both the perceived writer and the perceived reader emotion) and bi-representational design (part of the corpus was also annotated according to a categorical emotion model). Pearson’s correlation coefficient and **MAE** are the metrics employed as agreement measures, obtaining $r = 0.605$ and **MAE** = 0.335 for writer perspective and $r = 0.634$ and **MAE** = 0.386 for reader aspect.

Table 2.5: Dimensional Emotion Corpora (manual annotation)

Corpus	Source	Size	Emotion Dimensions	IAA
ANET (Bradley & Lang, 2007)	Psychological texts	120 sentences	<i>pleasure, arousal, dominance</i>	-
ANPST (Imbir, 2016)	Newspapers, movies, television, programs, etc.	718 sentences	<i>pleasure, arousal, dominance, origin, significance, source</i>	-
Facebook posts corpus (Preotjuc-Pietro et al., 2016)	Facebook Messages	2,895 posts	<i>valence, arousal</i> Circumplex Model of Affect (Russell, 1980)	$r = 0.757$ (valence) - $r = 0.872$ (arousal)
CVAT (Yu et al., 2016)	News articles, hotel reviews, forums, etc.	2,009 sentences	<i>valence, arousal</i> Circumplex Model of Affect (Russell, 1980)	0.56 MAE - 0.72 RMSE (valence) and 1.02 MAE - 1.27 RMSE (arousal)
EMOBANK (Buechel & Hahn, 2017)	News headlines, blogs, essays, etc..	10,548 sentences	<i>valence, arousal, dominance</i>	$r = 0.605$ and 0.335 MAE (write perspective) - $r = 0.634$ and 0.366 MAE (reader perspective)

2.3.2 Semi-Automatic/Automatic Corpora Annotation

As mentioned in the introductory section, there is a tendency of improving the emotion annotation process since the increase of the amount of data to analyze and the emergence of new genres implies the need of researching alternative techniques to build emotion resources more efficiently in term of cost and time. Consequently, it is possible to find, in the state-of-art, new approaches to tackle the emotion annotation task such as tools or methodologies to assist and guide the work of annotators, semi-automatic annotation methods where the most trivial and repetitive work is done automatically and the hardest annotation task is carried out for humans annotation or automatic annotation processes without human intervention.

While it is true that there are many semi-automatic and automatic methodologies to carry out emotion annotation, most of the approaches are focused on *distant supervision*. The intuition of *distant supervision* is that any piece of text (sentence, phrase or paragraph) that contains an emotion element is likely to express this emotion in some way (Mintz et al., 2009). This technique is widely applied in emotion annotation task since the launching of Twitter microblogging service⁴. Twitter is a social network service where users can post and interact with other users through messages called *tweets*. Moreover, the platform allows using *hashtags*, a word or phrase preceded by the character '#' to identify messages on a specific topic. Many works employ these hashtags or other kind of tags to apply *distant supervision*, as we will see in this section. This technique is also known as "emotion word-hashtag" (Mohammad, 2012a) or "harnessing" (Wang et al., 2012) in emotion annotation.

The great advantage of *distant supervision* is that manual annotation is not required because these labels are provided by the writers of the messages themselves. Thus, it is possible to obtain a large amount of the labeled data in a really efficient way in term of cost and time. However, the major drawback of this methodology is the data labeled contains noise (Read, 2005; M. Li et al., 2016). If there is a lot of noisy data, the resultant emotion corpus is less useful as it affects the performance of the automatic emotion system. Moreover, another limitation is that this technique can be applied only in those genres which the use of hashtags or emoticons is widespread.

⁴<https://twitter.com/>

This section presents the emotion corpora developed by semi/automatic techniques. As the previous section, the criteria of grouping is their emotion connotation: the categorical model and the dimension one and within these two groups, the resources are listed from oldest to latest.

2.3.2.1 Categorical Emotion Corpora

Read (2005) corpus is one of the first created by *distant supervision* in emotion annotation. To achieve that, they collected a corpus of 13,000 articles labeled with emoticons by downloading Usenet newsgroups. They extracted automatically the paragraph(s) containing the emoticon of interest. Each emoticon was associated with an emotion or a description of the emoticon and thus their categories were: SMILE :-), WINK ;-), FROWN :-), WIDE GRIN :-D, TONGUE STICKING OUT ;-P, SURPRISE :-O, DISAPPOINTED :-|, CRYING :'(, CONFUSED :-S, ANGRY :-@ and EMBARRASSED :-\$.

Mishne (2005) corpus is a collection of blog posts from LiveJournal⁵ annotated with mood categories. They used mood categories because LiveJournal is a free weblog service that includes an optional field indicating the "current mood". In this way, the blog posts are labeled with the writer's mood. To apply *distant supervision*, they used these mood categories and obtaining a corpus of 815,494 blog posts annotated with 132 common moods such as AMUSED, TIRED, HAPPY, BORED, CALM, SAD, etc.

Purver and Battersby (2012) corpus consists of tweets annotated with the Ekman's basic emotions. For building up the corpus, they collect Twitter messages marked with the author's own intended interpretation through emoticons or hashtags (*distant supervision*) corresponding to one of Ekman's emotion classes: ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE. For emoticons, Yahoo messenger classification was employed and for hashtags, they used emotion names themselves plus the main related adjective (#happy, #sad, #angry, #scared, #surprised, #disgusted). With respect to the size of the corpus, for emoticons they obtained a maximum of 837,849 (for happy) and a minimum of 10,539 for anger; for hashtags, a maximum of 10,219 for happy and a minimum of 536 for disgust. The total size of the corpus is not provided.

⁵<https://www.livejournal.com/>

Mood and Affects corpus. Choudhury et al. (2012) collected Twitter messages where each post contained one of 172 mood words defined previously as a hashtag at the end. The process to select the 172 mood words (hashtags) consisted of two main phases: 1) ensembling the words of different emotion lexicons (ANEW, LIWC, Emotion Annotation and Representation Language (EARL), a list of "basic emotions" provided by Ortony and Turner (1990) and a list of mood provided by the blogging website LiveJournal; and 2) performing a study on AMT to narrow they candidate set to truly mood-indicative words. Finally, they associated each mood with the affects defined by Positive and Negative Affect Schedule-Expanded (PANAS-X) (Watson & Clark, 1994). This resource defined 11 specific affects: FEAR, SADNESS, GUILT, HOSTILITY, JOVIALITY, SELF-ASSURANCE, ATTENTIVENESS, SHYNESS, FATIGUE, SURPRISE, and SERENITY. The corpus contains 6.8 million of tweets annotated with this 11 affects using emotion-word hashtags.

Twitter Emotion Corpus (TEC) or Hashtag Emotion Corpus (HEC) (Mohammad, 2012a) is a collection of 21,000 tweets from 19,000 different people annotated with Ekman basic emotions. Concretely, they apply *distant supervision* considering that any tweet that contains the following hashtags: #anger, #disgust, #fear, #happy, #sadness, and #surprise is likely to express this emotion in some way.

Wang et al. (2012) corpus. They also employed the hashtag phenomenon on Twitter to build their emotion corpus. Concretely, Wang corpus contains about 2.5 million tweets annotated automatically the basic emotions proposed by Shaver et al. (1987): JOY, SADNESS, ANGER, LOVE, FEAR, SURPRISE and they also added THANKFULNESS. The process of determination the hashtags used to collect the tweets consisted of collecting 7 sets of emotion words, one for each emotion. One source of the emotion words is Shaver, where the emotions are organized into a hierarchy in which the first layer contains six basic emotion and the second layer contains 25 secondary emotions that are subcategories of the six basic emotions. Moreover, the authors extended each list of words by including their lexical variants and removed ambiguous words.

Suttles and Ide (2013) corpus is a Twitter dataset of 5.9 million tweets annotated with the Plutchik's eight primary emotion: ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE, TRUST, and ANTICIPATION. It was built up

employing a *distant supervision* process consisted of two main steps: 1) collecting tweets without any sampling on specific query terms; and 2) selecting those tweets that contain any of the emotional tokens defined in their lexicon, including tweets labels appearing both within and at the end of the tweet. Their lexicon comprises a combination of emotional labels including 56 hashtags, 69 traditional emoticons derived from Wikipedia⁶, and 70 emoji labeled by them with the eight Plutchick emotion categories. The strategy for selecting the hashtags was to analyze which was the most frequent hashtags in their dataset that reflect actual user behavior.

Qadir and Riloff (2013) corpus. They present a bootstrapping algorithm to automatically learn emotion hashtags that subsequently are employed in the *distant supervision* technique. As the rest of works presented so far, Qadir corpus was collected considering that any tweet that contains an emotion hashtags is expressing this emotion. The main difference with the rest of corpora is that the selection process of the hashtags is through a bootstrapping method. To achieve that, firstly, they defined five "seed" hashtags for each emotion class: AFFECTION, ANGER/RAGE, FEAR/ANXIETY, JOY and SADNESS/DISAPPOINTMENT. These emotion categories were defined by them based on Parrott (2001)'s emotion taxonomy and how these emotions are expressed in tweets. After that, the algorithm runs in two steps: 1) for each seed hashtag, they search Twitter for tweets that contain the hashtag and label these tweets with that emotion class. They use these labeled tweets to train a supervised classifier for every emotion; 2) the emotion classifiers are applied to unlabeled tweets and those tweet labeled by the classifier are analyzed to extract the new hashtags. The process started with a seed labeled training dataset of 325,343 tweets and 2.3 unlabeled tweets.

The **EmotionExpert** game was developed by Munezero et al. (2013) with the aim of improving emotion annotations. It is a Facebook game consisting of in a manual annotation task of 100 public posts of the website `youopenbook.org`. The EmotionExpert game utilizes the primary emotions defined by Parrott (2001)'s hierarchy as emotion labels: ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE and LOVE. Moreover, with the aim of motivating and engaging players, the game used the Facebook social graph to encourage players to play against their Facebook friends. EmoExpert game has four game levels: BabyEmotioner, EmotionTrainee, EmotionGraduate

⁶https://en.wikipedia.org/wiki/List_of_emoticons

and Emotion Master. For evaluating the reliability of the annotation, a 'gold standard' of the 100 post was carried out by three experts in the field of emotion research. The average pairwise percentage agreement is the metric employed to calculate the agreement, obtaining 88,67% for the experts and 37,8% for Facebook players.

Dystemo corpus (Sintsova & Pu, 2016) consists of a 63,878 collection of tweets annotated using a *distant supervision* method where 167 emotion hashtags employed as labels. These emotional hashtags were defined for the 20 emotion categories from GEW. They started from a corpus of Twitter post about the 2012 Olympic Games obtained by querying Olympic-related keywords such as "Olympic" or "London2012". 250,000 tweets were chosen randomly from this corpus to create the "presumably-neutral" data.

Chinese Microblog corpus. M. Li et al. (2016) present a semi-automatic method employed to build an emotion corpus of 39,721 tweets. The first step is collect a corpus using *distant supervision* through hashtags associated with these emotional categories: LIKE, DISGUST, HAPPINESS, SADNESS, ANGER, SURPRISE, and FEAR. After that, a refinement process is applied in the resultant corpus with the aim of reducing the noise. This process consisting of three main steps: 1) a lexicon-based approach where they count words with emotional content to verify the most voted emotion correspond with the hashtag associated automatically; 2) an SVM classifier is trained with the tweets classified in the previous step and this classifier is employed to annotate the rest of the data; 3) those tweets that can be classified in the previous step are asked to one trained annotator for its manual annotation.

Tweet Emotion Intensity Dataset (Mohammad & Bravo-Marquez, 2017) is a collection of 7,097 tweets annotated for ANGER, FEAR, JOY, and SADNESS with Best-Worst Scaling (BWS) (Louviere, 1991) technique developed for the WASSA-2017 Shared Task on Emotion Intensity. The annotation methodology is a semi-automatic consisting of two main steps: 1) collecting tweets by *distant supervision* using a set of 50 to 100 emotion words extracted from Roget's Thesaurus for each emotion; 2) these tweets are annotated with a manual annotation process using BWS. With this technique, annotators were presented with four tweets at a time (4-tuples) and asked to select the speakers of the tweets with the highest and lowest emotion intensity. To

Table 2.6: Categorical Emotion Corpora (semi/automatic annotation)

Corpus	Source	Size	Emotion Categories	Method
Read (2005) corpus	News Articles	13,000 articles	11 categories	<i>distant supervision</i> from emoticons
Mishne (2005) corpus	Blog Posts	815,494 blogs	132 common moods (LiveJournal)	<i>distant supervision</i> from LiveJournal
Purver and Battersby (2012) corpus	Twitter Messages	-	Ekman (1992)'s emotions	<i>distant supervision</i> from hashtags
Mood and Affects corpus (Choudhury et al., 2012)	Twitter Messages	6.8 million on tweets	11 affects defined by PANAS-X (Watson & Clark, 1994)	<i>distant supervision</i> from hashtags
TEC (Mohammad, 2012a)	Twitter Messages	21,000 tweets	Ekman (1992)'s emotions	<i>distant supervision</i> from hashtags
Wang et al. (2012) corpus	Twitter Messages	2.5 million tweets	Shaver et al. (1987)'s emotions	<i>distant supervision</i> from hashtags
Suttles and Ide (2013) corpus	Twitter Messages	5.9 million tweets	Plutchik (1980)'s emotions	<i>distant supervision</i> from hashtags and emoticons
Qadir and Riloff (2013) corpus	Twitter Messages	2.6 million tweets	Parrott (2001)'s emotions	<i>distant supervision</i> from hashtags
EmotionExpert corpus (Munezero et al., 2013)	Facebook Posts	100 posts	Parrott (2001)'s emotions (primary emotions)	Facebook game
Dystemo corpus (Sintsova & Pu, 2016)	Twitter Messages	63,878 tweets	20 categories (GEW) (Scherer, 2005)	<i>distant supervision</i> from hashtags
Chinese Microblog corpus. M. Li et al. (2016)	Twitter Messages	39,721 tweets	7 categories	<i>distant supervision</i> + refinement process
Tweet Emotion Intensity Dataset (Mohammad & Bravo-Marquez, 2017)	Twitter Messages	7,097 tweets	4 emotions	<i>distant supervision</i> + manual annotation

carry out the second step, CF platform was employed and every 4-tuple was annotated by three independent annotators.

2.3.2.2 Dimensional Emotion Corpora

Généreux and Evans (2006) corpus. They employ the option "current mood" of LiveJournal that provides the mood of the users to automatically collect a corpus, as the Mishne corpus. However, their approach is based on dimensional emotion representation and use the typology of affect from Scherer (2005) to represent each mood in one particular quadrant of the two-dimensional Osgood's Evaluation-Activation space. Thus, Genereux corpus is a dataset of 156015 weblog post in English associated with these tags: Q1 (negative evaluation - active activity), Q2 (negative evaluation - passive activity), Q3 (positive evaluation - passive activity) and Q4 (positive evaluation - active activity).

EMOTEX corpus (Hasan et al., 2014) is a collection of 134100 tweets obtained using *distant supervision* technique through a set of 28 affect words as hashtags. They utilize the Circumplex Model of Affect, shown in Figure 2.2, and define four classes of emotion: HAPPY-ACTIVE, HAPPY-INACTIVE, UNHAPPY-ACTIVE, and UNHAPPY-INACTIVE, corresponding with the quadrants of the two-dimensional circular space defined by Russell (1980).

Table 2.7: Dimensional Emotion Corpora (semi/automatic annotation)

Corpus	Source	Size	Emotion Dimensions	Method
Généreux and Evans (2006) corpus	Blog Posts	156,015 blogs	<i>valence - arousal</i> \Rightarrow Q1, Q2, Q3, Q4	<i>distant supervision</i> from LiveJournal
EMOTEX corpus (Hasan et al., 2014)	Twitter Messages	134,100 tweets	<i>valence - arousal</i> \Rightarrow HAPPY-ACTIVE, HAPPY-INACTIVE, UNHAPPY-ACTIVE, UNHAPPY-INACTIVE	<i>distant supervision</i> from hashtags

2.3.3 Discussion

This section summarizes the most relevant emotion corpora developed for automatic emotion detection in text. The resources have been grouped by its creation process and in each of these groups, they have been classified by

its emotion connotation: categorical models or dimensional one.

The classification by their emotion connotation allows again verifying that the categorical model is the most employed in CL, because, as happens in emotion lexicons, the number of the categorical emotion corpora is substantially higher than the resources based on dimensional models. Focusing on categorical ones, it is possible to observe that Ekman's basic emotions are one of the most employed emotion theories (Strapparava & Mihalcea, 2007; Aman & Szpakowicz, 2007; Roberts et al., 2012; Mihalcea & Strapparava, 2012; Balabantaray et al., 2012; Purver & Battersby, 2012; Mohammad, 2012a).

The review of manually annotated corpora shows different methodologies to carry out the task. For instance, there are annotation tasks where annotators receive a training to understand correctly the objective, some tasks are carried out by experts and/or non-experts, other tasks use or not crowdsourcing platforms, etc. This is due to that there is no single method or process to develop manual annotation. However, as previously concluded in Section 2.2.3, the use of crowdsourcing platforms provides a set of advantages which makes it is one of the methods most used today when a manual annotation corpus is required.

On the subject of semi/automatic methodologies, *distant supervision* is the technique mainly applied to emotion corpora. This is due to its efficiency to built up large emotion corpora and the emergence of social networks as Twitter or Facebook. However, as mentioned in the introduction of this section, this technique has also associated a set of drawbacks that could affect the performance of emotion systems. While it is true that applying refinement processes to *distant supervision* like (Liew et al., 2016) or (Mohammad & Bravo-Marquez, 2017) could palliate these disadvantages, *distant supervision* has another limitation because it can be exclusively applied in genres where the use of hashtags or emoticons is widespread.

Focusing on categorical models, the classification by its creation process allows observing that the corpora developed between 2007 and 2011 have largely been annotated with manual processes. Instead, the methodologies employed for the development of corpora created between 2012 and 2017 are mostly semi/automatic techniques. This tendency is due to two main facts:

- **The disadvantages of manual labelling** (Fort et al., 2012). The cost and time, in terms of human effort, slows down the development of an accurate emotion recognition systems. Moreover, the annotation scheme, the difficulty of the task or the training of the annotators are factors that turn the task into a difficult assignment. These aspects are even more complex to define in emotion manual annotation task because of its highly subjective.
- **The exponential growth in the amount of subjective information on the Web 2.0.** This phenomenon creates a need for improving the emotion annotation task and therefore to propose effective and efficient new methodologies to tackle it.

This tendency is not evident in dimension emotion corpora because the number of these resources is low. However, there is recently an increase of these corpora but so far the majority of them have been manually annotated as was the case of first categorical emotion corpora.

2.4 Conclusion

In this second chapter, we presented the most relevant resources created in the framework of textual ER. In order to have a clear view of what has been done, we grouped all of them according to its creation process (manual or semi/automatic) and its emotion connotation (categorical or dimensional). We believe that this classification of resources is crucial to have first of all an overview of what has been done, but also to understand what is the contribution we bring with our research.

After having performed this exhaustive analysis, we can say that there is an evident lack of semi/automatic methodologies of textual emotion annotation for others genres apart from microblogging (Twitter) or social networks (Facebook) since the Web 2.0 offers other social media platforms (blogs, online news or commentaries about products, news or videos) that can not be analyzed with these techniques. Furthermore, most of the works done are focused on applying *distant supervision* technique without considering the drawbacks associated with this approach.

There is a need to develop new techniques able to build up large emotion corpora in different genres apart from Twitter or Facebook with high quality and high reliability. According to our opinion, after years of developing genre-dependent approaches, new flexible methodologies should be explored with the aim of palliating the drawbacks of manual labeling and to be able to analyze the data available in the Web 2.0.

Having taken into account this context, the methodologies presented in this research have been designed to overcome the abovementioned challenges: lack of semi/automatic genre-independent methodologies for large-scale annotation of emotional corpora with high standards of reliability.

As a result of the conclusions drawn from the state of the art and a reflection on the pending issues, next chapter explores the annotation techniques employed in other [NLP](#) tasks with the aim of analyzing alternative proposals to label emotion training data. This analysis is focused on two main techniques: bootstrapping approaches for [IL](#) and procedures where a pre-annotation process is employed.

Background in Annotation Techniques

“If I have seen further than others, it is by standing upon the shoulders of giants.”

ISAAC NEWTON

After having reviewed emotion resources and concluded that there is need to develop new techniques able to build up large emotion corpora in different genres with high quality and high reliability, this chapter presents a review of the annotation techniques employed in other Natural Language Processing (NLP) disciplines with the aim of simplifying and improving the annotation process and thus reducing time and cost of its development. This study has as main objective exploring alternative annotation techniques to tackle textual emotion labeling.

Supervised Machine Learning (ML) algorithms are widely used in NLP tasks since they usually lead better results than unsupervised approaches. Despite this, the supervised learning algorithm presents some difficulties such as the fact that it requires a large number of labeled training examples for accurate learning since these examples are employed by the learner algorithm to make predictions for all unseen examples. Thus, the application of supervised ML algorithms depends directly on the availability of the labeled training examples.

The creation of labeled training examples is a complex and expensive task in any **NLP** area, since most of them are manually annotated. For this reason, the problems of creating a labeled corpus also affect by other **NLP** tasks such as Named Entity Recognition (**NER**), Part-Of-Speech (**POS**) tagging or Semantic Frame/Role Labelling. Consequently, there is a number of techniques explored by other tasks that can be used to simplify the annotation process and efficiently develop the annotation. Examples of these techniques are the application of game design principles to the task, the pre-annotation of the data or bootstrapping approaches.

Concretely, this chapter presents a study of the two annotation techniques adopted in this dissertation to tackle emotion annotations for its usability and practicality demonstrated in other disciplines: bootstrapping technique for **IL** (Section 3.1) and pre-annotation process (Section 3.2).

3.1 Bootstrapping Technique for Intensional Learning

Bootstrapping is a strategy to automatically generate a number of sufficient instances from a small set of seed words, phrases or sentences. This technique was proposed to avoid, or at least considerably reduce, the need for manual corpora annotation. Hence, it has become an important topic in **NLP** since for many language-processing tasks there is an abundance of unlabeled data.

Despite there is a large variety of bootstrapping implementation, they all implement these two main steps:

1. An *initial step* where a labeled seed of words, phrase or sentence is created.
2. A *learning step* where a supervised classifier is trained from the data created in the previous step. This second step is possibly reiterated, but is not always necessary.

According to the terminology from computability theory and as [Gliozzo et al. \(2009\)](#) describe, all bootstrapping implementations can be classified into Extensional Learning (**EL**) and Intensional Learning (**IL**). **EL** is the standard example-based supervision mode where the seed is a small set of

examples (words, phrases or sentence) annotated with a category which has generally been manually associated. However, **IL** is a feature-based supervision where the user is expected to specify exact classification rules that operate in the feature space. The features may often be perceived as describing the *intension* of a category.

Between these two groups, **EL** is the most popular in **NLP**. These approaches generally are run with several iterations in step 2 where the most confident predictions of the algorithm are added to the initial seed in each iteration. The core feature of **EL** algorithms is that the output of the first iteration is used as the input of the second iteration, and so on. This kind of bootstrapping has been employed to extract automatically patterns to identify subjective words (Riloff et al., 2003; Banea et al., 2008) or for the construction of English and Italian corpora from the domain in Psychiatry via automated Google queries (Baroni & Bernardini, 2004).

However, although **IL** is less popular in **NLP** than **EL**, this strategy can be found in the literature as a technique for bootstrapping an **EL** algorithm (Yarowsky, 1995; Collins & Singer, 1999), but especially can be found to tackle **TC** task as we will see in the following paragraphs.

Mccallum and Nigam (1999) presented an alternative bootstrapping approach to **TC** consisting of using a small set of keywords per class, a class hierarchy and a large quantity of easily-obtained unlabeled documents. The first step in the bootstrapping is to use the keywords to generate preliminary labels for as many of the unlabeled document as possible by term-matching. Each class is given just a few keywords. These preliminary labels become the starting point for a bootstrapping process that learns Naive Bayes classifier using Expectation-Maximization (**EM**) and hierarchical shrinkage. The classifier learned by bootstrapping reaches 66% accuracy, a level close to human agreement.

Ko and Seo (2004a) proposed a new automatic **TC** method for learning from only unlabeled data using a bootstrapping framework and a feature projection technique. The input to the bootstrapping process consists in a large amount of unlabeled data and a small amount of seed information in the form of the title words associated with categories. At first, they automatically create keywords from a title word for each category. Then, centroid-context are extracted using the title word and keywords. Finally,

they build up context-clusters by assigning remaining contexts to each context-cluster using a similarity measure technique. This context-cluster are used to obtain labeled training data employing a Naive Bayes classifier. Once they obtained labeled data, they employ TCFP (Ko & Seo, 2004b), a feature projection technique for learning text classifiers. The results show reasonably comparable performed in comparison with that on the supervised Naive Bayes classifier. Moreover, it outperforms a clustering method proposed by Slonim et al. (2002).

Liu et al. (2004) presented an approach for TC. The proposed method labels a collection of representative words for each class to extract a set of labeled documents from unlabeled documents and form the initial training dataset. The method to obtain a set of representative words (ranked list) for each class consists of combining clustering and features selection. Then, the user select/labels some words from the ranked list for each class and the initial collection of documents is automatically labeled. Once the first step is finished, the EM algorithm is applied to build the classifier. The results show that this method for TC is highly effective and promising.

Gliozzo et al. (2009) proposed a generalized bootstrapping algorithm in which categories are described by relevant seed features. Their method includes two unsupervised steps to build the initial categorization step of the bootstrapping scheme: i) using Latent Semantic Analysis (LSA) space to obtain a generalized similarity measure between instances and features and ii) the Gaussian Mixture algorithm to obtain uniform classification probabilities for unlabeled sentences. The initial training set obtained by this first step is exploited to train a Support Vector Machine (SVM) classifier. The proposal was evaluated for TC and obtaining state-of-the-art performance using only the category names as initial seeds.

From these works and according to Gliozzo et al. (2009), it is possible to recognize a common structure for IL proposals based on a typical bootstrap schema (Yarowsky, 1995; Collins & Singer, 1999):

Step 1 *Initial similarity-based categorization.* This step is approached by applying a similarity criterion between the initial category seed and each unlabeled sentence. The result of this step is an initial categorization of (possibly a subset of) the unlabeled documents.

Step 2 *Training of a supervised classifier on the initially categorized set with one or more iterations.* The output of step 1 is exploited to train an (extensional) supervised classifier. Different learning algorithms have been tested, as [SVM](#) or Naive Bayes.

The core part of [IL](#) bootstrapping is step 1, that is, the initial unsupervised classification of the unlabeled dataset. This step has often been approached by simple method assuming that the supervised training algorithm would be robust enough to deal with noise from the initial set. Despite this assumption, the effectiveness of the first step is crucial for the satisfactory performance on the subsequent supervised training ([Gliozzo et al., 2009](#)).

To sum up, this study of the bootstrapping technique for [IL](#) shows its effectiveness and practicability for [TC](#) task since all of them obtain results reasonably comparable performed in comparison with that on supervised approaches. Hence, it is a technique to be considered when labeled data are lacking and too expensive to be create in large quantities. Moreover, one of the most relevant features of [IL](#) is that they are unsupervised proposals and thus build classifiers from unlabeled data. This is the main reason why we decide to evaluate [IL](#) for emotion annotation since we consider really interesting the fact that the annotations were not influenced by the human annotators' background.

3.2 Pre-annotation Process

Pre-annotation, or pre-tagging, is a procedure to automatically annotate a corpus by using an automatic system and to present these annotations to the human annotator. The human annotators then typically correct mistakes or omissions made by the automatic system, or alternatively make a choice between different options given by the automatic system ([Skeppstedt et al., 2017](#)). This technique has been widely studied in [NLP](#) tasks such as [NER](#), [POS](#) tagging or Word Sense Disambiguation ([WSD](#)), reporting a gain in time and quality in manual annotation tasks as we will see in the following paragraphs.

[Marcus et al. \(1993\)](#) work is one of the first approaches where the pre-annotation process is assessed for [POS](#) tagging during the creation of the Penn Treebank, a corpus consisting of over 4.5 million words of American English.

In this work, the pre-annotation stage is carried out with an automatic POS assignment provided by a cascade of stochastic and rule-driven taggers developed by themselves. Then, the results of the first step are given to annotators to manually correct. Finally, the manual correction is evaluated to determine how to maximize the speed, inter-annotator consistency, and accuracy of POS tagging. The experiment showed that manual tagging took about twice as long as correcting, with about twice the inter-annotator disagreement rate and an error rate that was about 50% higher.

Chou et al. (2006) present a semi-automatic process to construct a bio-medical proposition bank (BioProp) containing annotations of predicate-argument structures and semantic roles in a treebank schema. To achieve that, a Semantic Role Labeling (SRL) system trained on PropBank (**Palmer et al., 2005**), an annotated corpus of semantic roles on the Penn Treebank (**Marcus et al., 1993**), is used to pre-annotate BioProp. In a second phase, the incorrect tagging results are corrected by human annotators. The experimentation performed shows that the annotation effort could be reduced by 46% in case of employing the pre-annotation process.

Ganchev et al. (2007) evaluate a semi-automated process for NER. Its pre-annotation process is based on linear sequence model trained using a k-best MIRA learning algorithm (**Crammer et al., 2006; McDonald et al., 2005**). Linear sequence models score possible tag sequences for a given input. These proposals are asked to the human annotators to decide whether each mention produced by a high recall tagger is a true mention or a false positive. Their proposal demonstrates that can reduce the effort of extending a seed training corpus by up to 58%.

Rehbein et al. (2009) performed quite thorough experiments to assess the benefits of partial automatic pre-annotation on a frame assignment (WSD) task. They compare three annotation tasks to explore the effect of pre-annotation quality: i) applying Shalmaneser (**Erk & Pado, 2006**), an Automatic Semantic Role Labeling (ASRL) system, as the pre-annotation process, ii) using a enriched pre-annotation method created by manually inserting errors into the gold standard, and iii) carrying out the manual annotation task without pre-annotation. The results of this experimentation have not been able to show that automatic pre-annotation speeds up the labeling process. However, they show that pre-annotation has a positive

effect on the quality of human labeling and that even noisy and low-quality pre-annotation does not overall corrupt human judgment.

Fort and Sagot (2010) evaluate the influence of automatic pre-annotation on the manual POS annotation of a corpus, both from the quality and the time points of view, with specific attention to biases. To achieve that, they designed different experiment setups with the aim of analyzing: i) the impact of pre-annotation accuracy on precision and Inter-Annotator Agreement (IAA), ii) the impact of the pre-annotation accuracy on annotation time, and iii) the bias induced by pre-annotation. In all experiments, they use different taggers with varying degrees of accuracy built up training MElt POS tagger (Denis & Sagot, 2009) on increasingly larger parts of the POS tagger Penn Treebank. Their experiments confirmed and detailed the gain in quality and demonstrated that even a not so accurate tagger can help improve annotation speed.

Lingren et al. (2014) analyze the impact of pre-annotation on annotation speed and potential bias for clinical NER in clinical trial announcements. While the most of pre-annotation processes are based on ML systems with varying size of training data, in this work they use two dictionary-based methods to pre-annotate the text. With the evaluation performed, they concluded that the one with the pre-annotated text needed less time to annotate than the annotator with non-labeled text. Moreover, the pre-annotation did not reduce the IAA or annotator performance. Thus, the experiments demonstrate that the dictionary-based pre-annotation is a feasible and practical method to reduce the cost of clinical NER annotation.

Henriksson et al. (2015) report on the creation of an annotated corpus of Swedish health records for the purpose of learning to identify information pertaining to Adverse Drug Events (ADEs) present in clinical notes. To achieve that, three key tasks were tackled: recognizing relevant named entities, labeling attributes of the recognized entities, and the relationship between them. To speed up and facilitate the human annotation effort, the documents were pre-annotated for NER task using a condition random fields model trained on previously manually annotated health records from an internal medicine emergency unit in Stockholm (Skeppstedt et al., 2014). The F1-scores show better results than the values obtained in the original study (Skeppstedt et al., 2014) and the annotators appreciated the pre-annotations

because allowed them to focus on the more difficult task of assigning relations.

Bu et al. (2016) work also analyze the use of a pre-annotation process for DBpedia entity annotation. They evaluate three annotation approaches: W_{auto} , W_{naive} , and W_{free} . The W_{auto} method automatically generates possible candidates (DBpedia categories) using an entity typing tool, and then these candidates are proposed to human annotators with the aim of detecting and correcting errors. The W_{naive} process asks the human annotators to choose the DBpedia category by traversing from the root top-down until a specific category is selected. And W_{free} proposal tries to match between the text introduced by human annotators with a DBpedia category using a process based on textual similarity. While W_{auto} and W_{free} use automatic processes to help human annotators and improve the quality of annotations, W_{naive} is a completely manual process. Regarding the evaluation, W_{naive} obtains the best values in terms of errors detection and correction costs because it is carried out by humans. However, the best values of prediction cost are obtained by W_{free} and W_{auto} proposals, demonstrating the usability of automatic processes in manual annotation task.

From these works, it is possible to note that all pre-annotation processes contain two main stages: i) the data is pre-annotated with a pre-defined set of categories and ii) then humans annotators carry out a manual refinement process. With regards to the first stage, there are several pre-annotation processes but the majority is approached by ML systems with varying size of training data. About the manual refinement stage, the human annotator can correct mistakes or omissions made by the pre-annotation method or makes a choice between different options given by the first stage.

On the whole, the review allows us to note that this methodology has been widely applied in other NLP areas, obtaining encouraging results in terms of cost and time needed to developed labeled data. Moreover, the fact that an automatic process can help human annotators in labelling tasks is interesting from emotion detection point of view since if we want to have a human feedback on our resource, we need those human annotators participate in the annotation task. After all, we want to identify human emotions. Thus, we consider that the pre-annotation process is a suitable methodology to tackle emotion annotation.

3.3 Conclusion

As presented at the beginning of this chapter, the rationale beyond of this review is the need of exploring alternative annotation techniques to tackle textual emotion labeling.

In this chapter, we present a review of two techniques applied in other **NLP** tasks with the aim of simplifying and improving the annotation process and thus reducing time and cost of its development. These techniques are bootstrapping for **IL** and the pre-annotation process.

IL is a generalized bootstrapping algorithm that builds classifiers from unlabeled data, whereas the pre-annotation process is a semi-automatic technique where human annotators participate in the development process. Both of them have been assessed in different **NLP** areas such as **NER**, **POS** tagging or **WSD**, obtaining interesting results in terms of cost and time needed to developed labeled data.

Hence, the features of each technique, as well as, the usability and practicability demonstrate in other **NLP** tasks allow us to consider the bootstrapping technique for **IL** and the pre-annotation process suitable to tackle emotion annotation task with the aim of improving its development in terms of time and cost.

As a result of the conclusions draws from this review and a reflection on the pending issues in emotion resources draws from the previous chapter (Section 2.4), next chapter describes the first methodology proposed in this research: a **IL** proposal for emotion annotation. The process is described in detail and also an exhaustive evaluation is carried out to analyze its effectiveness. With this, our objective is to present the product of our research and to underline our contribution towards the improvement of the state of the art.

Intensional Learning for Emotion Annotation

“The technologies which have had the most profound effects on human life are usually simple.”

FREEMAN DYSON

This chapter provides our first proposal to efficiently tackle emotion annotation task in text: Intensional Learning (**IL**). It is an alternative bootstrapping approach proposed in (Gliozzo et al., 2009) consisting of two main steps:

Step 1 *Initial similarity-based categorization.* This step is approached by applying a similarity criterion between the initial category seed and each unlabeled sentence. The result of this step is an initial categorization of (possibly a subset of) the unlabeled documents.

Step 2 *Training of a supervised classifier on the initially categorized set with one or more iterations.* The output of step 1 is exploited to train an (extensional) supervised classifier. Different learning algorithms have been tested, as Support Vector Machine (**SVM**) or Naive Bayes.

Unlike the example-based supervision mode (Extensional Learning (**EL**)), **IL** approach is based on the classical rule-based classification method, where the user specifies exact classification rules that operate in the features space.

This is particularly relevant for Emotion Recognition (ER) since, in EL, the fact that the examples are manually annotated by humans implies that everyone's personal context can influence emotion interpretation. However, in IL, the influence of human emotion understanding is reduced since their participation is limited to the rules definition.

Given our dissertation objective of developing efficient techniques able to build up large emotion corpora in different genres with high quality and high reliability, the purpose of this chapter is to assess the usability to IL to achieve our main objective.

In this chapter, we will detail the entire process, starting with a complete description of the IL technique, following up with the evaluation performed, and finally the conclusions drawn from this experimentation.

4.1 Intensional Learning Process

As previously introduced, the common structure of IL is two main steps. Specifically, our bootstrapping approach is shown in Figure 4.1 and consists in:

- **Step 1** *Initial similarity-based categorization*. In particular, this step consists of two unsupervised sub-steps:
 - **Step 1.1**: the creation of an initial seed where an emotion lexicon is employed to annotate the sentence by its emotional words.
 - **Step 1.2**: the extension of the initial seed based on the measure of the semantic similarity between sentences.
- **Step 2** *Training of an (extensional) supervised classifier on the initially categorized set with one or more iterations*. Our supervised classifier (SVM) performs one iteration since, as explained by Gliozzo et al. (2009), the improvements reported by some works (Mccallum & Nigam, 1999; Liu et al., 2004) that perform an iterative re-estimation algorithm are small incremental and hence do not seem to justify the additional effort.

The process receives as input data a collection of unlabelled sentences or phrases, a set of emotion categories (e.g. Ekman (1992)'s basic emotions,

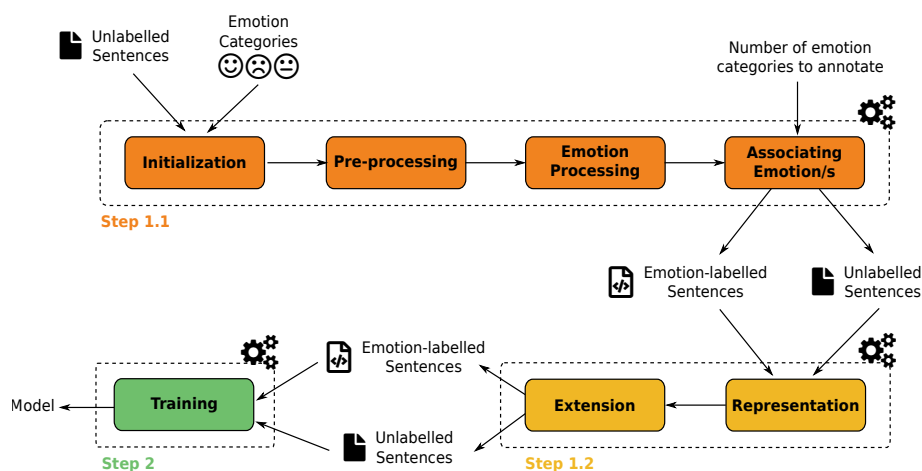


Figure 4.1: Overview of the Intensional Learning (IL) process.

Plutchik (1962)’s wheel of emotions or Izard (1971)’s emotions), and the number of emotion categories to annotate (one or more categories).

The adaptability of the proposal to the set of emotion categories, as well as, the number of emotion categories annotated (the dominant emotion or all of the emotions detected) is engaging and novel since this flexibility allows the use of this technique in different domains or applications. For instance, BOREDOM, ANXIETY and EXCITEMENT emotions are typically detected in education domain (Kim, 2011), whereas emotions like AMUSED or INSPIRED are analyzed in news domain¹. Moreover, this adaptability can be useful in those applications where the detection of the emotion intensity is important such as recommender systems.

In the following subsections, each step of our IL proposal is detailed.

4.1.1 Step 1.1: Selecting of Seed Sentences

This section describes the algorithm performed to create the initial seed consisting in the following steps:

- **Step 1 Initialization:** each sentence has an emotional vector associated with a value of each emotion (e.g. [ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE]) initialized to zero.

¹<http://www.rappler.com/>

- **Step 2** *Pre-processing*: each sentence is tokenized and lemmatized using a language analysis tool.
- **Step 3** *Emotion processing*: each word of the sentence is looked up in an emotion lexicon. If a word is in the lexicon, its emotional values are added to the emotional vector of the corresponding sentence.
- **Step 4** *Associating emotion/s*: if the process annotates the dominant emotion, each sentence is annotated with the emotion whose value is the highest in the emotional vector of the sentence. Instead, if the process annotates all of the emotions expressed in the sentence, each sentence is annotated with all of the emotions detected.

The simplicity of the algorithm provides adaptability and flexibility to the process because the requirements to run the algorithm is to provide: 1) an emotion lexicon annotated with the desired emotion categories, and 2) a language analysis tool to pre-process the text in the desired language.

Linguistic phenomena such as negation or irony have not been addressed in this approach because the objective of our research is to propose a technique for large-scale annotation in any genre with the aim of reducing cost and time-effort. The management of these phenomena introduces a high level of complexity in the approach since the detection of these aspects require in depth analysis of each genre, thereby that could hampering the achievement of our purpose.

Our particular approach is shown in Figure 4.2 and is implemented with Ekman's basic emotions as the set of emotion categories, *EmoLex* as emotion lexicon, and Stanford Core NLP Pipeline (Manning et al., 2014) as a language analysis tool.

As previously mentioned in Section 2.1.1, the most popular set of "basic" emotions is Ekman (1992)'s emotions since have been previously used in other computational approaches to emotion (Roberts et al., 2012; Mihalcea & Strapparava, 2012; Balabantaray et al., 2012; Purver & Battersby, 2012; Mohammad, 2012a). Moreover, these emotions have been most widely accepted by the different researchers (see Table 2.1). Consequently, we consider that the Ekman's emotions are the most suitable set of emotions to perform our proposal.

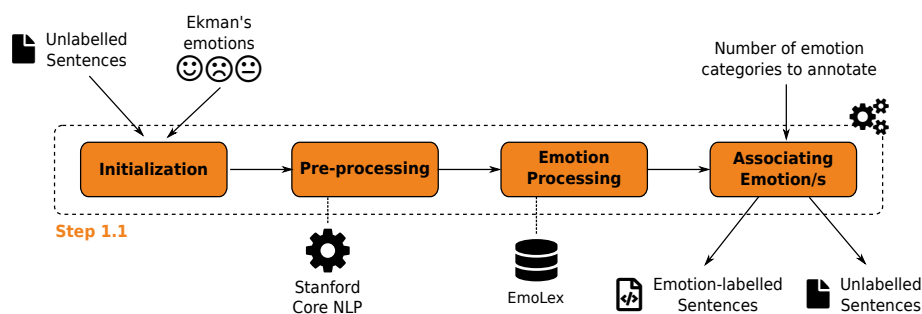


Figure 4.2: Workflow of the step 1.1 in Intensional Learning (IL) process.

The election of Stanford Core NLP Pipeline is due to the wide range of language technology tools it includes: POS tagger, the NER or the parser, as well as the number of languages supported: Arabic, Chinese, English, Spanish, French, and German. This allows us to provide versatility and adaptability to our proposal.

Regarding *EmoLex*, a lexicon of general domain introduced in Section 2.2.1, consisting of 14,000 English words manually compiled by humans and associated with Plutchik (1980)’s eight basic emotions and two polarities: POSITIVE and NEGATIVE. The fact that our proposal employs Ekman’s emotions implies that the lexicon is reduced to 3,462 English words since many words are NEUTRAL after removing ANTICIPATION and TRUST emotions. The coverage of this reduced version of *EmoLex* is shown in Table 4.1.

Table 4.1: Distribution of the emotion words in the reduced version of *EmoLex*.

Anger	Disgust	Fear	Joy	Sadness	Surprise
1,247	1,058	1,476	689	1,191	534

Taking into account the algorithm and the resources selected, two examples of the creation of the seed are shown in Figure 4.3:

- Sentence 1: “*We played fun baby games and caught up on some old time*”. Firstly, its emotional vector is initialized to zero, and after pre-processing, its emotional content is analyzed using *EmoLex*. In this case, the sentence contains three emotional words: ‘fun’, ‘baby’

and 'catch'. The values of these three words are added and the sentence has finally this vector: [0, 0, 0, 2, 0, 1] associated. This sentence will have JOY emotion associated because this emotion has the highest value associated when the process is detecting the dominant emotion and will have JOY and SURPRISE emotions associated when all of the emotions are detected.

- Sentence 2: “*My manager also went to throw a fake punch.*”. Firstly, its emotional vector is initialized to zero, and after pre-processing, its emotional content is analyzed using [EmoLex](#). In this case, the sentence contains one emotional word: 'punch'. The sentence has finally this vector: [1, 0, 1, 0, 1, 1] associated. Hence, if the process is detecting the dominant emotion, this sentence will be not associated any emotion, whereas this sentence will have ANGER, FEAR, SADNESS and SURPRISE emotions associated when the objective is to detect all emotions.

Due to the reduction of [EmoLex](#), its enrichment with synonyms has been considered relevant to test a different set of seeds. For this reason, [EmoLex](#) is extended automatically with WordNet ([WN](#)) ([Miller, 1995](#)) and the Oxford American Writer Thesaurus ([Aubur et al., 2004](#)) synonyms. Thus, three approaches are presented. The process of the seed creation is the same in all cases, however, each one employs different variants of [EmoLex](#):

- *Original*: the original version of [EmoLex](#) is employed.
- *Enriched WN*: an enriched version of [EmoLex](#) with [WN](#) synonyms is used as emotion lexicon.
- *Enriched Oxford*: the emotion lexicon used is an enriched [EmoLex](#) lexicon extended by Oxford synonyms.

The extension process of [EmoLex](#) is completely automatic and is explained in detail in the following subsection.

4.1.1.1 Enrichment Process of [EmoLex](#) with Synonyms

The enrichment of [EmoLex](#) consists in extending it with the synonyms of:

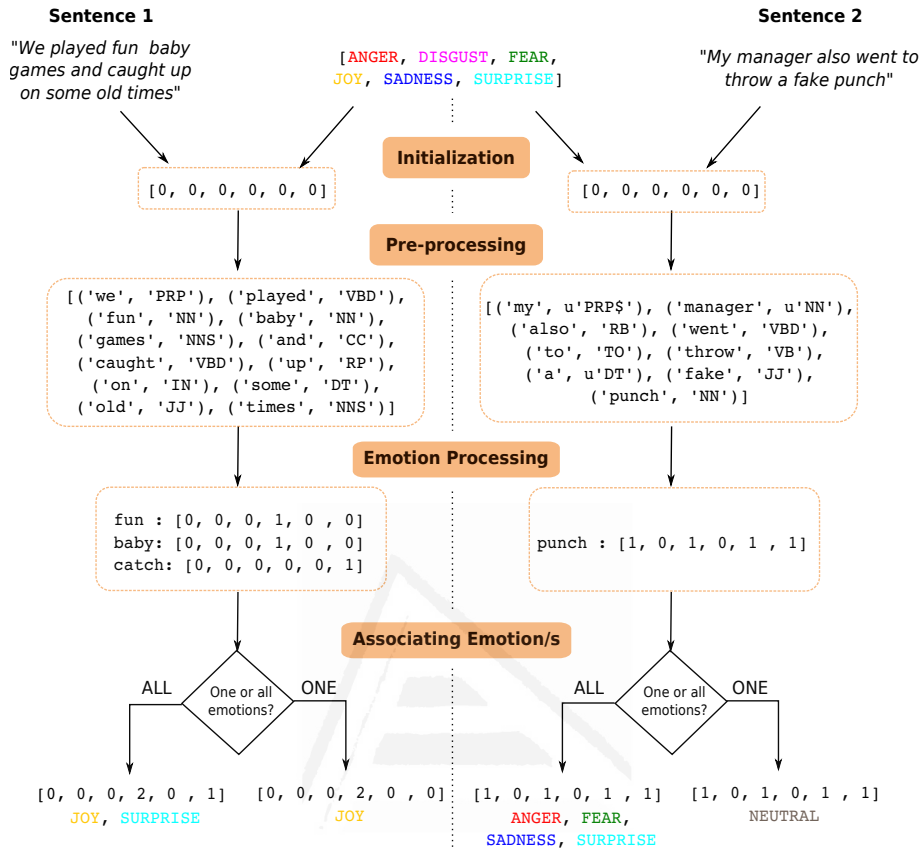


Figure 4.3: Examples of the process of selecting seed sentences (Step 1.1).

- *WordNet (WN)* (Version 3.0) (Miller, 1995): is a large lexical database that contains English words (nouns, verbs, adjectives and adverbs) grouped into sets of cognitive synonyms called *synsets*, each one expressing a distinct concept.
- *Oxford American Writer Thesaurus (1st Edition)* (Aubur et al., 2004): is a thesaurus that lists words grouped together according to the similarity of their meaning, providing a careful selection of the most relevant synonyms, as well as hints for choosing between similar words.

The process consists of several steps shown in Figure 4.4: 1) each word contained in *EmoLex* is looked up in *WN*/*Oxford* and the synonyms of the most frequent sense for *WN* and all of the senses for *Oxford* were obtained; 2) each synonym was associated with the emotions of the *EmoLex* word; 3) if the synonyms are not already in *EmoLex*, they are added. In case that a

synonym is already in **EmoLex**, the emotions associated will be a result of matching the emotion vector stored in **EmoLex** and the new emotion vector.

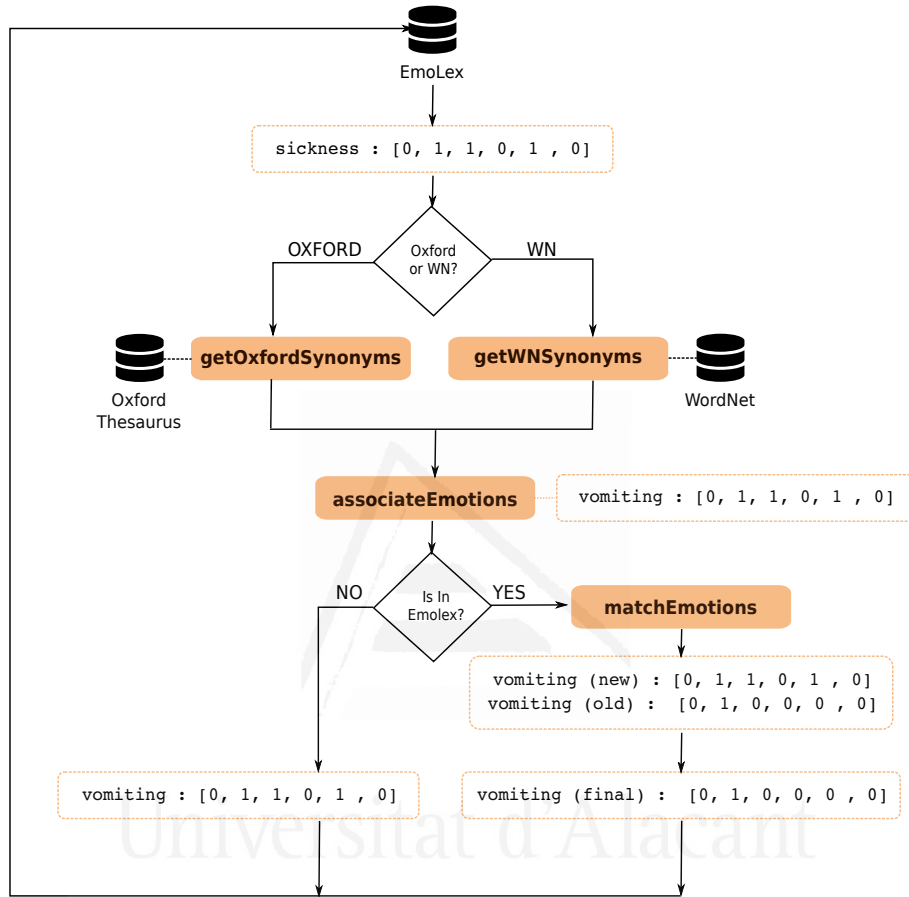


Figure 4.4: Process of the extension of **EmoLex** by **WN** and Oxford synonyms.

Once the process is completed, **EmoLex** was extended with 4,029 **WN** synonyms resulting in a lexicon of 7,491 words for the *Enriched WN* approach. With the extension by Oxford synonyms, the resultant lexicon for *Enriched Oxford* approach consists of 12,664 words where 9,202 are Oxford synonyms. The coverage of this extended versions of **EmoLex** is shown in Table 4.2.

The aim of getting the synonyms of the most frequent sense from **WN** and all of them from Oxford is to evaluate how this distinction affects to the resultant Emotion Recognition (**ER**) models.

Figure 4.4 shows the extension process with an example of obtaining the

synonyms for the word ‘sickness’ on Oxford. The first step gets the Oxford synonyms and for each synonym (in this example the synonym ‘vomiting’): 1) associate the emotions of ‘sickness’, this is, **DISGUST**, **FEAR** and **SADNESS**; and 2) check if ‘vomiting’ is already in **EmoLex**. If it is not, the emotions associated will be the same as ‘sickness’. In another case, their emotional vector will contain the emotion in common between the vector saved in **EmoLex** (old) and the new emotional vector (new). In this case, ‘vomiting’ will be associated with **DISGUST** emotion.

Table 4.2: Distribution of the emotion words in the enriched version of **EmoLex** with **WN** and Oxford synonyms.

Extension	Anger	Disgust	Fear	Joy	Sadness	Surprise
WN	2,683	2,327	3,009	1,524	2,418	1,188
Oxford	3,563	3,259	4,271	3,482	3,644	1,864

4.1.2 Step 1.2: Seed Extension via Semantic Similarity

After obtaining the initial seed sentences, the next step will consist into increasing the number of annotated sentences with the help of Distributional Semantic Models (**DSMs**).

Many approaches have been suggested to determine the semantic similarity between text such as approaches based on lexical matching, handcrafted patterns, syntactic parse trees, external sources of structured semantic knowledge and distributional semantics (**Kenter & de Rijke, 2015**). Our proposal is focused on distributional semantics because we aim to employ a generic model that does not require lexical and nor linguistic analysis and does not use external sources of structured semantic knowledge.

DSMs are based on the assumption that the meaning of a word can be inferred from the way it is used. Therefore, these models dynamically build semantic representations (high-dimensional semantic vector spaces) through a statistical analysis of the contexts in which the words occur². Finally, each word is represented by a real-valued vector called *word vector* or *word*

²<http://wordspace.collocations.de/doku.php/course:acl2010:start>

embedding and the geometric properties of high-dimensional semantic vector spaces prove to be semantically and syntactically meaningful (Mikolov et al., 2013; Pennington et al., 2014), thus words that are semantically or syntactically similar tend to be close in the semantic space.

Latent Semantic Analysis (LSA) and Word2Vec (W2V) algorithms incorporate this intuition. On the one hand, LSA (Deerwester et al., 1990) builds a word-document co-occurrences matrix and performing a dimensional reduction by a Singular Value Decomposition (SVD) on it to get a lower-dimensional representation. On the other hand, W2V algorithm (Mikolov et al., 2013) learns a vector-space representation of the terms by exploiting a two-layer neural network. There are two architectures of W2V: Continuous Bag-Of-Words (CBOW) that predicts the current word based on the context; and Skip-gram (SKIP) which predicts surrounding words given the current word. In this research, both algorithms are employed to build DSMs for several reasons: (i) both methods allow us to employ generic model to calculate the semantic similarity by measuring the distance between the word vectors in the extension of the seed; (ii) LSA and W2V algorithms have demonstrated their effectiveness in calculating the semantic similarity in many NLP tasks such as e-learning (Villalón et al., 2008), Text Categorization (TC) (Gliozzo et al., 2009; L. Li et al., 2014); Emotion Recognition (ER) (Predoiu et al., 2014); or Sentiment Analysis (SA) (García Pablos et al., 2015); (iii) allow us to compare LSA, a consolidated and traditional method, and W2V, a recent technique based on neural networks, in textual ER task.

Compositional Distributional Semantic Models (CDSMs) are employed to determine semantic similarity of sentences/phrases by word embedding. These models are an extension of DSMs that characterize the semantics of entire phrases or sentences. This is achieved by composing the distributional representations of the words that sentences contain (Marelli et al., 2014). Among these models, the approach employed in our research has been used in (Banea et al., 2014) and it is called *VectorSum*. This method consists of adding the vectors corresponding to non-stop words in Bag-Of-Words (BOW) A and B , resulting in a vector V_A and V_B , respectively. The selection of this approach as CDSMs is due to its simplicity and because as Banea et al. (2014) demonstrated, these vectors are able to capture the semantic meaning associated with the contexts, enabling us to gauge their relatedness using cosine similarity.

With these models, the seed sentences will be extended based on the semantic similarity between annotated and non-annotated sentences. In particular, the extension process of the seed has two main steps (Figure 4.5) consisting in:

- **Step 1 Representation:** annotated and non-annotated sentences are represented by distributional vectors employing different DSMs (Table 4.3). The representation of each sentence is achieved by adding the distributional vectors corresponding to non-stop words of each sentence.
- **Step 2 Extension:** this process calculates the cosine similarity between the vectors of labeled and unlabelled sentences. When the similarity is higher than 80%, the non-annotated sentences are annotated with the emotions of the annotated one. If non-annotated sentences could be matched to two or more annotated sentences, the process selects the annotated sentence whose similarity with non-annotated one is higher.

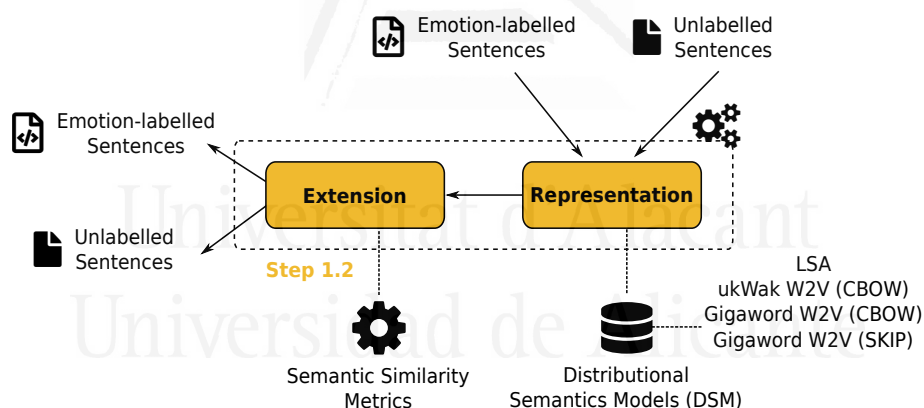


Figure 4.5: Workflow of the step 1.2 in Intensional Learning (IL) process.

The similarity threshold was empirically determined. These experiments showed that the employment of thresholds lower than 80% added noise to the seed. Consequently, the use of a strict similarity (80%) allows us to ensure that the seed is extended with high confidence.

Once the process is completed, we have labeled and unlabeled data that make up our emotion corpus annotated automatically.

Table 4.3: Distributional Semantic Models (DSMs) parameters for IL approaches

DSMs	Parameters	Authors
LSA	<ul style="list-style-type: none"> • source: British National Corpus (BNC) • size: 400-dimensional • pre-processing: lemma#pos 	(Gliozzo & Strapparava, 2009)
ukWak W2V (CBOW)	<ul style="list-style-type: none"> • source: BNC and WackyPedia/ukWaC • size: 300-dimensional • pre-processing: words • architecture: CBOW • context windows: 5 words • sub-sampling: $1e - 05$ • negative examples: 10 	(Dimu & Baroni, 2015)
Gigaword W2V (CBOW)	<ul style="list-style-type: none"> • source: New York Times Newswire Service from the Annotated English Gigaword • size: 100-dimensional • pre-processing: lemma#pos • architecture: CBOW • context windows: 5 words • sub-sampling: $1e - 3$ • negative examples: 5 	HLT-NLP group FBK (Italy)
Gigaword W2V (SKIP)	<ul style="list-style-type: none"> • source: New York Times Newswire Service from the Annotated English Gigaword • size: 100-dimensional • pre-processing: lemma#pos • architecture: SKIP • context windows: 5 words • sub-sampling: $1e - 3$ • negative examples: 5 	HLT-NLP group FBK (Italy)

4.1.3 Step 2: Training supervised classifiers

The second step of IL process consists of exploiting a set of supervised classifiers with the annotated and the non-annotated sentences from the previous step as shown Figure 4.6.

Our proposal performs Support Vector Machine (SVM) algorithm since it

has previously given good performance in textual ER experiments (Aman & Szpakowicz, 2007; Ghazi et al., 2010; Mihalcea & Strapparava, 2012; Anusha & Sandhya, 2015). Concretely, the Sequential Minimal Optimization (SMO) algorithm, the Platt (1999)’s implementation for SVM, is performed with the default parameters (Poly kernel and C value: 1.0) using Weka (Hall et al., 2009). With respect to the features, the sentences are represented as a vector of words weighted by their counts (*StringToWordVector* filter from Weka).

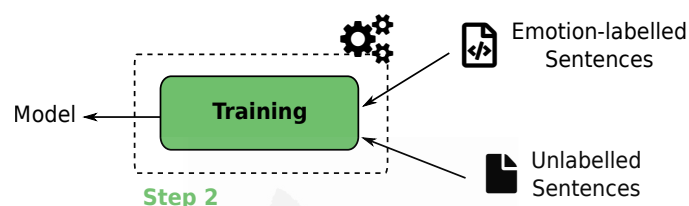


Figure 4.6: Workflow of the step 2 in Intensional Learning (IL) process.

4.2 Evaluation

4.2.1 Data Description

In order to assess the usability for different genres, the approaches are evaluated against two emotion corpora: Aman and Affective Text corpora, previously introduced in Section 2.3.1 (Table 2.4).

Aman corpus (Aman & Szpakowicz, 2007). This dataset contains sentence-level annotation of 4,000 phrases from blog posts collected directly from Web. This corpus was manually developed by four annotators who received no training, though they were given samples of annotated sentences to illustrate the kind of annotations required. It was annotated with the emotion intensity (high, medium, or low) and eight categories: the six emotion proposed by Ekman (1992), MIXED EMOTIONS and NO EMOTION. Despite the initial objective of labelling those sentences with more than one emotion (MIXED EMOTIONS), the gold standard is annotated with Ekman’s emotions and NO EMOTION categories. The distribution of the corpus is shown in Table 4.4.

Affective Text corpus (Strapparava & Mihalcea, 2007). It contains sentence-level annotations of 1,250 short texts from news headlines, which

Table 4.4: Distribution of the sentences per emotion on Aman corpus, a corpus of blog posts annotated with Ekman’s basic emotions.

Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral	Total
179	172	115	536	173	115	2,800	4,090

were drawn from major newspapers such as New York Times, CNN and BBC News, as well as from the Google News. They organized a manual annotation task constituted of six annotators who were instructed to select the appropriate emotions. The annotators assigned a value for each Ekman (1992)’s basic emotion and a value for valence. Hence, each headline had associated a value for each emotion and another one for valence. With respect to the evaluation, they carry out two assessments: fine-grained and coarse-grained. The fine-grained evaluation was conducted using the Pearson (1956)’s correlation between the system and the gold standard scores. In the coarse-grained evaluation, each emotion of the gold standard was mapped to a 0/1 classification ($0=[0,50)$, $1=[50,100]$), and each valence was mapped to a -1/0/1 ($-1=[-100, -50]$, $0=(-50,50)$, $1= [50,100]$). For our assessment, we use the gold standard of course-grained evaluation whose distribution is shown in Table 4.5.

Table 4.5: Distribution of the sentences per emotion on Affective Text corpus, a corpus of headlines annotated with Ekman’s basic emotions.

Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral	Total
41	21	124	148	145	50	796	1,250

These corpora are selected because of several reasons: (i) both corpora are manually annotated allowing us to compare automatic to manual annotation; (ii) they are relevant to emotion detection task since they have been employed in many works to detect emotions (Keshtkar & Inkpen, 2010), (Chaffar & Inkpen, 2011), (Mohammad, 2012b); (iii) both corpora are English sentences annotated with Ekman (1992)’s basic emotions; and (iv) these corpora allow us to test our approaches about corpora with different sources of information:

news headlines and blog posts from Web. Thus, the usability and effectiveness of our approach can be assessed.

4.2.2 Methodology

The evaluation methodology is divided into two steps:

1. Training of an supervised classifier from the corpus annotated automatically to evaluate its usability (Step 2 of [IL](#)).
2. Assessing the quality of automatic annotations through measuring the agreement between our corpus developed by [IL](#) (automatic annotation) and the gold standard of Aman and Affective Text corpora (manual annotation).

With regards to automatic emotion classification, a [SMO](#) multi-classifier is employed on Aman corpus because of it is annotated with the dominant emotion. On Affective Text corpus, six [SMO](#) binary classifiers are applied since each sentence can be annotated with one or more emotions. For the evaluation, the versions of the corpora (Aman corpus and Affective Text corpus) automatically annotated with our approaches are performed with a 10-fold cross-validation. Specifically, precision (P), recall (R) and F1-score (F1) are calculated in each model, as well as the macro-average of each one of these metrics. The election of the macro-average instead of the micro-average is because macro-averaged give equal weights to the scores of different classes and it does not give larger classes more weight in the calculation of the average. In the calculation of average scores, the [NEUTRAL](#) class is included since we consider important that the classifier be able to distinguish between emotional and non-emotional content. This evaluation allows us to analyze the results obtained by Machine Learning ([ML](#)) algorithms when an emotion corpus annotated automatically is employed.

Concerning agreement evaluation, the Inter-Annotator Agreement ([IAA](#)) between the automatic annotation and the gold standard of each corpus is measured with [Cohen \(1960\)](#)'s kappa metric, one of the most popular metrics employed to compare the extent of consensus between annotators in classifying items. This assessment indicates us how well our process annotates since the automatic annotations are directly compared to the gold

standard of each corpus. If there is a disagreement between automatic and manual annotations, this indicates that it has been mistakes of the creation of the seed and thus there are incorrect associations between sentences and emotions.

4.2.3 Results

We first list the results obtained by each classifier trained on the gold standard corpora manually annotated with same algorithms, set of features and evaluation (10-fold cross-validation) in Table 4.6.

Table 4.6: Results for the SMO multi-classifier on the gold standard of Aman corpus and the six SMO binary-classifiers on the gold standard of Affective Text corpus. Precision, recall, F1-score per class and their macro-average scores.

	Aman Corpus			Affective Text Corpus		
	P	R	F1	P	R	F1
Anger	0.538	0.274	0.363	0.946	0.962	0.953
Disgust	0.714	0.320	0.442	0.986	0.988	0.985
Fear	0.672	0.357	0.466	0.876	0.902	0.881
Joy	0.720	0.513	0.599	0.843	0.879	0.850
Sadness	0.577	0.260	0.359	0.904	0.913	0.897
Surprise	0.553	0.226	0.321	0.967	0.970	0.962
Neutral	0.798	0.955	0.869	-	-	-
Macro-avg.	0.653	0.415	0.488	0.920	0.936	0.921

As for the results achieved by each classifier in all of our approaches of corpora annotated automatically, they are shown in the tables below. Tables 4.7 and 4.8 detail results obtained with all of the DSMs on Aman corpus and Tables 4.9 and 4.10 show the results on Affective Text corpus. Precision (P), recall (R) and F1-score (F1) are shown per class, as well as their macro-average scores in the *original* approach and the *enriched* ones.

With respect to the comparison between automatic and manual annotations, the IAA in terms of Cohen (1960)'s kappa achieved by each one of our approaches when they are compared to the gold standard of both corpora are shown in Tables 4.11 - 4.12.

Table 4.7: Results for the SMO multi-classifier trained on the corpus developed applying LSA and ukWak W2V (CBOW) models on Aman corpus. Precision, recall, F1-score per class and their macro-average scores.

	LSA model (Aman corpus)									ukWak W2V (CBOW) (Aman corpus)								
	Original			Enriched WN			Enriched Oxford			Original			Enriched WN			Enriched Oxford		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.198	0.137	0.162	0.444	0.348	0.391	0.338	0.330	0.334	0.184	0.152	0.167	0.413	0.330	0.367	0.360	0.356	0.358
Disgust	0.250	0.068	0.107	0.308	0.178	0.225	0.353	0.120	0.179	0.121	0.047	0.067	0.350	0.286	0.315	0.200	0.098	0.132
Fear	0.401	0.236	0.297	0.392	0.303	0.342	0.412	0.251	0.312	0.289	0.179	0.221	0.409	0.282	0.334	0.336	0.219	0.265
Joy	0.574	0.571	0.572	0.677	0.702	0.689	0.565	0.604	0.584	0.507	0.586	0.544	0.680	0.796	0.733	0.520	0.600	0.557
Sadness	0.247	0.107	0.149	0.467	0.269	0.341	0.591	0.462	0.519	0.307	0.226	0.260	0.406	0.241	0.303	0.552	0.586	0.568
Surprise	0.459	0.224	0.301	0.366	0.152	0.214	0.359	0.192	0.250	0.345	0.185	0.241	0.294	0.103	0.153	0.376	0.229	0.285
Neutral	0.706	0.846	0.770	0.559	0.676	0.612	0.551	0.668	0.604	0.608	0.702	0.652	0.587	0.573	0.580	0.596	0.554	0.574
Macro-avg.	0.405	0.313	0.337	0.459	0.375	0.402	0.453	0.375	0.397	0.337	0.297	0.307	0.448	0.373	0.398	0.420	0.377	0.391

Table 4.8: Results for the SMO multi-classifier on the corpus developed applying Gigaword W2V (CBOW & SKIP) models on Aman corpus. Precision, recall, F1-score per class and their macro-average scores.

	Gigaword W2V (CBOW) (Aman corpus)									Gigaword W2V (SKIP) (Aman corpus)								
	Original			Enriched WN			Enriched Oxford			Original			Enriched WN			Enriched Oxford		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.113	0.074	0.089	0.541	0.400	0.460	0.399	0.385	0.392	0.139	0.094	0.112	0.465	0.354	0.402	0.383	0.385	0.384
Disgust	0.250	0.052	0.086	0.262	0.129	0.173	0.235	0.080	0.119	0.176	0.037	0.061	0.365	0.256	0.301	0.160	0.073	0.100
Fear	0.419	0.233	0.300	0.387	0.287	0.329	0.300	0.172	0.219	0.336	0.223	0.268	0.388	0.286	0.329	0.257	0.167	0.203
Joy	0.554	0.423	0.480	0.674	0.706	0.690	0.550	0.556	0.553	0.528	0.597	0.560	0.688	0.748	0.717	0.557	0.609	0.582
Sadness	0.305	0.105	0.157	0.496	0.298	0.372	0.554	0.459	0.502	0.273	0.143	0.188	0.435	0.213	0.286	0.544	0.417	0.472
Surprise	0.407	0.222	0.287	0.406	0.160	0.230	0.338	0.150	0.208	0.353	0.156	0.217	0.250	0.092	0.134	0.359	0.168	0.229
Neutral	0.719	0.876	0.790	0.591	0.679	0.632	0.540	0.648	0.589	0.629	0.751	0.685	0.538	0.636	0.583	0.485	0.595	0.534
Macro-avg.	0.395	0.284	0.313	0.480	0.380	0.412	0.417	0.350	0.369	0.348	0.286	0.299	0.447	0.369	0.393	0.392	0.345	0.358

Table 4.9: Results for the SMO six binary-classifiers on the corpus developed applying LSA and ukWak W2V (CBOW) models on Affective Text Corpus. Precision, recall, F1-score per class and their macro-average scores.

	LSA model (Affective Text corpus)									ukWak W2V (CBOW) (Affective Text corpus)								
	Original			Enriched WN			Enriched Oxford			Original			Enriched WN			Enriched Oxford		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.722	0.737	0.720	0.705	0.705	0.701	0.769	0.763	0.760	0.754	0.765	0.746	0.722	0.720	0.716	0.772	0.766	0.763
Disgust	0.870	0.883	0.855	0.830	0.831	0.808	0.803	0.824	0.797	0.875	0.884	0.853	0.827	0.829	0.806	0.799	0.822	0.792
Fear	0.764	0.762	0.754	0.672	0.671	0.671	0.726	0.724	0.722	0.768	0.768	0.761	0.697	0.696	0.696	0.753	0.749	0.747
Joy	0.815	0.826	0.803	0.769	0.771	0.759	0.717	0.722	0.712	0.823	0.835	0.813	0.769	0.772	0.758	0.745	0.749	0.739
Sadness	0.778	0.789	0.774	0.730	0.737	0.727	0.755	0.757	0.748	0.805	0.812	0.793	0.738	0.744	0.732	0.756	0.756	0.745
Surprise	0.851	0.852	0.819	0.822	0.823	0.801	0.799	0.811	0.790	0.853	0.857	0.826	0.817	0.823	0.805	0.802	0.813	0.791
Macro-avg.	0.800	0.808	0.788	0.755	0.756	0.745	0.762	0.767	0.755	0.813	0.820	0.799	0.762	0.764	0.752	0.771	0.776	0.763

Table 4.10: Results for the SMO six binary-classifiers on the corpus developed applying Gigaword W2V (CBOW & SKIP) models on Affective Text Corpus. Precision, recall, F1-score per class and their macro-average scores.

	Gigaword W2V (CBOW) (Affective Text corpus)									Gigaword W2V (SKIP) (Affective Text corpus)								
	Original			Enriched WN			Enriched Oxford			Original			Enriched WN			Enriched Oxford		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.765	0.775	0.755	0.714	0.713	0.708	0.777	0.772	0.769	0.738	0.748	0.730	0.697	0.696	0.693	0.767	0.761	0.759
Disgust	0.884	0.889	0.862	0.832	0.834	0.813	0.804	0.825	0.797	0.856	0.868	0.834	0.818	0.822	0.801	0.792	0.815	0.784
Fear	0.782	0.782	0.775	0.692	0.691	0.690	0.726	0.724	0.722	0.770	0.768	0.762	0.694	0.693	0.693	0.719	0.716	0.715
Joy	0.824	0.837	0.815	0.772	0.774	0.761	0.724	0.729	0.718	0.813	0.826	0.804	0.771	0.771	0.757	0.728	0.732	0.722
Sadness	0.805	0.812	0.795	0.729	0.737	0.725	0.757	0.758	0.748	0.788	0.796	0.780	0.723	0.729	0.717	0.737	0.740	0.732
Surprise	0.868	0.868	0.838	0.812	0.819	0.799	0.814	0.822	0.799	0.854	0.857	0.824	0.824	0.828	0.813	0.804	0.815	0.792
Macro-avg.	0.821	0.827	0.807	0.759	0.761	0.749	0.767	0.772	0.759	0.803	0.811	0.789	0.755	0.757	0.746	0.758	0.763	0.751

Table 4.11: IAA in terms of Cohen’s kappa on the comparison of the annotation of the *Original* and *Enriched* approaches to the gold standard of Aman corpus.

	Cohen's kappa values (Aman corpus)											
	LSA			ukWak W2V (CBOW)			Gigaword W2V (CBOW)			Gigaword W2V (SKIP)		
	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford
Anger	0.9368	0.9051	0.8882	0.9193	0.9004	0.8713	0.9430	0.9089	0.8875	0.9328	0.9044	0.8675
Disgust	0.9495	0.9417	0.9537	0.9452	0.9392	0.9529	0.9507	0.9430	0.9527	0.9460	0.9412	0.9514
Fear	0.9226	0.8919	0.9323	0.9315	0.9099	0.9328	0.9380	0.9136	0.9343	0.9106	0.9009	0.9223
Joy	0.7719	0.6041	0.7241	0.6987	0.5359	0.7219	0.8053	0.6414	0.7443	0.7281	0.5752	0.6942
Sadness	0.9285	0.9193	0.8033	0.8750	0.9119	0.7425	0.9340	0.9173	0.8396	0.9131	0.9066	0.8230
Surprise	0.9186	0.9512	0.9345	0.9014	0.9522	0.9338	0.9368	0.9557	0.9325	0.9146	0.9509	0.9295

Table 4.12: IAA in terms of Cohen’s kappa on the comparison of the annotation of the *Original* and *Enriched* approaches to the gold standard of Affective Text Corpus.

	Cohen's kappa values (Affective Text corpus)											
	LSA			ukWak W2V (CBOW)			Gigaword W2V (CBOW)			Gigaword W2V (SKIP)		
	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford	Original	Enriched WN	Enriched Oxford
Anger	0.6896	0.5544	0.5312	0.6976	0.5600	0.5448	0.7024	0.5632	0.5496	0.6800	0.5480	0.5344
Disgust	0.8552	0.7422	0.7888	0.8560	0.7448	0.7904	0.8568	0.7456	0.7896	0.8416	0.7336	0.7832
Fear	0.6576	0.5476	0.5792	0.6696	0.5504	0.5832	0.6712	0.5520	0.5848	0.6520	0.5464	0.5752
Joy	0.7704	0.6902	0.6456	0.7752	0.6928	0.6488	0.7792	0.6928	0.6480	0.7688	0.6856	0.6392
Sadness	0.7576	0.6693	0.6576	0.7712	0.6752	0.6616	0.7720	0.6768	0.6624	0.7544	0.6664	0.6520
Surprise	0.7856	0.7182	0.7328	0.7912	0.7208	0.7352	0.7976	0.7224	0.7392	0.7904	0.7168	0.7352

4.2.4 Analysis

4.2.4.1 Aman corpus

As for the **SMO** classifier results (Tables 4.7 and 4.8), the systems perform reasonably well since the macro-average F1-scores achieved by all approaches are near 40%, obtaining the best value (41.2%) in Gigaword **W2V (CBOW)**. Even though these scores do not perform above the results of the original Aman corpus (48.8%), they are remarkable taking into account that they have been achieved by corpora automatically labeled with an unsupervised technique.

About the agreement evaluation, the results indicate "perfect agreement" according to **Landis and Koch (1977)** since most of the approaches reach values higher 80%. However, this agreement is "substantial" for **JOY** emotion because their values are between 60% and 80%. This could indicate that the process of the creation of the seed introduces false **JOY** sentences. The possible causes of this problem could be:

1. The fact that linguistic phenomena such as negation or irony have not been addressed.
2. The use of a general domain lexicon where "generic" words like *child*, *found*, *clean*, etc. are associated with **JOY** emotion.

All this leads to the number of sentences annotated with **JOY** is higher than they should be and the agreement is worse.

Focusing on the comparative between *Original* and *enriched* approaches (*Enriched WN*, *Enriched Oxford*), this must be done per emotion since there are different situations:

- **ANGER**, **FEAR**, and **SADNESS**. They obtain improvements in *enriched* approaches in F1-scores. Respect to agreement values in these emotions, the best performs are showed in *Original* approach in the majority of these emotions, however the scores achieved by *enriched* ones are higher 80%, thus the quality of the annotations is high.
- **DISGUST**. It is an emotion where the improvements in F1-score in *enriched* approaches is also shown in agreement values.

- **SURPRISE.** It is one of the emotions more complicated to classify, however, the *Enriched Oxford* approach in ukWak **W2V (CBOW)** and Gigaword **W2V (SKIP)** models perform best with respect to *Original* ones. Moreover, the agreement values in these models remain higher than 90%.
- **JOY.** In this case, the best F1-scores are achieved by *enriched approach*. However, the agreement values in these approaches are around 60% ("substantial agreement"). The possible false JOY sentences, previously mentioned, impact positively in the F1-scores since the algorithm tend to classify by the most frequent class and negatively in the agreement evaluation.

In general, the results show benefits in *enriched* approaches and thus demonstrate the usability of extending the seed in Aman corpus, since most of the best F1-scores have been achieved by enriched ones and their agreement values remain high. However, the use of **WN** or Oxford synonyms should be analyzed in depth since the results vary depending on each emotion. Therefore, it is not possible to conclude which is the best resource for extending the seed, as well as the influence of extending **EmoLex** by the most frequent sense or by all senses.

Finally, regarding the **DSMs** employed, in the F1-scores, as well as the agreement, there are no significant differences between the models that allow us to conclude that one model is better than the rest.

4.2.4.2 Affective Text corpus

In terms of the classification results (Tables 4.9 and 4.10), the systems perform well since the macro-average F1-scores are around 80%, obtaining the best value (80.7%) in GigaWord **W2V (CBOW)**. As in Aman corpus, these results do not outperform the results of the original Affective Text corpus (92.1%), however, our proposals provide remarkable benefits in terms of cost and time.

With reference to the agreement evaluation (Table 4.12), the results show a "substantial agreement" according to **Landis and Koch (1977)** since the scores are between 65% and 85%. Unlike Aman corpus, in this corpus, the worst values are achieved by **FEAR** emotion, therefore, it could indicate that

the process of the creation of the seed introduces false FEAR sentences. This could be due to the fact that the coverage of FEAR words in EmoLex is the highest respect to other emotions as shown Table 4.1.

Focusing on the comparative between *Original* and *enriched* approaches (*Enriched WN*, *Enriched Oxford*), the situation is different to Aman corpus. In this case, the results only show best F1-scores in *enriched* approaches for ANGER emotion. However, these improvements are not reflected in agreement evaluation. Therefore, in this corpus, the extension of EmoLex is not recommended since the *Original* approach perform best for the majority of the emotions in both evaluations: classification and agreement. This allow us to conclude that the use of the resources to extend EmoLex and the election of these resources would depend on the genre of text to annotate.

Finally, as on Aman corpus, regarding the DSMs employed, in the F1-scores, as well as the agreement, there are no significant differences between the models that allow us to conclude that one model is better than the rest.

4.3 Conclusion

As presented in the introductory section, the rationale beyond our research is the need to tackle the annotation task of emotions automatically due to the cost and time associated with the manual annotation process.

In this chapter, we presented a bootstrapping technique for IL for emotion annotation with two main steps: 1) *an initial similarity-based categorization* where a set of seed sentences is created and this seed is extended by the semantic similarity; 2) *train an (extended) supervised classifier on the initially categorized set with one or more iterations*.

As first step, our proposal is described with the explanation of each phase of the methodology and giving concrete examples. After that, we described how the methodology is assessed and which the datasets employed.

Finally, the last part of this chapter has been dedicated to the analysis of the results obtained in each corpus. This analysis allows us to verify the appropriateness and reliability of our approach and obtain the following main conclusions:

1. The viability of [IL](#) bootstrapping technique to automatically label emotion corpora reducing the cost and time-consuming is demonstrated, since the classification and agreement evaluation performed on both corpora achieved promising results with high benefits in terms of cost and time.
2. The results do not allow us to conclude which [DSMs](#) is better for extending the seed since there are no significant differences between the models. Thus, we can conclude that the step 1.2 of the process is independent of the [DSMs](#) employed, providing flexibility to our proposal.
3. About the use of [EmoLex](#), the results have been satisfactory taking into account it is a general domain resource and it has been applied in two different genres: headlines and blog posts. However, in order to improve the results, it would be recommendable to employ domain-depended resources.
4. The improvement of *enriched* approaches has been demonstrated for several emotions in Aman corpus, thus the process of extension could be beneficial depending on the genre of text analyzed. Hence, the usability of these approaches will be analyzed in depth in future works.

These encouraging results allow us to confirm how automatic processes can improve the challenging task of emotion annotation in text. Hence, in our second proposal to efficiently tackle this task, we assess EmoLabel: a semi-automatic methodology where an automatic process is included in order to help human annotators. This proposal is presented in the next chapter.

EmoLabel: Semi-Automatic Methodology for Emotion Annotation

“Do not fear mistakes. There are none.”

MILES DAVIS

This chapter provides our second proposal to efficiently tackle emotion annotation task in a text. We present EmoLabel, a semi-automatic methodology based on an automatic pre-annotation process. Its building up consists of two main phases shown in Figure 5.1:

Phase 1 *Pre-annotation Process.* This step is approached by applying an automatic process to annotate the unlabelled sentences with a reduced number of emotion categories.

Phase 2 *Manual Refinement.* The output of phase 1 is examined in a manual refinement process where human annotators determine which are the emotion/s associated with each sentence. In our proposal, this phase has as objective the detection of the dominant emotion between the pre-defined set of possibilities.

By means of proposing innovation in terms of annotation methodol-

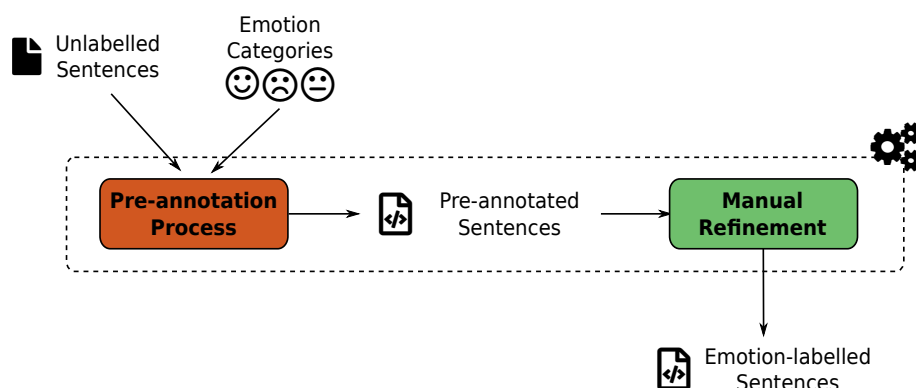


Figure 5.1: Overview of EmoLabel methodology.

ogy, our aim is to automatically pre-annotate those emotion categories that are more related to each sentence, since the number of coding categories influences reliability estimation. As [Antoine et al. \(2014\)](#) concluded, annotation agreement increases significantly when the number of classes decreases. Hence, our hypothesis is that suggesting a reduced number of categories could help to human annotators in their decision on which is the dominant emotion in the second phase of EmoLabel and this will improve its reliability.

Thus, our purpose is to assess the usability of EmoLabel to build up emotion corpora since our objective is the development of efficient techniques able to build up large emotion corpora in different genres with high quality and high-reliability standards.

In this chapter, we will detail the entire process, starting with a complete description of each phase of EmoLabel, following up with the evaluation performed, and finally the conclusions drawn from this experimentation.

5.1 Phase 1: Pre-annotation Process

This section describes the first phase of EmoLabel: the pre-annotation process where the number of emotion categories is automatically reduced. We have compared two pre-annotation processes: an *unsupervised* approach based on Distributional Semantic Models (DSMs) and a *supervised* method based on Machine Learning (ML), explained in Section 5.1.1 and 5.1.2, respectively.

As input data, both processes receive a collection of unlabelled sentences and a set of emotion categories (e.g. Ekman (1992)’s basic emotions, Plutchik (1962)’s wheel of emotions or Izard (1971)’s emotions) as shown Figure 5.2. This adaptability of EmoLabel allows the use of the processes proposed in different domains or application where the set of emotion categories is different. For instance, in the educational domain where the emotions typically detected are BOREDOM, ANXIETY, and EXCITEMENT (Kim, 2011), or in news domains where emotions such as AMUSED or INSPIRED¹ are frequently analyzed.

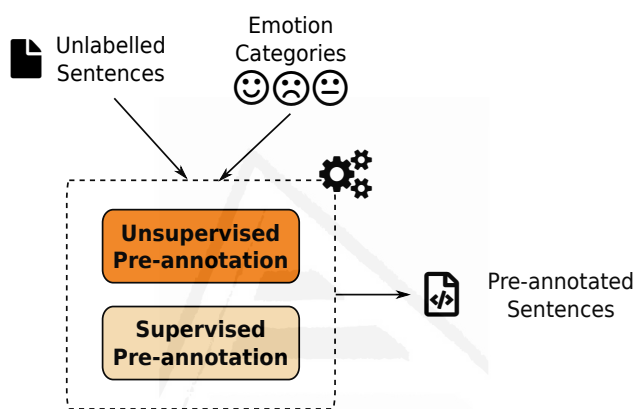


Figure 5.2: Overview of pre-annotation process (Phase 1).

Concretely, Ekman (1992)’s basic emotions are the set of emotions chosen for performing the evaluation of our proposal since it is the most popular set of emotions in computational approaches and is the set most widely accepted by different researchers (Table 2.1).

5.1.1 Unsupervised Pre-annotation

The human intervention in an *unsupervised* approach is minimum and thus it is an interesting proposal for emotion annotation since the influence of everyone’s personal context in the emotional interpretation is reduced. Therefore, the creation of an *unsupervised* pre-annotation process has been considered suitable for our proposal.

Given the encouraging results obtained in the previous chapter (Section 4.2.3), in this approach, we have considered relevant the use of distributed

¹<http://www.rappler.com/>

representations of emotions and the sentences to develop the process. Hence, the *unsupervised* approach is based on Distributional Semantic Models (DSMs).

As was introduced in Chapter 4 (Section 4.1.2), DSMs are based on the assumption that the meaning of a word can be inferred from its usage. Therefore, these models dynamically build semantic representations (high-dimensional semantic vector spaces) through a statistical analysis of the contexts in which words occur (Lapesa & Evert, 2014). Finally, each word is represented by a real-valued vector called *word vector* or *word embedding* whose geometric properties prove to be semantically and syntactically meaningful (Mikolov et al., 2013; Pennington et al., 2014). Thus, words that are semantically and syntactically similar tend to be close in the semantic space.

A big advantage of using these representations that encode semantic information is that they can be generated from large corpora of unlabelled text, and can be trained on very large corpora in a reasonable amount of time. Thus, it is a simple way to filter the number of emotion categories that can be associated with each sentence and reduce the ambiguity of the second phase of EmoLabel.

The process consists of two main steps shows in Figure 5.3: the representation of emotion categories and sentences in a semantic space (Step 1) and the association between emotions and sentences (Step 2).

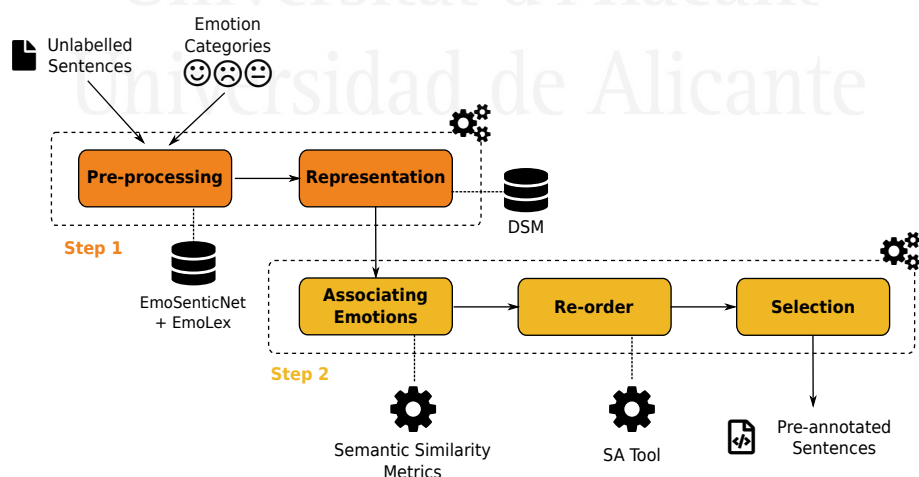


Figure 5.3: Overview of *unsupervised* pre-annotation process.

Step 1: Emotion Categories and Sentences in Semantic Space

The first step towards data annotation consists in encoding the emotions and the sentences in a semantic space with the help of distributed representations. This step is split into two main sub-steps shown in Figure 5.3:

- **Step 1.1 *Pre-processing***: for emotions, it consists in building up a bag of words related to each emotion by exploring an emotion lexicon and adding those words associated with only one of the Ekman (1992)'s basic emotions to create an accurate seed without ambiguous words. While the *pre-processing* of sentences consists of tokenizing and lemmatizing each sentence and build up a bag of words from the lemmas. A graphical representation of this step is shown in Figure 5.4.
- **Step 1.2 *Representation***: it consists in creating emotion vectors and sentence vectors by replacing each word in every bag of words with its vector representation. Following this, for each emotion and sentence, a single vector is obtained by applying averaging as a compositional function. The DSMs employed are detailed in Table 5.2.

In terms of the emotion lexicon employed in the *pre-processing*, the approach proposed employs a union of two emotion lexicons (*EmoSenticNet* + *EmoLex*) previously presented in Section 2.2 (Table 2.3):

- *EmoSenticNet* (ESN) (Poria et al., 2013): is a lexical resource of 13,189 words that automatically assigns qualitative emotions label and quantitative polarity scores to SenticNet concepts (Cambria et al., 2014). Ekman (1992)'s emotions: ANGER, FEAR, DISGUST, SADNESS, SURPRISE, or JOY is the set of emotions employed for labelling the concepts.
- *NRC Emotion Lexicon* (EmoLex) (Mohammad & Turney, 2013): is a lexicon of general domain consisting of 14,000 English words manually compiled and associated with the Plutchik (1980)'s eight basic emotions and two sentiments: POSITIVE and NEGATIVE. The fact that our proposal employs Ekman's emotions implies that the lexicon is reduced to 3,462 English words. The coverage of this reduced version of EmoLex is shown in Table 4.1.

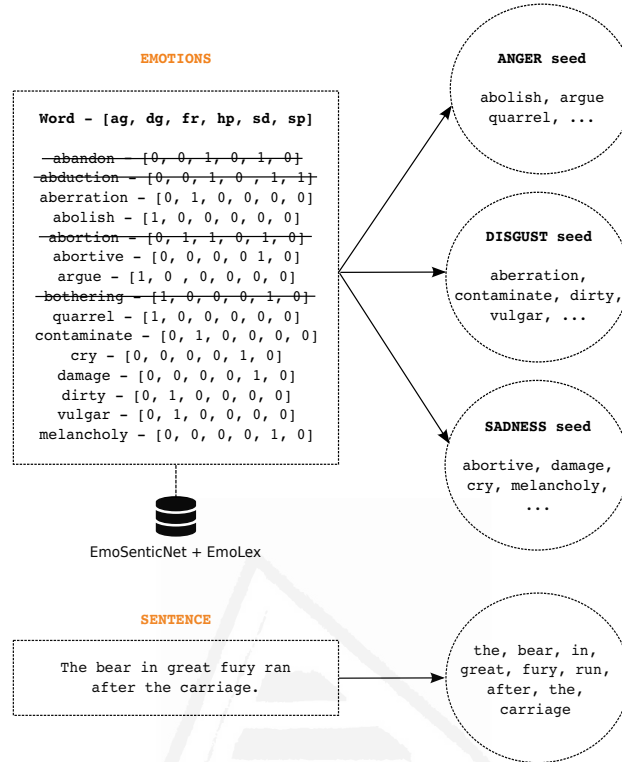


Figure 5.4: Graphical representation of the pre-processing step (1.1).

In both resources, each word has an emotion vector associated where each position represents an emotion: [ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE]. For the union of the emotion lexicons, if a word is stored in both lexicons, the word will be associated with the emotions in common. For instance, the word 'sterile' has the vector [0, 1, 1, 0, 1, 0] associated in [ESN](#) and the vector [0, 0, 0, 0, 1, 0] associated in [EmoLex](#). Considering both vectors, the resultant vector will be [0, 0, 0, 0, 1, 0]. The distribution of this resource is shown in Table 5.1.

Table 5.1: Distribution of the emotion words annotated with only one emotion in the resultant lexicon (*EmoSenticNet + EmoLex*)

Anger	Disgust	Fear	Joy	Sadness	Surprise	Total
120	168	185	1357	386	65	2281

The fact that the emotion lexicon was the result of combining two emotion resources allows us to employ a lexicon more precise since when a word is in the two lexicons, the emotion associated is verified in both lexica.

Regarding the **DSMs** employed in *Representation* step, our approaches have been evaluated using four semantic spaces (Table 5.2):

- *Vector Space Model (baseline)*: a simple semantic space is built by a Vector Space Model (**VSM**) created with *EmoSenticNet+Emolex*, the emotion lexicon explained above (Table 5.1). In this space, the emotions and sentences are represented by a vector that contains information about which *EmoSenticNet+Emolex* words occur in each sentence or emotion.
- *Affective Space* (**Cambria et al., 2015**): this set is the 100-dimensional vector space representation of AffectNet (a matrix of affective commonsense knowledge in which common-sense concepts are linked to semantic and affective features).
- *GloVe vectors* (**Pennington et al., 2014**): here, two set of vectors are employed depending on the test corpus. For Aman corpus, the 300-dimension GloVe vectors trained on 42 billion tokens of web data from Common Crawl² are applied. And for EmoTweet-5, the 200-dimension Glove vectors trained on 2 billion tweets (27 billion tokens) is used.
- *Ultradense Sentiment Analysis Word Embeddings* (**Rothe et al., 2016**): these pre-trained embeddings are the results of learning an orthogonal transformation of the embedding space that focuses the information relevant for a task. For this evaluation, two sets of ultradense vectors are employed. Specifically, for Aman corpus, the 300-dimension Google News vectors are applied. And for EmoTweet-5, the 400-dimension embeddings on a Twitter corpus of size 5.4 billion of tweets. Both sets of vectors are focused on Sentiment Analysis.

Step 2: Associating Sentences with Emotions

Once the emotions and the unlabelled sentences are represented by distributional vectors, the next step consists in associating the sentences with the emotions. This is carried out in three steps shown in Figure 5.3:

²<http://commoncrawl.org/>

Table 5.2: DSMs features for EmoLabel approaches

DSMs	Features	Authors
VSM (baseline)	<ul style="list-style-type: none"> source: <i>EmoSenticNet+Emolex</i> size: 2281-dimensional 	our baseline
Affective Space	<ul style="list-style-type: none"> source: AffectNet size: 100-dimensional 	(Cambria et al., 2015)
GloVe vectors	<ul style="list-style-type: none"> source: web data from Common Crawl size: 300-dimensional 	(Pennington et al., 2014)
	<ul style="list-style-type: none"> source: Twitter Messages size: 200-dimensional 	
Ultradense SA Word Embeddings	<ul style="list-style-type: none"> source: Google News size: 300-dimensional 	(Rothe et al., 2016)
	<ul style="list-style-type: none"> source: Twitter Messages size: 400-dimensional 	

- **Step 1** *Associating Emotions-Sentences*: because all emotions and sentences are created using the same distributed vectors and compositional function, the vector space in which they are placed is also comparable. Hence, in this step, a first emotional ranking for each sentence is proposed by measuring the cosine distance between emotions and sentences.
- **Step 2** *Re-order*: the order of the emotions proposed by the system in the previous step is re-ordered according to the polarity and subjectivity values of each sentence because, as we conclude in our preliminary work (Canales et al., 2017), this information is useful in the pre-annotation process. For that, the Sentiment Analysis (SA) tool from Pattern (De Smedt & Daelemans, 2012) is employed, which returns an averaged (polarity, subjectivity) tuple for a given string.
- **Step 3** *Selection*: in this step, the pre-annotated emotions are finally chosen. The system selects the first three emotions of the resultant ranking of the previous step. Concretely, the process pre-annotates with three emotions because it is half of the number of Ekman (1992)'s

basic emotions. This criterion was empirically determined, showing that the annotation of the half of emotion categories obtained a suitable balance between the reduction of the number of categories and the accuracy of the pre-annotation process. If the process would work with a greater or less group of emotion categories, the number of emotions pre-annotated would be increased or reduced respectively.

In the second sub-step (*re-order*), about the classification of the Ekman (1992)'s six basic emotions according to the polarity, we assume that JOY belongs to the positive class, while the other five emotions have negative polarity, except for SURPRISE since it can be employed from the positive and negative point of view. Hence, when SURPRISE is the first emotion proposed by the system and the subjective value is not zero, the polarity information is employed to re-order the rest of the emotions.

The re-ordering of the emotions is carried out considering the following conditions:

- If the subjective value is zero, the sentence will be considered NEUTRAL and thus this category is proposed in the first place. Alternatively, the polarity value will evaluate it.
- If the polarity value is POSITIVE (higher than zero), the emotion considered positive (JOY) is proposed in the first position.
- If the polarity value is NEGATIVE (less than zero), the emotions considered negatives (ANGER, DISGUST, FEAR, SADNESS) are proposed before the positive ones. The order between these emotions is determined by the semantic similarity obtained when emotion word vectors are compared to sentence vector.

Table 5.3 shows examples of how the polarity and subjectivity information is employed in the pre-annotation process and the emotion proposed by the system for each sentence.

5.1.2 *Supervised Pre-annotation*

As previously mentioned in Section 1.1, there has been applied a wide variety of Natural Language Processing (NLP) techniques to tackle the

Table 5.3: Examples of *unsupervised* pre-annotation process. The *1st ranking* column shows the order proposed by the system before employing the polarity and subjective information. The *Emotion proposed* column shows the pre-annotated emotions by the system after re-ordering the first ranking.

Sentence	1st ranking	Polarity	Subjectivity	Emotions proposed
This was the best summer I have ever experienced.	joy, disgust, sadness, fear, surprise, anger	0.9	0.6	joy, disgust, sadness
I hate fucking pills.	anger, surprise, fear, disgust, sadness, joy	-0.7	0.85	anger, surprise, fear
Had a lovely birthday yesterday with Alex and Christine.	sadness, joy, disgust, surprise, fear, anger	0.5	0.75	joy, sadness, disgust
I'm becoming a broken toy and now that I have had twelve (I counted) vials of blood drawn, I just feel like I'm completely useless.	joy, sadness, disgust, fear, surprise, anger	-0.15	0.48	sadness, disgust, fear
You don't know their middle name or the age of their sister.	joy, disgust, sadness, fear, surprise, anger	0.0	0.0	neutral, joy, disgust

textual Emotion Recognition (ER) task. However, learning from annotated data (*supervised* learning) leads to better results than learning from raw data (*unsupervised* learning) (Kim, 2011). Thus, the number of emotion recognition systems based on *supervised* approach is higher than *unsupervised* ones. The accuracy of these systems varies from 60%-70% when they try to determine the dominant emotion (Aman & Szpakowicz, 2007; Ghazi et al., 2010; Wang et al., 2012), which indicates that this task is unresolved.

Despite this, these approaches could be employed in emotion annotation for automatically reducing the number of emotion categories as shown Figure 5.5. This is the intention of the methods presented in this section whose

objective is to evaluate the usability of the *supervised* approach in the pre-annotation task.

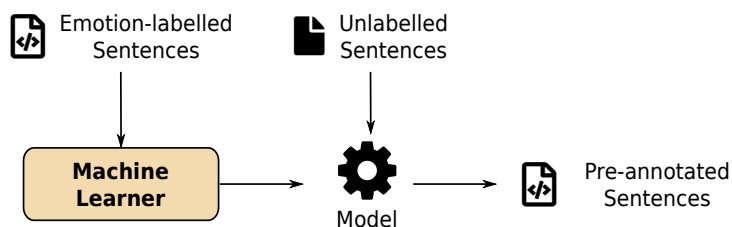


Figure 5.5: Overview of *supervised* pre-annotation process.

With this starting point in mind, three different experiments are performed:

- **Count-Emotion-Words-per-Emotion (*CountWordEmo*):** the first experiment consists in the classification with an 8-feature array where the six first positions represent the number of words associated with each emotion (ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE) and the other two positions contain the polarity and subjectivity values obtained with SA Tool (De Smedt & Daelemans, 2012) for each sentence.
- **Emotion-Lexicon-Words (*EmoLexicon*):** the second experiment consists in the classification with features derived from the emotion lexicon. The features here are the tokens that are common between the lexicon and the chosen dataset.
- **Unigrams (*1-grams*):** this last experiment is a corpus-based classification which uses unigrams. Unigram models have been extensively applied in text classification, and have shown good results in SA classification tasks (Kennedy & Inkpen, 2006).

In *CountWordEmo* and *EmoLexicon* features, the emotion lexicon employed is the union of two emotion lexicons: *EmoSenticNet* + *EmoLex*, the same resource that has been employed in the *unsupervised* pre-annotation process (Section 5.1.1).

As machine-learning algorithm, all experiments apply a Support Vector Machine (SVM) multi-class classifier using the scikit-learn (Pedregosa et al., 2011) package throughout.

5.2 Phase 2: Manual Refinement

Once the unlabeled sentences have been pre-annotated, a manual labeling task is performed by humans annotators with the aim of determining which are the emotion/s associated with each sentence. The number of emotion categories finally annotated will depend on our goals. In our proposal, this phase has as objective the detection of the dominant emotion between the pre-defined set of possibilities.

In order to evaluate the impact of the pre-annotation on the quality of the resulting corpus and on the time needed to annotate, three different experimental setups have been designed (Figure 5.6):

1. **Pre-ML**: in this setup, the best model of the *supervised* pre-annotation (**ML** approach) is used to select the pre-annotated emotions in each sentence.
2. **Pre-WE**: in this setup, the best model of the *unsupervised* pre-annotation (*word embedding* approach) is used to select the pre-annotated emotions proposed to human annotators.
3. **No-pre**: in this setup, no pre-annotation process is employed. Thus, all emotion categories employed are showed to human annotators, as Figure 5.8 shows.

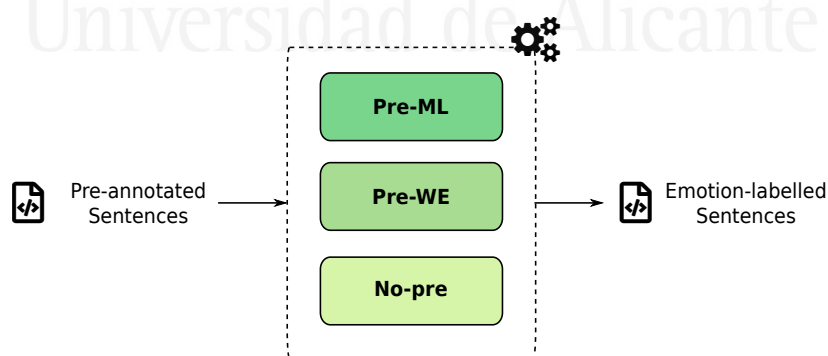
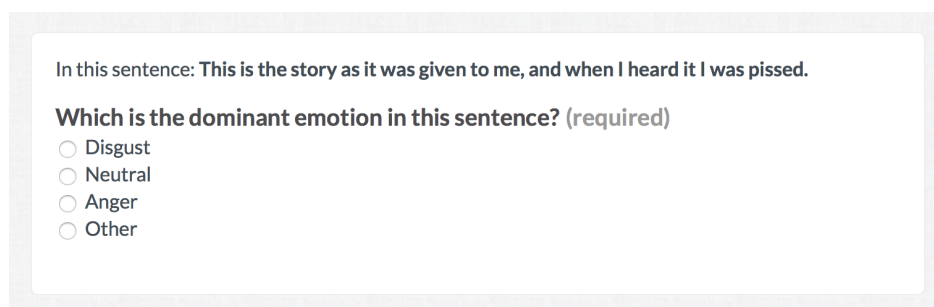


Figure 5.6: Overview of manual refinement (Phase 2).

When the pre-annotation process is employed (*Pre-ML* and *Pre-WE* tasks), the emotions proposed by the system are shown in first place to

humans annotators, who also have the possibility of selecting another emotion (no automatically pre-selected). To do this, they have to choose the option 'Other' and the rest of emotions are displayed, as Figure 5.7 shows.

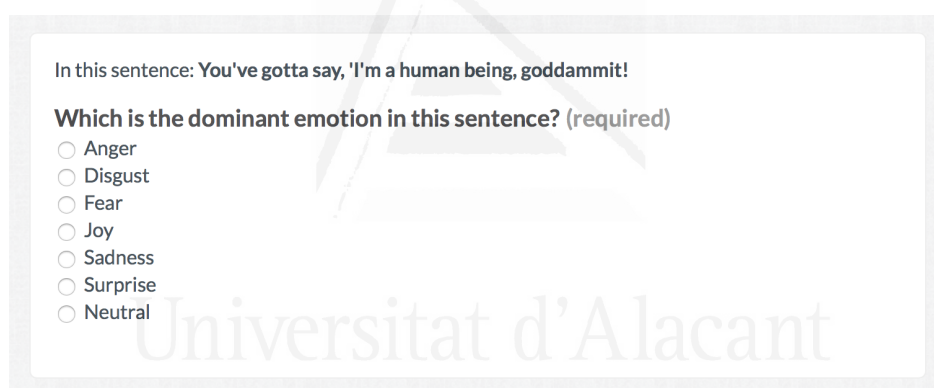


In this sentence: **This is the story as it was given to me, and when I heard it I was pissed.**

Which is the dominant emotion in this sentence? (required)

- Disgust
- Neutral
- Anger
- Other

Figure 5.7: An example of how sentences are shown to the human annotators when the pre-annotation process is employed.



In this sentence: **You've gotta say, 'I'm a human being, goddammit!**

Which is the dominant emotion in this sentence? (required)

- Anger
- Disgust
- Fear
- Joy
- Sadness
- Surprise
- Neutral

Figure 5.8: An example of how sentences are shown to the human annotators when the pre-annotation process is no employed.

All manual annotation tasks were carried out by three annotators with a good knowledge of English language.

In a previous experiment, this phase was designed using three different datasets ($D1$, $D2$, $D3$) for each setup ($Pre-ML$, $Pre-WE$, $NO-pre$). However, we detected that the random selection of the sentences that populate each corpus may negatively or positively affect the results achieved in each setup. Consequently, we decided to apply cross-validation so that the results are not affected by chance bias and the annotators' learning curve. Hence, each annotator carried out three labeling tasks in the order described in Table 5.4. By doing this, each dataset was annotated with all setups and were labeled by

different annotators. For instance, Annotator 1 performed three tasks where *D1* is pre-annotated with ML approach (Pre-ML), *D2* with *unsupervised* pre-annotation (Pre-WE) and *D3* is not pre-annotated (No-pre).

Table 5.4: Cross-validation setup

	<i>D1</i>	<i>D2</i>	<i>D3</i>
Annotator 1	Pre-ML	<u>Pre-WE</u>	<i>No-pre</i>
Annotator 2	<i>No-pre</i>	Pre-ML	<u>Pre-WE</u>
Annotator 3	<u>Pre-WE</u>	<i>No-pre</i>	Pre-ML

Furthermore, due to the difficulty of manual emotion annotation, three training tasks were performed in order to ensure a correct understanding of the task. In each training, the three annotators labeled 21 sentences, three per emotion and three for the NEUTRAL category. After each training, we met for resolving doubts and clarifying aspects related to the annotation guide (Appendix A). Table 5.5 shows the Fleiss (1971)' kappa values reached between the three annotators in each training and Table 5.6 shows the Inter-Annotator Agreement (IAA) reached between each annotator and the Aman corpus' gold standard.

Table 5.5: IAA in terms of Fleiss' kappa between the three annotators in each training task.

Training 1	Training 2	Training 3
0.4512	0.655	0.610

Considering the conclusions drawn from the revision of the state of the art about the benefits of crowdsourcing platforms in manual annotation tasks (Section 2.2.3), we consider Figure Eight (F8)³ (earlier called CrowdFlower (CF)) as the most suitable tool to implement this second phase of EmoLabel.

F8 platform allows accessing an online workplace to clean, label and enrich data. A big advantage of this platform is that there are thousands of

³<https://www.figure-eight.com/>

Table 5.6: IAA in terms of Fleiss' kappa between each annotator and the Aman corpus' gold standard in each training task.

Annotator	Training 1	Training 2	Training 3
1	0.444	0.667	0.833
2	0.556	0.611	0.611
3	0.500	0.722	0.778

people available to read content and score it, with a relatively inexpensive cost. Moreover, F8 offers the possibility of sending the job/task exclusively to your team (using internal contributors option). In this research, we chose to use internal contributors due to the need to control that all tasks were annotated by the same annotators. Although the external contributors were not used, this tool provides us the following advantages: (i) low level of complexity for the creation of the questionnaires and the tasks, (ii) user-friendliness of the application for annotators, and (iii) adaptability of the platform to different types of devices.

5.3 Evaluation

The assessment of EmoLabel requires an intrinsic and extrinsic evaluation. The intrinsic evaluation involves assessing the pre-annotation process whereas the extrinsic one has as objective the evaluation of annotators' performance in the second phase of the methodology.

5.3.1 Data Description

In order to assess the usability for different genres, the approaches are evaluated against two emotion corpora: Aman corpus and EmoTweet-28 corpus, previously introduced in Section 2.3.1 (Table 2.4).

Aman corpus (Aman & Szpakowicz, 2007, 2008). This dataset contains sentence-level annotation of 4,000 sentences from blog posts collected directly from Web. This corpus was manually developed by four annotators who

received no training, though they were given samples of annotated sentences to illustrate the kind of annotations required. It was annotated with the emotion intensity (high, medium, or low) and eight categories: the six emotion categories proposed by Ekman (1992), MIXED EMOTIONS and NO EMOTION. Despite the initial objective of labelling those sentences with more than one emotion (MIXED EMOTIONS), the gold standard is annotated with Ekman’s emotions and NO EMOTION categories. The distribution of the corpus is shown in Table 4.4.

EmoTweet-28 corpus (Liew et al., 2016). This dataset consists of a collection of 15,553 tweets annotated with 28 emotion categories. The corpus contains annotations for four facets of emotion: valence, arousal, emotion category and emotion cues. The Amazon Mechanical Turk (AMT) platform was employed in the manual annotation tasks where each tweet was annotated by at least three annotators. As this research works with Ekman (1992)’s basic emotions, a reduced corpus of EmoTweet-28 is employed (EmoTweet-5). This corpus contains those tweets annotated with ANGER, FEAR, JOY, SADNESS, SURPRISE and the same proportion of NEUTRAL tweets as the original corpus. Finally, EmoTweet-5 comprises 5,931 tweets which distribution per emotion is shown in Table 5.7.

Table 5.7: Distribution of the sentences per emotion on EmoTweet-5, a reduced version of EmoTweet-28 that contains tweets annotated with Ekman’s basic emotions.

Anger	Fear	Joy	Sadness	Surprise	Neutral	Total
986	180	1306	350	179	2,930	5,931

These corpora were chosen for two main reasons: (i) both corpora have been employed in relevant emotion studies as a benchmark (Keshtkar & Inkpen, 2010; Chaffar & Inkpen, 2011; Mohammad, 2012b; Liew et al., 2016); and (ii) it is possible to assess the effectiveness of the pre-annotation process in different social media genres that allows people to post messages to share information, opinions, and emotions: blogs and tweets.

5.3.2 Intrinsic Evaluation

The objective of the intrinsic evaluation is to assess which is the best pre-annotation process that will be employed in the second phase of EmoLabel. To achieve that, the evaluation is carried out comparing the emotions proposed by each method with the gold standard of the test corpora. Specifically, the intrinsic evaluation has been carried out in the two corpora: Aman corpus (Aman & Szpakowicz, 2007) and EmoTweet-5 (Liew et al., 2016).

For the evaluation purpose, the 30% of data of each corpus is employed because in the *supervised* approach the 70% of data is applied for training. Moreover, the use of the same test data for all approaches allows that the results comparability.

As far as the number of emotion categories pre-annotated are concerned, in Aman corpus the sentences will be pre-annotated with three emotions since the corpus is labeled with six categories, while in EmoTweet-5 the sentences will be pre-annotated with two emotions.

Concerning the evaluation methodology, the pre-annotation process is assessed measuring the precision (P), recall (R), and F1-score (F1) of the emotions proposed by our system against the gold standard of the test corpora, as well as the macro-average of each of these metrics for each model. As the process pre-annotates the half of the number of emotion categories, if the correct emotion (the gold standard) is one of the pre-annotated emotions, the prediction will be considered as correct. In the calculation of average scores, the NEUTRAL class is included since we consider important that the pre-annotation process is able to distinguish between emotional and non-emotional content.

5.3.2.1 Unsupervised Pre-annotation

The results of the *unsupervised* pre-annotation process for each DSMs (Table 5.2) are shown in Table 5.8 for Aman corpus (Aman & Szpakowicz, 2007), and Table 5.9 shows the results for EmoTweet-5 (Liew et al., 2016).

The results of the *unsupervised* pre-annotation on Aman corpus shows that considering the macro-average F1-score, all models outperform significantly the baseline. Although, the best result is obtained by *Glove* model due to its

recall and precision values in JOY and SADNESS emotions. From these high values, we may draw that these emotions are frequently found between the emotions proposed by the system. In terms of the rest of the models:

- *Ultradense SA* model. It also obtains high recall values in JOY and SADNESS emotions and moreover, it reaches the best values for F1-measure for ANGER and DISGUST, two of the emotions hard to detect in text.
- *Affective Space*. It is interesting to highlight the results obtained by FEAR and SURPRISE considering that *Affective Space* is a set of 100-dimension vectors and the vocabulary represented in this space is smaller compared to the rest of the models.

Table 5.8: Results for the *unsupervised* pre-annotation using different distributional representations on Aman corpus. Precision, recall, F1-score per class and their macro-average scores.

	<i>Unsupervised Pre-annotation - Aman corpus</i>											
	Baseline			Affective Space			GloVe			Ultradense SA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.35	0.22	0.27	0.17	0.02	0.03	1.00	0.37	0.54	0.58	0.65	0.61
Disgust	0.49	0.38	0.43	0.21	0.85	0.33	1.00	0.23	0.38	0.61	0.54	0.57
Fear	0.34	0.32	0.33	0.84	0.76	0.80	0.96	0.68	0.79	0.55	0.35	0.43
Joy	0.30	0.83	0.44	0.31	0.94	0.47	0.71	0.94	0.81	0.60	0.94	0.73
Sadness	0.40	0.60	0.48	0.65	0.25	0.36	0.87	0.75	0.80	0.62	0.87	0.72
Surprise	0.16	0.18	0.17	0.55	0.62	0.58	0.07	1.00	0.13	0.10	0.97	0.17
Neutral	0.86	0.58	0.69	0.92	0.48	0.63	0.93	0.48	0.63	0.95	0.48	0.63
Macro-avg.	0.42	0.44	0.40	0.52	0.56	0.46	0.79	0.64	0.58	0.57	0.69	0.55

Regarding the results of the *unsupervised* pre-annotation on EmoTweet-5 (Liew et al., 2016), they shows that *GloVe* and *Ultradense SA* outperform significantly the baseline whereas *Affective Space* does not improve it. This can be due to the fact that the vocabulary of *Affective Space* is formal and the language employed in Twitter is more informal, not carefully edited or with grammatical errors. Hence, the results emphasize the importance of using *DSMs* adapted to the genre when the process runs with social media texts. With regard to the rest of the models:

Table 5.9: Results for the *unsupervised* pre-annotation using different distributional representations on EmoTweet-5 corpus. Precision, recall, F1-score per class and their macro-average scores.

	<i>Unsupervised Pre-annotation - EmoTweet-5</i>											
	Baseline			Affective Space			GloVe			Ultradense SA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.41	0.11	0.18	1.00	0.01	0.01	1.00	0.07	0.14	0.77	0.30	0.43
Fear	0.29	0.31	0.30	0.16	0.59	0.25	0.52	0.22	0.31	0.54	0.35	0.43
Joy	0.40	0.92	0.56	0.57	0.94	0.71	0.75	0.96	0.84	0.58	0.96	0.73
Sadness	0.23	0.11	0.15	0.92	0.11	0.20	0.67	0.48	0.56	0.51	0.69	0.59
Surprise	0.09	0.20	0.13	0.06	0.44	0.11	0.05	0.63	0.10	0.07	0.46	0.13
Neutral	0.71	0.45	0.55	0.73	0.44	0.55	0.74	0.44	0.55	0.76	0.44	0.55
Macro-avg.	0.36	0.35	0.31	0.57	0.42	0.31	0.62	0.47	0.42	0.54	0.53	0.48

- *Ultradense SA* model. In this corpus, its recall improvements are confirmed for JOY, SADNESS and SURPRISE, and moreover, the interesting values obtained by ANGER and DISGUST continue to be noted when EmoTweet-5 is employed for the evaluation.
- *GloVe* model. As happens in Aman corpus, in this model the best performs are achieved in JOY and SADNESS emotions for its high recall values.

Furthermore, the results reflect that another important factor in *unsupervised* approach is the coverage of the lexicon employed. For instance, the high coverage of JOY emotion allows to *Affective Space* model achieving good results in this emotion despite this space is not adapted to the genre. However, the low coverage in SURPRISE emotion could explain that the *unsupervised* pre-annotation approach is not able to detect this emotion in this genre since the best F1-value obtained is 13%.

Comparing both evaluations, the results show that the best models are *GloVe* and *Ultradense SA* and therefore the need of using word embeddings with the following features: (1) embeddings built from a large amount of data for representing a large vocabulary; (2) with high dimensionality to codify more semantic features because the better performances; and (3) adapted to the genre of the text that we want to annotate in order to the semantic

space be representative. Moreover, the results highlight the importance of the lexicon coverage of the lexicon employed in the *unsupervised* pre-annotation. Finally, it is interesting to mention the improvements obtained by *Ultradense SA* in ANGER and DISGUST emotions since this enhancement is shown regardless the genre employed.

5.3.2.2 *Supervised Pre-annotation*

As mentioned previously, a multi-classifier SVM is applied using three sets of features: *CountWordEmo*, *EmoLexicon*, and *Unigrams (1-grams)* described in Section 5.1.2. For the evaluation of the *supervised* approach, the datasets are split in 70% for training and 30% for testing. And, the optimal set of hyperparameters for each SVM was determined based on an exhaustive search through the parameter space using 10-fold cross-validation. Using this, the parameters selected in each SVM are described below:

- Aman Corpus
 - *CountWordEmo*: an RBF kernel, C value: 1, gamma value: 0.001
 - *EmoLexicon*: an Linear kernel, C value: 1
 - *Unigrams (1-grams)*: an Linear kernel, C value: 1
- Emo-Tweet-5
 - *CountWordEmo*: an Linear kernel, C value: 10
 - *EmoLexicon*: an RBF kernel, C value: 100, gamma value: 0.001
 - *Unigrams (1-grams)*: an RBF kernel, C value: 100, gamma value: 0.001

The results of the *supervised* pre-annotation process for each set of features are shown in Table 5.10 for Aman corpus, and in Table 5.11 for EmoTweet-5.

The results of the *supervised* pre-annotation on Aman corpus show that considering the macro-average F1-score, the best result is obtained by the *1-grams* model due to the fact that its F1-score is higher than 75% for all the emotions. With respect to *CountWordEmo* and *EmoLexicon*, the results show these models are not able to detect emotions like FEAR and SURPRISE

because these set of features are heavily dependent on the coverage of the lexicon employed.

Table 5.10: Results for the *supervised* pre-annotation using different set of features on Aman corpus. Precision, recall, F1-score per class and their macro-average scores.

	<i>Supervised Pre-annotation - Aman corpus</i>								
	CountWordEmo			EmoLexicon			1-grams		
	P	R	F1	P	R	F1	P	R	F1
Anger	1.00	0.39	0.56	0.95	0.72	0.82	0.90	0.65	0.75
Disgust	1.00	0.40	0.58	1.00	0.21	0.35	0.94	0.65	0.77
Fear	0.00	0.00	0.00	0.75	0.09	0.16	0.92	0.65	0.76
Joy	1.00	0.99	0.99	0.98	0.99	0.98	0.96	0.96	0.96
Sadness	1.00	0.56	0.72	0.88	0.29	0.43	0.97	0.73	0.84
Surprise	1.00	0.09	0.16	1.00	0.09	0.16	1.00	0.65	0.79
Neutral	0.85	1.00	0.92	0.85	1.00	0.92	0.92	1.00	0.96
Macro-avg.	0.84	0.49	0.56	0.92	0.48	0.55	0.95	0.75	0.83

As for the results of the *supervised* pre-annotation on EmoTweet-5, the conclusion is the same as the one for on Aman corpus since the best performance is obtained by *1-grams* and *CountWordEmo* and *EmoLexicon* continue having problems to detect **FEAR** and **SURPRISE**. In general, the results on EmoTweet-5 are worse than on Aman corpus due to the fact that Twitter is a platform where the text with grammatical errors or not carefully edited is more frequently. These worse results are most noticeable in those set of features that are exclusively dependent of the lexicon because its coverage in this genre is low.

Comparing the evaluation in both corpora, the results allow concluding that the set of features employed needs to contain information about the text to be processed and not depend exclusively on an emotion lexicon. Naturally, if the *supervised* emotion approach is improved with an advanced set of features or algorithms, the pre-annotation would improve. However, our aim is to assess the viability of a *supervised* emotion model for pre-annotation, thus a sophisticated feature engineering has not been carried out.

Table 5.11: Results the *supervised* pre-annotation using different set of features on EmoTweet-5 corpus. Precision, recall, F1-score per class and their macro-average scores.

	<i>Supervised Pre-annotation - EmoTweet-5</i>								
	CountWordEmo			EmoLexicon			1-grams		
	P	R	F1	P	R	F1	P	R	F1
Anger	0.77	0.87	0.82	0.69	0.47	0.56	0.81	0.83	0.82
Fear	0.00	0.00	0.00	1.00	0.02	0.04	0.86	0.24	0.38
Joy	0.87	0.66	0.75	0.88	0.87	0.87	0.94	0.87	0.90
Sadness	1.00	0.08	0.15	0.87	0.23	0.37	0.88	0.53	0.66
Surprise	0.00	0.00	0.00	0.75	0.06	0.12	0.73	0.23	0.35
Neutral	0.62	0.99	0.76	0.58	0.93	0.71	0.74	0.97	0.84
Macro-avg.	0.54	0.43	0.41	0.79	0.43	0.45	0.83	0.61	0.66

5.3.3 Extrinsic Evaluation

The extrinsic evaluation has as objective the assessment of the work of the annotators in the second phase of EmoLabel. To achieve that, a manual annotation task is carried out for three annotators.

Aman corpus is the dataset employed to assess this phase. Correctly, the test data (30%) previously used for evaluating the pre-annotation processes. This data is split into three datasets of 100 sentences each one (D_1 , D_2 , D_3) whose distribution per emotion is shown in Table 5.12. The distribution has done in an equitable way with the aim of having the same number of sentence for each emotion.

Table 5.12: Distribution of the number of sentences per emotion annotated in each manual task.

Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral	Total
16	15	11	16	15	11	16	100

As previously introduced in Section 5.2, the manual annotation task is split into several sub-tasks where all datasets are labeled by all annotators using different setups (Table 5.4): (i) Pre-ML setup where the *supervised*

pre-annotation is employed; (ii) Pre-WE setup which uses the *unsupervised* pre-annotation; and (iii) No-pre setup where the pre-annotation method is not applied. Considering the results of intrinsic evaluation on Aman corpus, the best pre-annotation methods are selected for these tasks. Thus, the approach *1-grams* is applied for Pre-ML setup and the *GloVe* model is used for Pre-WE setup.

Regarding the agreement metrics employed, the k -coefficient metrics are well-known in NLP for measuring Inter-Annotator Agreement (IAA) because these are designed for nominal-scaled variables. Thus, the manual annotation process is assessed calculating Fleiss (1971)’s kappa between each annotator and the Aman corpus’ gold standard. The results of the agreement achieved by each annotator in each setup are shown in Table 5.13.

Table 5.13: IAA in terms of Fleiss’ kappa between each annotator and the Aman corpus’ gold standard.

Annotator	Pre-ML	Pre-WE	No-pre	Macro-avg
1	0.588	0.649	0.659	0.632
2	0.637	0.626	0.578	0.614
3	0.635	0.544	0.555	0.578
Macro-avg	0.620	0.606	0.597	

Regarding the agreement evaluation, all tasks reach macro-average scores of 0.60, a "substantial agreement" according to Landis and Koch (1977). Therefore, the results demonstrate that the pre-annotation process does not reduce the IAA or annotation performance. Moreover, it is interesting to mention that two of the three annotators reach their best agreement values in the tasks with Pre-ML, a fact which shows that an accurate pre-annotation process could help human annotators to effectively label emotions.

In terms of time effort, the F8 platform records the time the annotator submitted the judgment and the time at which the annotator started working on the judgment for each page. This allows measuring the time required to complete each task. The macro-average time obtained by each annotator in each setup is shown in Table 5.14.

As for time effort, the macro-average time shows that Pre-ML reduces annotation time by near 20% (19,1%) with respect to the second-best time

(No-pre). And as happen in agreement evaluation, two of the three annotators obtain a time gain of more than 20% (42,6% for Annotator 1 and 23,5% for Annotator 3) when the pre-annotation process is applied (Pre-ML) with respect to their tasks with No-pre. Hence, the evaluation performed demonstrates that the pre-annotation process reduces the annotation time required in emotion labeling.

Table 5.14: Annotation time of each annotator in all manual annotation tasks.

Annotator	Pre-ML	Pre-WE	No-pre	Macro-avg
1	02:01	04:37	03:31	3:53
2	04:09	03:55	03:48	4:00
3	04:27	04:57	05:49	4:00
Macro-avg	03:32	04:29	04:22	

As far as the comparison between the pre-annotation methods, whereas there are no significant differences in terms of agreement values, in time evaluation, Pre-ML reduces annotation time by 24,8% with respect to Pre-WE. This indicates that the use of inaccurate pre-annotation methods may worsen annotation time and thus it does not help human annotators.

Concerning the labeling performed by each annotator, it is of remarkable interest the fact that Annotator 3 reach their best agreement and time scores when the *supervised* pre-annotation process is employed (Pre-ML) since this annotator had some difficulties in understanding the task (Table 5.13). It may, therefore, be concluded that pre-annotation could be used as a strategy to improve the performance of inaccurate annotators. This is an important factor if we want to carry out emotion annotation in crowdsourcing platforms (AMT or F8) with external contributors, since in this kind of tools we cannot know the background of the annotators in detail.

5.4 Conclusions

As presented in the introductory section of this chapter, the rationale behind our research is the need to simplify the emotion annotation task so that to improve its reliability and efficiency.

In this chapter, we presented EmoLabel: a semi-automatic methodology consisting in two phases: (1) an automatic process to pre-annotate the unlabelled sentences with a reduced number of emotion categories; and (2) a manual refinement process where human annotators will determine which is the dominant emotion between the pre-defined set of possibilities. Two pre-annotation strategies are presented: *unsupervised* proposal with the aim of minimizing the human intervention and *supervised* method where simple emotion models are build up, exploiting corpora or models previously developed.

As first step, our proposal is described with the explanation of each phase of the methodology and giving concrete examples. After that, we described how the methodology is assessed and which the datasets employed.

Finally, the last part of this chapter has been dedicated to the analysis of the results achieved in each assessment in terms of intrinsic and extrinsic evaluation. This study allows verifying the appropriateness and reliability of our methodology in emotion annotation and obtaining the following main conclusions:

1. The benefits of pre-annotation processes in emotion labeling are demonstrated since the results on annotation time show a gain of near 20% when the pre-annotation process is applied (Pre-ML) with respect to No-pre. Moreover, the experiments performed show that all tasks reach "substantial agreement" and therefore the pre-annotation process does not reduce the IAA or annotator performance.
2. With respect to the intrinsic evaluation, the gains of the *supervised* pre-annotation method in terms of and time with respect to the *unsupervised* pre-annotation process, allow concluding that the use of this method is more helpful for annotators than the *unsupervised* approach. Consequently, the existing *supervised* emotion detection systems developed so far could be employed to annotate new data.
3. The improvements reached by Annotator 3 (with the lowest performance) (Table 5.13) in terms of time and agreement demonstrate the usability of our methodology with inaccurate annotators, since his best performances are obtained when a pre-annotation process is employed (Pre-ML).

These encouraging results demonstrate that the use of pre-annotation processes provides benefits in the challenging task of textual emotion annotation.



Universitat d'Alacant
Universidad de Alicante

Conclusions and future perspectives

“We can only see a short distance ahead, but we can see plenty there that needs to be done.”

ALAN M. TURING

Due to the need to develop new techniques able to build up large emotion corpora with high quality and high reliability, this work has been focused on one of the most important challenges in textual Emotion Recognition (ER): gathering data with emotion labels. Our main motivation was the difficulties associated with the development of emotion corpora demonstrated by the most relevant research carried out so far. It is true that the problems of creating a labeled corpus such as the time and cost required for its development are shared by other Natural Language Processing (NLP) tasks; however, in textual ER these problems became more challenging because of the detection of emotion in text can be difficult even for humans, producing higher costs and time for its development and problems to obtain good Inter-Annotator Agreement (IAA).

With this context, this dissertation addressed the emotion annotation task, providing automatic and semi-automatic techniques/methodologies with the intention of contributing to efficiently tackle the emotion annotation. Our focus was textual emotion annotation of English data in any genre, at

sentence level and employing a set of distinct emotion categories (categorical emotion models) as labels.

In Chapter 2, we carried out an exhaustive analysis of the state of art with a special focus on the creation of linguistic resources for textual ER. First of all, the emotion theories outlined by psychologists for representing the emotion are described with the aim of properly introducing emotion resources. Then, an extensive review of exiting emotion lexicons and emotion corpora classified according to their creation process (manual or semi/automatic) and their emotion connotation (categorical and dimensional models) are presented. From the exhaustive analysis performed, we were able to conclude that there is an evident lack of semi/automatic methodologies of textual emotion annotation for different genres.

Considering the difficulties in the creation label corpora and the fact that some of these problems are shared by other NLP tasks, a review of the annotation techniques is presented in Chapter 3. This study has as main objective exploring alternative annotation techniques that reduce time and cost of its development to tackle textual emotion labeling. Concretely, this chapter presents a study of two annotation techniques employed in other disciplines which have been applied in this dissertation: bootstrapping technique for Intensional Learning (IL) and the pre-annotation process, given their contribution in other manual annotation tasks.

Firstly, we study the use of a bootstrapping technique for IL as a method for emotion annotation (Chapter 4). This technique was previously applied by Gliozzo et al. (2009) in Text Categorization (TC) and consists of two main steps: 1) *an initial similarity-based categorization*, and 2) *training an (extended) supervised classifier on the initially categorized set with one or more iterations*. It is an unsupervised approach which allows that the influence of emotion interpretation by humans is minimized. Moreover, the technique proposed is adaptable to the set of emotion categories, as well as, to be flexible to the number of emotion categories annotated (the dominant emotion or all of the emotions detected) which makes it a novel proposal.

In addition, in order to evaluate alternative proposals to tackle emotion annotations, this dissertation studies the usability and effectiveness of a semi-automatic methodology to improve the labeling task: EmoLabel (Chapter 5). It consists of two main phases: 1) *an automatic process to pre-annotate*

the unlabelled sentences with a reduced number of emotion categories, and 2) *a refinement manual process* where human annotators will determine which are the emotion/s associated with each sentence. As the [IL](#) proposal, EmoLabel is adaptable to the different sets of emotion labels and it also allows the possibility of annotating one or more emotions in the second phase of EmoLabel, depending on our objective and thus the design of the second phase (manual task). In our proposal, the annotators label the dominant one.

In summary, this dissertation studies two proposals with the aim of efficiently tackling the challenging task of textual emotion annotation. Each one has different features and thus distinct benefits and drawbacks depending on our objectives.

Our main contributions in the form of conclusions (Section [6.1](#)), the list of relevant publications related to this thesis (Section [6.1.1](#)), and the work for the future (Section [6.2](#)) are provided in the next sections.

6.1 Contributions

The main contributions and conclusions gathered from the development of this research work are summarized in the following points:

- **Analysis of the state of art with special focus on the creation of linguistic resources for textual [ER](#)**

This analysis allowed us to verify that the categorical emotion models (particularly, [Ekman \(1992\)](#)'s basic emotions) are the most popular between the computational approaches since the majority of them employs this emotion model due to its simplicity in tackling emotion analysis from the human and computational points of view. Moreover, the chronological analysis of emotion resources allows us to note a tendency of applying semi/automatic techniques for emotion annotation. This is due to two main facts: the disadvantages of manual labeling and the exponential growth in the amount of subjective information on the Web 2.0 (blogs, social networks, microblogging, etc.).

- **Research into Annotation Techniques (Bootstrapping technique for IL and Pre-annotation) for textual ER**

This research studies different efficient methods in terms of time and cost of building up resources. Given the emotion annotation problems of tackling the task in an efficient way and with high reliability, we explore alternative annotation techniques employed in other NLP areas in order to improve textual emotion annotation. Concretely, this research is focused on two methods: the bootstrapping technique for IL and the pre-annotation process. These methods have demonstrated its usability and practicability in other NLP tasks which enable us to consider them suitable to tackle emotion annotation task, obtaining improvements in its development process.

- **Proposal and development of IL technique for textual ER**

A bootstrapping technique for IL is presented with the aim of tackling emotion annotations. It is an unsupervised proposal that builds classifiers from unlabeled data. This is one of the most attractive features of this technique for emotion annotation because it allows us to build up emotion corpora where the influence of human annotators is minimized. Moreover, its simplicity and flexibility to apply to other emotion categories or genres make it an attractive technique to consider when labeled data are lacking and too expensive to be created in large quantities.

- **Proposal and development of Pre-annotation processes for textual ER (EmoLabel)**

EmoLabel is a semi-automatic methodology where an automatic pre-annotation process is carried out with the aim of helping humans to decide the dominant emotion of each sentence. While it is true that this proposal is not as efficient as IL technique in terms of cost and time since requiring the participation of human annotators, we consider that it is important to explore alternative emotion techniques where human annotator participates. After all, we want to identify human emotions. Moreover, EmoLabel provides adaptability and versatility to use other sets of emotion categories and the number of categories associated with each sentence.

- **Evaluation of IL technique**

In order to verify the appropriateness of IL for emotion annotation, two evaluations have been carried out. On the one hand, an emotion model is built from the corpus annotated automatically to evaluate the usability of this corpus. On the other hand, the quality of automatic annotations is assessed through the measure of agreement between the corpus developed with our approach (automatic annotation) and the gold standard of Aman corpus and Affective Text corpus (manual annotation). Both evaluations allow us to verify the viability of IL as a technique to automatically emotion corpora reducing the cost and time-consuming for its development since both evaluations obtain encouraging results.

- **Evaluation of EmoLabel**

With the aim of performing an in-depth assessment of EmoLabel an intrinsic and extrinsic evaluation is required. The objective of the intrinsic evaluation is to assess which is the best pre-annotation process that will be employed in the second phase of EmoLabel. To achieve that, the evaluation is carried out comparing the emotions proposed by each method with the gold standard of the test corpora. The extrinsic evaluation has as objective the assessment of the work of the annotators in the second phase of EmoLabel. To this end, a manual annotation task is carried out for three human annotators. According to the extrinsic evaluation, the experiments performed show the benefits of pre-annotation processes in emotion labeling since the results of annotation time show a gain of near 20% when the pre-annotation process is applied (Pre-ML) with respect to no pre-annotation (No-pre). Moreover, the experiments performed show that all tasks reach "substantial agreement" and therefore the pre-annotation process does not reduce the IAA or annotator performance.

6.1.1 Publications

Part of the contents of this dissertation have been published in several journals and conference events. These publications are now listed showing, in brackets, the chapter to which they are related:

- Canales, L., & Martínez-Barco P. (2014, June) Detección de perfiles de usuario en la Web 2.0 desde el punto de vista emocional. In *Proceedings of V Jornadas TIMM (TIMM14)*. Seville, Spain. [Chapter 1]
- Canales, L., & Martínez-Barco P. (2014, October). Emotion Detection from text: A Survey. In *Proceedings of the 5th Information Systems Research Working Days (JISIC 2014)*. Quito, Ecuador. [Chapter 2]
- Canales, L. (2015, September). Detección de perfiles emocionales de usuario en la Web 2.0. In *Proceedings of the Doctoral Symposium of XXXI edition of the Spanish Society for Natural Language Processing (SEPLN 2015)*. Alicante, Spain. [Chapter 1]
- Canales, L., Strapparava, C., Boldrini, E., & Martínez-Barco P. (2016, May). A Bootstrapping Technique to Annotate Emotional Corpora Automatically. In *Proceedings of the Workshop on Emotion and Sentiment Analysis (ESA - LREC 2016)*. Portorož, Slovenia. [Chapter 4]
- Canales, L., Strapparava, C., Boldrini, E., & Martínez-Barco P. Bootstrapping Technique + Embeddings = Emotional Corpus Annotated Automatically. In *Future and Emerging Trends in Language Technology. Machine Learning and Big Data*. Book chapter. [Chapter 4]
- Canales, L. (2016, September). eMotion: Mejora de la detección de perfiles emocionales de usuario a través de una técnica bootstrapping para la anotación automática de categorías emocionales. In *Proceedings of the Doctoral Symposium of XXXII edition of the Spanish Society for Natural Language Processing (SEPLN 2016)*. Salamanca, Spain. [Chapter 4]
- Canales, L., Strapparava, C., Boldrini, E., & Martínez-Barco P. (2016, October). Exploiting a Bootstrapping Approach for Annotating Emotions in Texts Automatically. In *Proceedings of the 3ed IEEE International Conference on Data Science and Advanced Analytics (DSAA 2016). Special Session on Emotion and Sentiment in Intelligent Systems and Big Social Data Analysis (SentISData)*. Montreal, Canada. [Chapter 4]
- Canales, L., Strapparava, C., Boldrini, E., & Martínez-Barco P. (2016, December). Innovative Semi-Automatic Methodology to Annotate

Emotional Corpora. In *Proceedings of the Workshop PEOPLES: Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (COLING 2016)*. Osaka, Japan. [Chapter 5]

- Canales, L., Daelemans, W., Boldrini, E., & Martínez-Barco P. (2017, September) Towards the Improvement of Automatic Emotion Pre-annotation with Polarity and Subjective Information. In *Proceedings of the 11th biennial Recent Avances in Natural Language Processing Conference (RANLP 2017)*. Varna, Bulgaria. [Chapter 5]
- Canales, L. (2017, September). Metodología semi-automática para la anotación de corpus emocionales. In *Proceedings of the Doctoral Symposium of XXXIII edition of the Spanish Society for Natural Language Processing (SEPLN 2017)*. Murcia, Spain. [Chapter 5]
- Canales, L., Strapparava, C., Boldrini, E., & Martínez-Barco P. (2017). Intensional Learning to Efficiently Build up Automatically Annotated Emotion Corpora. *IEEE Transactions on Affective Computing Journal*. [Chapter 4]
- Canales, L., Daelemans, W., Boldrini, E., & Martínez-Barco P. (2018). EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text. *Journal of Data Mining and Knowledge Discovery*. Submitted [Chapter 5]

6.2 Future work

This thesis has carried out a deep research work in the challenging task of textual emotion labeling, drawing relevant conclusions from the application of alternative methods as the bootstrapping technique for IL and the pre-annotation process to efficiently tackle the annotation emotion in text. However, much is the future work that has to be done. This work can be divided into the following groups:

- **Improving IL technique**
Given the core part of bootstrapping technique for IL is the initial unsupervised classification and in order to reduce the false sentences

annotated by this initial process, an improvement would be exploring alternative methods of the seed creation. For instance, considering phenomena as negations or modifiers, adding the analysis of emoticons or smileys to enrich the process, or generating by a Natural Language Generation (NLG) system a set of simple sentences with emotional content (implicit emotional vocabulary) which is later enriched with real sentences by semantic similarity. Moreover, it seems promising to apply this proposal in other Web 2.0 genres such as Twitter messages, Facebook posts, commentaries from news or forums where there is a high emotional content since these genres allow people to post messages to share information, opinion, and emotions.

- **Improving EmoLabel methodology**

The future research in EmoLabel is focused on fully exploiting the great potential of pre-annotation to build up a large amount of emotion annotated data which allows us to apply Machine Learning (ML) or/and Deep Learning (DL) algorithms for the creation of accurate emotion detection systems. To achieve that, we plan the development of the second phase of EmoLabel with new data extracted from different Web 2.0 genres in crowdsourcing platforms (Figure Eight (F8)) with more contributions (internal or external). Moreover, the encouraging results achieved by the *supervised* approach in the pre-annotation process open up the possibility of reusing existing emotion models as IBM Emotional Tone Analyser¹, whose macro-average F1-score is around 60-70%, to pre-annotate new data.

- **Exploring both proposals in other languages**

Mainly, the automatic analysis of emotion in text so far has been focused on English due to the lack of emotion resources in other languages. Because of this and the encouraging results achieved by our proposals (Bootstrapping technique for IL and EmoLabel), it is of remarkable interest to further explore the application of them in other European languages such as Spanish, Italian, or Dutch, as well as in Asian languages as Bangla (or Bengali) or Hindi to analyze how cultural influences affect emotion detection. To this end, it is important that the development of these resources are carried out jointly with

¹<https://tone-analyzer-demo.ng.bluemix.net/>

native people since the fact that the relation of a word to emotion concepts may depend on ideology and in general on cultural aspects (Strapparava, 2016).

- **Analyzing other alternatives for emotion annotation**

While we have evaluated two effective annotation techniques, we not discard and could be attractive the assessment of other alternatives for emotion annotation as Active Learning (AL) or applying game design principles to the task. As for AL strategy, we will apply a method which uses the confidence estimation of classification models to determine if a sentence needs to be reviewed by human annotators or not. This will allow us to reduce the number of sentences used in manual annotation task. To this end, we can use tools like PAL (Skeppstedt et al., 2017), a tool for pre-annotation and AL. About applying game design to annotation task, the idea is that human annotators participate in the emotion labeling without being aware of they are labeling a text with the aim of not affecting their emotional interpretation of the text. To achieve that, it would be interesting to create a mobile app that queries the user about the emotional content of their texts in a non-intrusive way.

- **Studying which are the most appropriate emotion categories for text**

Focusing on categorical emotion models, Ekman's basic emotions are the most popular set employed in computational approaches. However, this emotion model was originally derived from facial expressions and physiology and thus is not based on language theories. During the development of this thesis, we found difficulties in detecting emotions such as DISGUST, FEAR or SURPRISE in text as many other researchers. Thus, the analysis of which are the emotions expressed in text as the Liew (2015)'s study and a definition of a most representative set of categories for textual analysis seems promising and would be a great contribution to the research community.

- **Studying the benefits of emotion analysis in other disciplines.**

Improving the emotion annotation methods will allow us to build up a large amount of data with emotional content which will be used to improve the performance of DL algorithms where huge amounts of data

are required for its training. Moreover, the creation of an accurate emotion recognition system to evaluate and represent people's emotions from comments on the Social Web jointly with the geographic and temporal information available in these genres will allow us to create user emotion profiles which bring substantial benefits to different tasks as suicide prevention, identification of cyberbullying, contribution towards the improvement of people motivation, or e-learning environment.



Universitat d'Alacant
Universidad de Alicante

Annotation Guidelines

In order to evaluate the impact of pre-annotation on the quality of the resulting corpus and on the time needed to annotate, in the second step of EmoLabel (Section 5.2), three different experimental setups have been designed: two with pre-annotation (Pre-ML and Pre-WE) and one without pre-annotation (No-pre). The annotation guidelines of the three tasks are the same except the Section *Options shown* and *Steps* since the emotion options shown to the annotators and the steps to carry out the task slightly vary when the pre-annotation is employed or not. The annotation guidelines employed in these tasks are shown below:

The task

Social Media is a phenomenon that has recently expanded throughout the world and quickly attracted billions of users. Blogs are a Social Media platform that allows users posting messages to share information and opinions. These messengers usually have high emotional content that we want to analyze.

- **GOAL OF THE TASK:** the goal of this task is to determine which is the emotion expressed in sentences extracted from blog posts.
- **LEVEL OF ANNOTATION:** each sentence is annotated at sentence level. This is, you have to think the emotion considering the whole

sentence and choose a label for each sentence.

- **EMOTION ANNOTATION**: although it is obviously possible to express more than one emotion per sentence, the objective is to identify the dominant emotion, the strongest emotion, per sentence.

Categories

Each sentence will be classified according to one of Ekman basic emotions or the NEUTRAL category shown below:

- **ANGER (choler, ire)**: a strong feeling that makes you want to hurt someone or be unpleasant because of something unfair or unkind that has happened.
 - My mother truly wants to murder my father, and then gets pissed off when i don't agree with her.
 - Again, makes me so angry.
 - I emailed the seller and told him about it and he was very arrogant and rude.
- **DISGUST (dislike, revulsion)**: a strong feeling of disapproval and dislike at a situation, person's behavior, etc.
 - The job was frustrating today.
 - I remember walking off the field disgusted with myself.
 - I come home after two weeks and our place is fucking disgusting.
- **FEAR (terror, panic, phobia)**: an unpleasant emotion or thought that you have when you are frightened or worried by something dangerous, painful, or bad that is happening or might happen.
 - I'm still pretty freaked out by the piano exam next week.
 - I can't wait, but I'm nervous as hell now.
 - Afraid that you will be hurt yet again.
- **JOY (happy, felicity)**: the feeling of being happy, pleasure or satisfaction.
 - It actually was quite a bit of fun.

- He really laughed a lot today.
- It's pretty sweet and I feel very blessed that I have it.
- **SADNESS (unhappy, depressed)**: the feeling of being sad, not satisfied or unhappy.
 - I can't relate to 99% of humanity.
 - I felt really awful throughout the whole day too.
 - I'll miss you always forever and ever.
- **SURPRISE (amazing)**: the feeling caused by something unexpected happening.
 - I can't believe she is FINALLY here!!!
 - I just stood there in shock!
 - I haven't even seen snow in years and to have it in October it too much to wrap around my little head!
- **NEUTRAL**: nothing remarkable is happening. **NEUTRAL** is used in the place where there is no emotion present in the sentence or where there is no emotion discernible in the sentence.
 - So Adam and I went to App State to go hiking.
 - When they were together everyone was equal.
 - I actually thought I was making a difference.

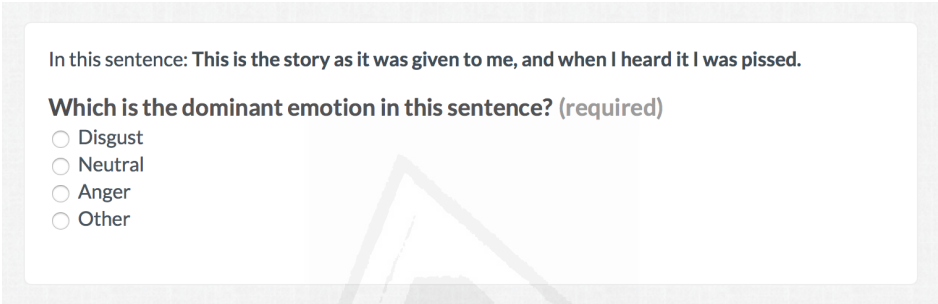
The sentences may involve:

- Explicit expressions of one of the emotions.
- Some information that allows one to infer that the sentence expresses a particular emotion.
- The sentence does not involve any relation to emotions. In this case, the sentence will be considered **NEUTRAL**.

Options shown

Option shows in Pre-ML and Pre-WE tasks

With the aim of facilitating your task, an automatic system will pre-select three of them and will be shown. If the emotion that you consider the dominant is not between these pre-selected categories, you can choose the option Other and the rest of the categories will be shown. You can see an example below:



In this sentence: **This is the story as it was given to me, and when I heard it I was pissed.**

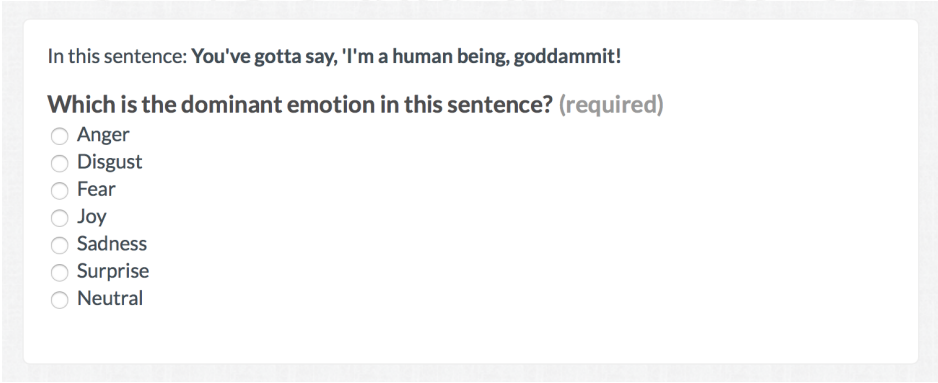
Which is the dominant emotion in this sentence? (required)

- Disgust
- Neutral
- Anger
- Other

Figure A.1: An example of how the sentences are shown to the human annotators when the pre-annotation process is employed.

Option shows in No-pre task

For each sentence, the six Ekman basic emotions will be shown plus the Neutral category as you can see in the image below:



In this sentence: **You've gotta say, 'I'm a human being, goddammit!**

Which is the dominant emotion in this sentence? (required)

- Anger
- Disgust
- Fear
- Joy
- Sadness
- Surprise
- Neutral

Figure A.2: An example of how the sentences are shown to the human annotators when the pre-annotation process is no employed.

Time

Moreover, in this task, we want to evaluate the time needed to annotate a set of sentences. Thus, it is important that if you have to stop the task because of different causes: receive a call, you need a break, etc. **You have to click on 'Give up' before leaving the task.**

Important Notes

- This document provides the annotations of specific examples, but do not state that a sentence which contains specific words should always be annotated a certain emotion.
- When you think about which is the dominant emotion in a sentence think which is the emotion that the majority of people would associate.
- This task is about the emotions expressed by the text and not about the emotion that you can feel when you read it from a personal point of view. **Try to be the most objective possible.**
- Most importantly, **try not to over-think the answer, follow your first intuition.**

Remember

- The sentences must be annotated **considering only the information that you have in the sentences** without thinking what happened before or after the event or situation expressed in the sentence.
- If you have doubts about which emotion should be selected, **the vocabulary used in the sentence could help you to decide the most suitable emotion.**
- The beginning of a sentence usually contains the main clause which the most important one. Then, **if a sentence expresses two emotions, you should annotate the emotion of the main clause.**

Steps

Steps in Pre-ML and Pre-WE tasks

1. Read the sentence.
2. Think briefly which is the dominant emotion in the sentence which is the strongest emotion expressed in the sentence.
3. Choose between the three emotional categories pre-selected by the system if between them there is the dominant emotion that you consider.
4. If the emotion that you consider dominant is not between the pre-selected categories, choose 'Other' option and select the dominant emotion you consider most appropriate.

Steps in No-pre task

1. Read the sentence.
2. Think briefly which is the dominant emotion in the sentence which is the strongest emotion expressed in the sentence.
3. Choose between the emotional categories which one is the dominant emotion.

Resumen

B.1 Introducción

Como parte esencial de las relaciones humanas, el análisis de las emociones ha sido un tema cautivador en disciplinas como: la neurociencia, la ciencia cognitiva, la psicología o las ciencias de la conducta. Este interés también ha atraído a investigadores del campo de *Inteligencia artificial* (del inglés Artificial Intelligence, [AI](#)), ya que las emociones son cruciales para mejorar la experiencia de los usuarios en la *Comunicación mediante ordenadores* (del inglés Computer-Mediated Communication, [CMC](#)) y la *Interacción persona-ordenador* (del inglés Human-Computer Interaction, [HCI](#)) ([Cowie et al., 2001](#)), donde el lenguaje juega un rol importante.

El lenguaje es un medio de comunicación humano, tanto escrito como hablado, para expresar nuestras ideas, nuestros pensamientos y más importante, nuestras emociones. Basándonos en las funciones del lenguaje definidas por el modelo de [Jakobson \(1960\)](#), se puede observar la importancia de la relación entre lenguaje y emoción, ya que identifica la función emotiva como una de las seis funciones del lenguaje. Por lo tanto, el lenguaje es una poderosa herramienta para comunicar y transmitir nuestras emociones.

En el campo [HCI](#), el análisis emocional ha sido evaluado usando diferentes *Interfaces de usuario* (del inglés User Interface, [UI](#)) como: las expresiones faciales, la voz y el texto ([Kim, 2011](#)). La importancia del texto como

medio de comunicación con los ordenadores se ha incrementado notablemente con la aparición de la Web 2.0 o la Web social (del inglés *Social Web*). A diferencia de la Web 1.0 donde los usuarios eran agentes pasivos que se limitaban a leer o recibir información, la Web 2.0 les permite comunicarse y compartir información en Internet usando los ordenadores, teléfonos móviles o cualquier dispositivo con conexión a Internet. Hay muchas plataformas de redes sociales como: Facebook¹, Instagram² o Youtube³; Blogs como la plataforma Blogger⁴ o WordPress⁵ donde la gente publica sus experiencias en diferentes publicaciones; o servicios de microblogging como Twitter⁶ que son blogs donde los usuarios comparten pequeños fragmentos sobre sus ideas o pensamientos (frases, imágenes o videos) (Kaplan & Haenlein, 2011).

Como muestran las estadísticas, el fenómeno de las plataformas sociales se ha extendido a través de todo el mundo y rápidamente, ha atraído millones de usuarios (Farzindar & Inkpen, 2015). Por ejemplo, el último ranking sobre el uso de redes sociales publicado por Statista⁷, el portal de estadísticas más grande del mundo, publicado en enero de 2018, sitúa a Facebook como la red social con más usuarios (2.167 millones), en segundo lugar está YouTube con 1.500 millones usuarios e Instagram se sitúa en séptimo lugar con más de 800 millones de usuarios activos. Como consecuencia y debido al uso masivo de estas redes sociales por parte de los usuarios, ha habido un crecimiento exponencial de la información subjetiva en la Web 2.0.

De forma paralela al crecimiento de la información subjetiva, investigadores en *Procesamiento del Lenguaje Natural* (PLN) (del inglés Natural Language Processing, NLP) han mostrado un creciente interés en desarrollar métodos para extraer automáticamente el conocimiento de estas nuevas fuentes. El PLN es un campo de investigación que se ocupa de la investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio de lenguajes naturales. Dada la importancia de las emociones en el lenguaje, dentro del PLN ha nacido una sub-disciplina cuyo objetivo es la identificación y extracción de la subjetivi-

¹<https://www.facebook.com/>

²<https://www.instagram.com/>

³<https://www.youtube.com/>

⁴<https://www.blogger.com>

⁵<https://www.wordpress.com/>

⁶<https://twitter.com/>

⁷<https://es.statista.com/>

dad y contenido emocional del texto, llamada *Análisis de Sentimientos* (AS) (del inglés Sentiment Analysis, SA).

El objetivo principal del AS es la identificación de sentimientos, opiniones y emociones en el texto. La mayoría de los trabajos en esta disciplina se han centrado tradicionalmente en el reconocimiento de la polaridad del sentimiento (POSITIVO, NEGATIVO, NEUTRO). Estos trabajos se enmarcan dentro de la tarea de *Minería de Opiniones* (del inglés Opinion Mining, OP). Sin embargo, el reconocimiento de tipos de emociones como categorías emocionales (IRA, ASCO, MIEDO, etc.) o dimensiones emocionales (*placer, activación, dominancia*) ha aumentado recientemente, ya que reconocer las emociones transmitidas por un texto puede conducir a una mejor comprensión del contenido del texto (Aman, 2007). Este análisis es conocido como *Reconocimiento de Emociones* (RE) (del inglés Emotion Recognition, ER) y es donde se enmarca este trabajo de tesis.

Recientemente, ha habido un interés creciente en el RE en el texto por parte de la comunidad científica, debido principalmente a la aparición de los nuevos géneros de la Web 2.0 y su potencial para aportar beneficios sustanciales a diferentes sectores como la prevención del suicidio (Cherry et al., 2012; Desmet & Hoste, 2013), identificación de casos de ciberacoso (Dadvar et al., 2013), o las aportaciones en la mejora de la motivación de los estudiantes (Suero Montero & Suhonen, 2014).

Motivación

La tarea de RE en texto escrito ha sido abarcada utilizando diferentes técnicas de PLN, incluyendo el uso de *Aprendizaje automático* (del inglés Machine Learning, ML), métodos basados en reglas o aproximaciones basadas en conocimiento. Sin embargo, la mayoría de ellas se han basado en algoritmos de aprendizaje automático, debido principalmente a su escalabilidad, capacidad de aprendizaje y su rápido desarrollo.

El *Aprendizaje automático* es una disciplina científica cuyo objetivo es desarrollar y estudiar algoritmos que permitan a los ordenadores *aprender* a partir de la *experiencia*. Esta experiencia son datos que los algoritmos utilizan para mejorar el rendimiento o para hacer predicciones precisas (Mohri et al., 2012). Este conjunto de datos, llamado *datos de entrenamiento* (del inglés *training data*) debe ser etiquetado cuando usamos aprendizaje automático

supervisado, mientras que el aprendizaje *no supervisado* recibe datos no anotados. El escenario más común en el RE en texto es el uso de algoritmos de aprendizaje automático *supervisados* ya que estos algoritmos conducen a mejores resultados que el resto de alternativas.

Centrándonos en el RE en texto, los algoritmos de aprendizaje *supervisado* consisten en inferir una función a partir de un conjunto de datos anotados con la emoción correcta (*datos de entrenamiento*). Después del entrenamiento, el modelo es capaz de predecir la emoción de nuevos ejemplos. El éxito de las predicciones hechas por el modelo dependen directamente de la calidad y el tamaño de nuestros *datos de entrenamiento*. Por lo tanto, el conjunto de datos utilizado en el entrenamiento es crucial para la creación de un sistema de RE preciso que pueda generar resultados fiables.

Este requisito de calidad y tamaño del conjunto de *entrenamiento* es incluso más importante en la nueva disciplina de *Aprendizaje profundo* (del inglés Deep Learning, DL). Es una parte de la familia de algoritmos de *Aprendizaje automático* que utiliza un nivel jerárquico de redes neuronales para realizar el proceso de aprendizaje (Deng & Yu, 2014). Una de las características más relevantes de este tipo de algoritmos es que no necesitan un proceso de diseño de características. Sin embargo, esta propiedad implica que el conjunto de *datos de entrenamiento* de las arquitecturas de *Aprendizaje profundo* requieren mayores cantidades de datos que los algoritmos tradicionales de *Aprendizaje automático*.

Sin embargo, la creación de un conjunto de datos etiquetados para el RE en texto no es trivial, ya que la detección de emociones en texto puede ser difícil incluso para los seres humanos, porque los contextos personales de cada persona pueden influir en la interpretación de las emociones. Muchas de las investigaciones llevadas a cabo hasta el momento, han mostrado las dificultades relacionadas con esta tarea, como: la detección de un buen *Acuerdo entre anotadores* (del inglés Inter-Annotator Agreement, IAA) o el tiempo necesario para su desarrollo. Como consecuencia, la obtención de datos con contenido emocional se ha convertido en una de las tareas más desafiantes de la sub-disciplina de RE en texto.

Definición del problema y objetivo

Teniendo en cuenta las dificultades del RE en texto y con el fin de disminuir y contrarrestar el desafío de la anotación de emociones, esta investigación abarca el análisis de diferentes aproximaciones semiautomáticas con el objetivo de mejorar la anotación de emociones en texto escrito. Más específicamente, se han investigado dos técnicas que han demostrado su usabilidad y efectividad en otras tareas de PLN: bootstrapping basado en Intensional Learning (IL) y un proceso de pre-anotación.

Estas técnicas han sido evaluadas con el objetivo de proporcionar un método capaz de anotar eficientemente grandes cantidades de texto en inglés en cualquier género textual y con sólidos estándares de fiabilidad. Estos requisitos incrementan la dificultad de la tarea debido a que se ha abarcado desde un punto de vista general, es decir, independientemente del género y del conjunto de etiquetas emocionales empleadas.

La tarea de anotación de emociones se lleva a cabo a nivel de frase porque en géneros como blogs o cuentos, un análisis más detallado es beneficioso, ya que a menudo hay una progresión de las emociones en el texto narrativo (Kim, 2011). Además, en redes sociales como Twitter o Facebook, las personas expresan sus opiniones y/o emociones a través de comentarios cortos. El conjunto de etiquetas empleado son las seis emociones básicas definidas por Ekman (1992): IRA, ASCO, MIEDO, ALEGRÍA, TRISTEZA, y SORPRESA, porque este ha sido el conjunto de emociones más empleado en los enfoques computacionales y además, es el más aceptado por diferentes investigadores, como veremos en el próximo capítulo. Existen diferentes perspectivas desde las cuales se pueden analizar las emociones en el texto: *escritor*, *lector*, *texto*. La perspectiva del *escritor* se refiere a cómo se siente alguien mientras produce una afirmación, mientras que la perspectiva del *lector* es cómo se siente alguien después de leer un texto. Por último, en cuanto a la perspectiva del *texto* no se especifica a ninguna persona real en cuanto a la percepción de una emoción y se considera que la emoción es una propiedad intrínseca de una oración. Nuestros enfoques se han desarrollado teniendo en cuenta la perspectiva del *texto* porque nuestro objetivo es analizar la orientación emocional del texto en sí mismo, sin considerar el contexto emocional del escritor o lector.

B.2 Contribuciones

Las aportaciones de la presente investigación son descritos de una manera breve en esta sección. En concreto estos aportes se pueden agrupar en dos grandes bloques: *i*) la aplicación de la técnica bootstrapping basada en **IL** en la tarea de anotación de emociones en texto, desarrollada en el Capítulo 4; y *ii*) el desarrollo de una nueva metodología de etiquetado de emociones semiautomática: *EmoLabel*, la cual es presentada en el Capítulo 5.

Los siguientes apartados resumen de manera breve las propuestas, desarrollos y resultados obtenidos para cada una de las aportaciones.

B.2.1 *Intensional Learning* para la anotación de emociones

Nuestra primera aproximación para abarcar y mejorar la tarea de anotación de emociones en texto es una aproximación bootstrapping basada en **IL**, previamente propuesta por [Gliozzo et al. \(2009\)](#), que consta de dos pasos principales:

Paso 1 *Categorización inicial basada en similitud.* Este paso se aborda aplicando un criterio de similitud entre una semilla etiquetada inicialmente y cada oración no anotada. El resultado de este paso es una categorización inicial de los documentos no anotados.

Paso 2 *Entrenamiento de un clasificador supervisado con una o más iteraciones utilizando el conjunto de datos categorizado en el paso anterior.* La salida del paso 1 se utiliza para entrenar un clasificador supervisado. En este paso se pueden utilizar diferentes algoritmos como *Máquinas de vector de soporte* (del inglés Support Vector Machine, **SVM**) o Naive Bayes.

A diferencia de las aproximaciones tradicionales de bootstrapping basadas en ejemplos, conocidos como Extensional Learning (**EL**) en la terminología de la teoría de la computabilidad, **IL** se basa en el método clásico de clasificación basado en reglas, donde el usuario especifica reglas de clasificación exactas que operan en el espacio de características. Esta propiedad es particularmente relevante para el RE en texto, ya que en **EL**, el hecho de que los ejemplos sean anotados manualmente por humanos implicada que el contexto personal

de cada uno de ellos puede influir en la interpretación de las emociones. Sin embargo, en [IL](#), la influencia de la comprensión personal de las emociones se reduce, ya que su participación se limita a la definición de las reglas.

Dada esta característica y teniendo en cuenta nuestro objetivo de desarrollar técnicas eficientes capaces de construir corpus con contenido emocional en diferentes géneros, el trabajo presentado en este Capítulo ha consistido en el diseño y evaluación de una propuesta de [IL](#) para la anotación de emociones en texto. Concretamente, nuestra propuesta se muestra en la Figura [B.1](#) y consiste en:

- **Paso 1** *Categorización inicial basada en similitud.* En nuestro enfoque, este paso está compuesto por dos sub-pasos no supervisados:
 - **Paso 1.1:** creación de la semilla etiquetada inicialmente. Para ello, en este paso se emplea un lexicón emocional y las frases son anotadas en función de las palabras emocionales que contengan.
 - **Paso 1.2:** extensión de la semilla inicial obtenida en el Paso 1.1 utilizando una métrica de similitud semántica entre oraciones.
- **Paso 2** *Entrenamiento de un clasificador supervisado con una o más iteraciones utilizando el conjunto de datos categorizado en el paso anterior.* Nuestro enfoque utiliza un clasificador supervisado [SVM](#) con una iteración.

El proceso recibe como datos de entrada una colección de oraciones no etiquetadas, un conjunto de categorías emocionales (por ejemplo, las emociones básicas de [Ekman \(1992\)](#), las de [Plutchik \(1962\)](#) o las emociones de [Izard \(1971\)](#)) y el número de categorías emocionales que deseamos anotar (la emoción dominante o todas las detectadas en la frase). La adaptabilidad de la propuesta al conjunto de categorías de emociones, así como al número de categorías anotadas, es una de las aportaciones más novedosas de esta propuesta, ya que esta flexibilidad permite el uso de esta técnica en diferentes dominios y aplicaciones. Por ejemplo, las emociones como **ABURRIMIENTO**, **ANSIEDAD** e **INTERÉS** se detectan típicamente en el dominio de la educación ([Kim, 2011](#)), mientras que las emociones como **DIVERTIDO** o **INSPIRADO** se analizan en el dominio de las noticias online⁸. Además, esta adaptabilidad

⁸<http://www.rappler.com/>

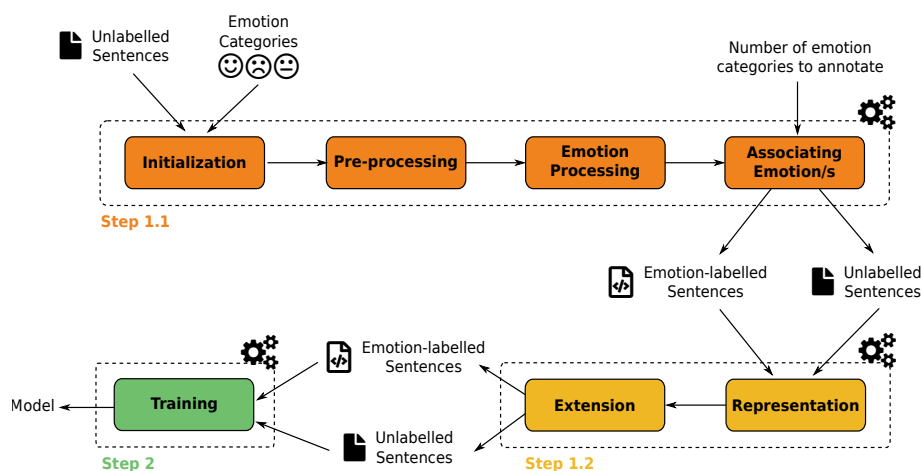


Figura B.1: Descripción general del proceso bootstrapping basado en Intensional Learning (IL).

puede ser útil en aquellas aplicaciones en las que la detección de la intensidad de las emociones es importante, como son los sistemas de recomendación.

Como hemos mencionado anteriormente, en el paso 1.1 (creación de la semilla) se emplea un lexicón emocional para etiquetar las frases en función de las palabras emocionales que contengan. El lexicón empleado en nuestra propuesta es [EmoLex](#), un lexicón de dominio general anotado con las emociones de [Plutchik \(1980\)](#), entre las que se encuentran las de [Ekman \(1992\)](#), el conjunto con el que trabajamos en esta disertación. Con el objetivo de evaluar diferentes aproximaciones, se evalúan dos propuestas más en las que se utilizan versiones extendidas de [EmoLex](#) utilizando la base de datos léxica WordNet ([WN](#)) y un tesoro de [Oxford](#). Por tanto, se presentan tres aproximaciones donde el proceso es el mismo pero emplean diferentes versiones de [EmoLex](#):

- *Original:* se utiliza la versión original de [EmoLex](#).
- *Enriquecida WN:* se emplea una versión del lexicón de [EmoLex](#) extendida automáticamente con sinónimos de [WN](#).
- *Enriquecida Oxford:* la versión del lexicón empleada es una versión extendida de [EmoLex](#) con sinónimos de [Oxford](#).

Una vez creadas las semillas, en el paso 1.2 se procede a la extensión de las

mismas utilizando una métrica de similitud. Si bien hay diferentes enfoques para determinar la similitud semántica en el texto (Kenter & de Rijke, 2015), nuestro enfoque utiliza semántica distribucional, ya que nuestro objetivo es utilizar un modelo genérico que no requiera análisis léxico ni lingüístico y que no utilice fuentes externas de conocimiento semántico. Los *Modelos de Semántica Distribucional* (MSD) (del inglés, *DSMs*, *Distributional Semantic Models*) se basan en la suposición de que el significado de una palabra se puede inferir desde la forma en que se usa. Por lo tanto, estos modelos construyen dinámicamente representaciones semánticas (espacios vectoriales semánticos con muchas dimensiones) a través de un análisis estadístico de los contextos en los que ocurren las palabras. Concretamente, cada una de las semillas es extendida utilizando cuatro modelos que incorporan esta intuición: un modelo Latent Semantic Analysis (LSA) y tres modelos Word2Vec (W2V). Las características de cada uno de ellos se muestran en la Tabla 4.3.

Una vez diseñadas y desarrolladas nuestras propuestas, estas fueron evaluadas con el objetivo de verificar la usabilidad de técnica IL en el etiquetado de las emociones en textos de diferentes géneros. Para ello, todas las aproximaciones presentadas en este Capítulo fueron evaluadas con dos corpus de emociones: Aman y Affective Text corpus. Aman (Aman & Szpakowicz, 2007) es una colección de 4.000 frases de publicaciones realizadas en blogs recopiladas directamente de la Web y anotadas manualmente con las seis emociones básicas de Ekman (1992). Mientras que Affective Text (Strapparava & Mihalcea, 2007), es un corpus con 1.250 titulares de noticias periodísticas que fueron extraídas de los principales periódicos como New York Time, CNN y BBC News, que están etiquetados manualmente con las emociones de Ekman.

Respecto a la metodología de evaluación, esta se divide en dos partes:

- Entrenamiento de un clasificador supervisado con el corpus anotado automáticamente resultante del paso 1 de IL, para evaluar su usabilidad.
- Cálculo del acuerdo (IAA) entre los corpus anotados automáticamente y las versiones *gold standard* de cada uno de ellos, con el objetivo de evaluar la calidad de las anotaciones automáticas.

Una vez realizada la experimentación, los resultados nos permiten inferir una serie de conclusiones de gran importancia:

1. Se demuestra la viabilidad y usabilidad de la técnica bootstrapping basada en **IL** para el etiquetado automático de las emociones, ya que la evaluación de clasificación y acuerdo realizada en ambos corpus lograron resultados prometedores con altos beneficios en términos de coste y tiempo de desarrollo.
2. En cuanto a los MSD, los resultados obtenidos no muestran diferencias significativas entre los modelos. Por lo que podemos concluir que el paso 1.2 (extensión de la semilla) es independiente del MSD empleado, lo que proporciona flexibilidad a nuestra propuesta.
3. Respecto al lexicón empleado, los resultados han sido satisfactorios teniendo en cuenta que es un recurso de dominio general y se ha aplicado en dos géneros diferentes: titulares y publicaciones de blogs. Sin embargo, para mejorar los resultados, sería recomendable emplear lexicones adaptados al dominio.
4. Las mejoras de los enfoques *enriquecidos* se han demostrado para varias emociones en Aman corpus, por lo que el proceso de extensión podría ser beneficioso según el género textual analizado. Por lo tanto, la usabilidad de estos enfoques se analizará en profundidad en trabajos futuros.

B.2.2 *EmoLabel*: metodología semi-automática para la anotación de emociones

La pre-anotación es un procedimiento para etiquetar automáticamente un corpus utilizando un sistema automático, que posteriormente es revisado por un anotador humano. Los anotadores humanos usualmente corrigen errores u omisiones realizadas por el sistema automático, o hacen una elección entre las diferentes opciones dadas por el sistema automático (Skeppstedt et al., 2017). Esta técnica ha sido ampliamente utilizada en otras tareas de PLN como el *Reconocimiento de entidades* (del inglés **NER**, Named Entity Recognition), el *Etiquetado gramatical* (del inglés **POS**, Part-Of-Speech tagging), o la *Desambiguación lingüística* (del inglés **WSD**, Word Sense Disambiguation), proporcionando una ganancia en tiempo y coste en la tarea de anotación manual.

Dadas las dificultades asociadas al proceso de anotación de emociones en texto y los beneficios proporcionados por la pre-anotación en otras tareas de PLN, en el Capítulo 5 presentamos nuestra propuesta para abordar de forma eficiente el etiquetado de emociones: *EmoLabel*, una metodología semiautomática basada en un proceso de pre-anotación automática. El proceso consta de dos fases principales que se muestran en la Figura B.2:

Fase 1 *Proceso de pre-anotación*. Esta fase es llevada a cabo aplicando un proceso automático para anotar las oraciones no etiquetadas con un conjunto reducido de categorías emocionales.

Fase 2 *Refinamiento manual*. El resultado de la Fase 1 es examinado por anotadores humanos que determinan cuáles son finalmente las emociones asociadas a cada oración. En nuestra propuesta, esta fase tiene como objetivo identificar cual es la emoción dominante en cada una de las oraciones.

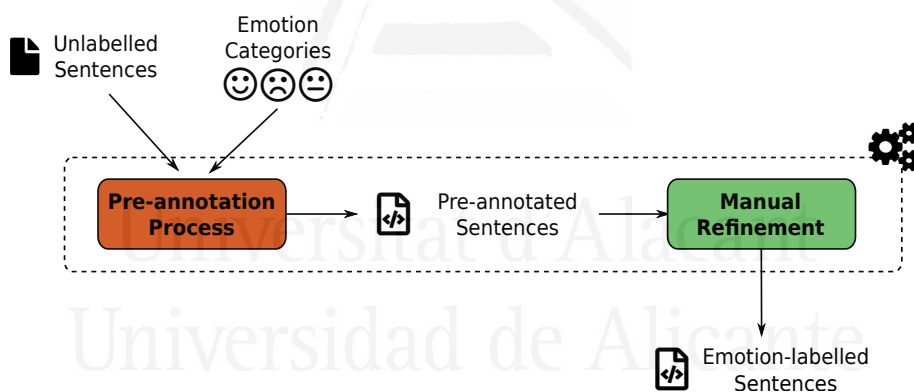


Figura B.2: Descripción general de la metodología *EmoLabel*.

Para la primera fase de *EmoLabel*, diseñamos dos procesos de pre-anotación automáticos: una aproximación *no supervisada* basada en MSD (DSMs) y un enfoque *supervisado* basado en *Aprendizaje automático* (ML). Ambos reciben como parámetros de entrada: una colección de oraciones no etiquetadas y un conjunto de categorías emocionales (por ejemplo, las emociones de Ekman (1992) o de Plutchik (1980)). Esta adaptabilidad de *EmoLabel*, como ocurre con nuestra primera propuesta basada en IL, permite que los procesos propuestos se puedan emplear en diferentes dominios y/o aplicaciones.

Pre- anotación *no supervisada*

Como mencionamos en la propuesta anterior, la intervención humana en una aproximación *no supervisada* es mínima y, por lo tanto, es una propuesta interesante para la anotación de emociones, ya que el contexto personal de cada anotador no influye en la interpretación emocional de las oraciones. Esa característica junto con los resultados obtenidos en nuestra primera propuesta, nos permite considerar relevante el desarrollo de una propuesta de pre- anotación *no supervisada* basada en MSD.

La gran ventaja del uso de estas representaciones que codifican la información semántica es que pueden generarse a partir de grandes corpus de texto no etiquetado y en un período de tiempo razonable. Por lo tanto, es una manera simple de filtrar el número de categorías de emociones que se pueden asociar a cada oración y, de esta manera, reducir la ambigüedad de la segunda fase de *EmoLabel*.

Pre- anotación *supervisada*

En la tarea de RE en texto se han aplicado una amplia variedad de técnicas de PLN para abordarla. Sin embargo, la mayoría de ellas se han llevado a cabo utilizando aprendizaje automático *supervisado* dado que conduce a mejores resultados que las aproximaciones *no supervisadas* (Kim, 2011). Por ello, el número de sistemas de RE en texto escrito basados en estos enfoques es mayor. La precisión de estos sistemas varía entre un 60% y 70% cuando intentan determinar la emoción dominante (Aman & Szpakowicz, 2007; Ghazi et al., 2010; Wang et al., 2012), lo que indica que es una tarea no resuelta.

A pesar de ello, estos enfoques podrían emplearse en procesos de pre- anotación de emociones para reducir automáticamente el número de categorías emocionales. Este es el objetivo de los métodos presentados en este Capítulo, los cuales son evaluados en las tareas de pre- anotación.

Con este propósito en mente, se proponen tres aproximaciones:

- *CountWordEmo*: en esta propuesta, el conjunto de características está compuesto por un vector de 8 componentes donde las seis primeras representan cada una de las emociones de Ekman (1992) (IRA, ASCO,

MIEDO, ALEGRÍA, TRISTEZA, SORPRESA) y las otras dos componentes contienen los valores de polaridad y subjetividad de cada oración proporcionados por una herramienta de *Análisis de Sentimientos* (De Smedt & Daelemans, 2012).

- *EmoLexicon*: en esta aproximación, el conjunto de características son derivadas del lexicón de emoción empleado. Por lo que, las características son los tokens en común entre el lexicón y el conjunto de datos elegido.
- *1-grams*: en esta propuesta, se utilizan unigramas como características. Los modelos basados en unigramas se han aplicado ampliamente en la clasificación de textos y han mostrado buenos resultados en tareas de clasificación (Kennedy & Inkpen, 2006).

Como algoritmo de aprendizaje, en todas las propuestas se ha utilizado un multi-clasificador de *Máquinas de soporte vectorial* (del inglés Support Vector Machine, SVM) utilizando el entorno *scikit-learn* (Pedregosa et al., 2011).

Refinamiento manual

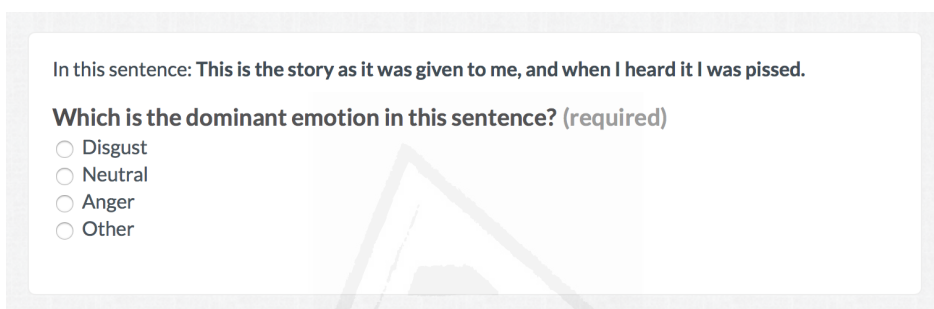
Una vez que las oraciones no etiquetadas han sido anotadas con uno u otro proceso de pre-anotación, los anotadores humanos realizan una tarea de refinamiento con el objetivo de determinar cuáles son las emociones asociadas a cada oración. La cantidad de categorías emocionales finalmente etiquetadas dependerá de nuestros objetivos. En nuestra propuesta, esta fase tiene como objetivo la detección de la emoción dominante.

Para evaluar el impacto de la pre-anotación sobre la calidad del corpus resultante y el tiempo empleado en la tarea de anotación, se han diseñado tres configuraciones diferentes:

- **Pre-ML**: en esta configuración, el mejor modelo de pre-anotación *supervisada* se utiliza para seleccionar el conjunto de emociones pre-anotadas en cada oración.
- **Pre-WE**: en esta configuración, se utiliza el mejor modelo de pre-anotación *no supervisada* para seleccionar las emociones propuestas a los anotadores humanos.

- **No-Pre:** en esta configuración, no se emplea ningún proceso de pre-annotación. Por lo tanto, todas las categorías de emociones empleadas se muestran a anotadores humanos como muestra la Figura B.4.

Cuando se emplean las configuraciones con pre-annotación (Pre-ML y Pre-ML), las emociones propuestas por el sistema se muestran en primer lugar a los anotadores humanos, pero también tienen la posibilidad de seleccionar otra emoción no pre-seleccionada automáticamente. Para ello, deben elegir la opción 'Other' y el resto de emociones se mostrarán (Figura B.3).

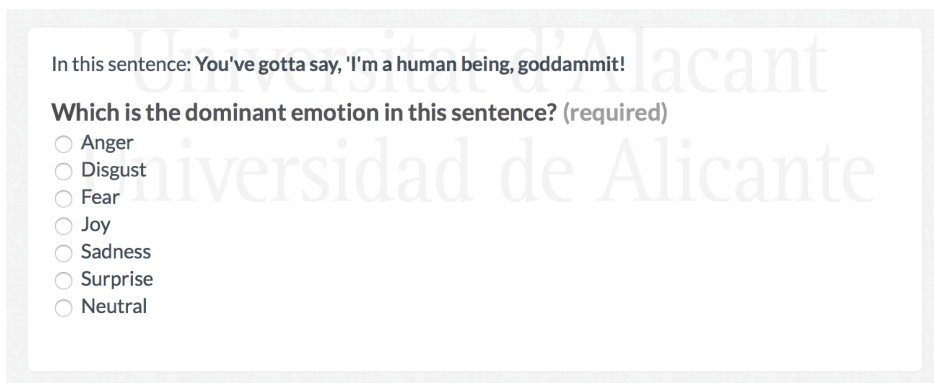


In this sentence: **This is the story as it was given to me, and when I heard it I was pissed.**

Which is the dominant emotion in this sentence? (required)

- Disgust
- Neutral
- Anger
- Other

Figura B.3: Un ejemplo de como se muestran las frases a los anotadores humanos cuando se utiliza el proceso de pre-annotación.



In this sentence: **You've gotta say, 'I'm a human being, goddammit!**

Which is the dominant emotion in this sentence? (required)

- Anger
- Disgust
- Fear
- Joy
- Sadness
- Surprise
- Neutral

Figura B.4: Un ejemplo de como se muestran las frases a los anotadores humanos cuando no se utiliza el proceso de pre-annotación.

Todas las tareas de anotación manual fueron llevadas a cabo por tres anotadores con un buen conocimiento del idioma inglés.

En un experimento previo, esta fase se diseñó utilizando tres conjuntos de datos diferentes ($D1$, $D2$, $D3$) para cada configuración (Pre-ML, Pre-

WE, No-pre). Sin embargo, detectamos que la selección aleatoria de las oraciones que componen cada uno de los conjuntos de datos podían afectar negativa o positivamente a los resultados obtenidos en cada configuración. En consecuencia, decidimos aplicar la *Validación cruzada* (del inglés *cross-validation*) para que los resultados no se vieran afectados por el sesgo del azar y la curva de aprendizaje de los anotadores. Por lo tanto, cada anotador llevó a cabo tres tareas de anotación en el orden descrito en la Tabla B.1. De esta manera, cada conjunto de datos fue anotado con todas las configuraciones y por todos los anotadores. Por ejemplo, el anotador 1 realizó tres tareas donde *D1* está pre-anotado con el enfoque *supervisado* (Pre-ML), *D2* con el enfoque *no supervisado* (Pre-WE) y el *D3* no está anotado previamente (No-pre).

Tabla B.1: Configuración de la *Validación cruzada*

	<i>D1</i>	<i>D2</i>	<i>D3</i>
Annotator 1	Pre-ML	<u>Pre-WE</u>	No-pre
Annotator 2	No-pre	Pre-ML	<u>Pre-WE</u>
Annotator 3	<u>Pre-WE</u>	No-pre	Pre-ML

La evaluación de *EmoLabel* requiere una evaluación intrínseca y extrínseca. La evaluación intrínseca implica la evaluación de los procesos de pre-anotación automáticos para determinar cuál de ellos se emplea en la segunda fase de *EmoLabel*. Todas las aproximaciones de pre-anotación fueron evaluadas con dos corpus de emociones: Aman y una versión reducida de EmoTweet-28 (EmoTweet-5). Como hemos mencionado anteriormente, Aman ([Aman & Szpakowicz, 2007](#)) es una colección de 4.000 frases de publicaciones realizadas en blogs recopiladas directamente de la Web y anotadas manualmente con las seis emociones básicas de [Ekman \(1992\)](#). En cambio, la versión reducida de EmoTweet-28 ([Liew et al., 2016](#)), comprende casi 6.000 tweets etiquetados manualmente con cinco de las seis emociones básicas de Ekman: IRA, MIEDO, ALEGRÍA, TRISTEZA y SORPRESA. En cuanto a la metodología de evaluación, el proceso de pre-anotación se evalúa midiendo precisión, cobertura y medida-F (del inglés, *F-score*) de las emociones propuestas por nuestro sistema contra las versiones *gold standard* de cada uno de los corpus. Como el proceso

de pre-annotación etiqueta un subconjunto de categorías emocionales, si la emoción correcta (la contenida en *gold standard*) es una de las emociones pre-annotadas, la predicción se considera correcta.

En cuanto a la evaluación extrínseca, esta tiene como objetivo la evaluación del rendimiento de los anotadores en la segunda fase de la metodología. En esta fase, el corpus empleado es el de Aman. Por un lado se evalúa la calidad del corpus resultante calculando el acuerdo (IAA) entre cada anotador y el *gold standard* de Aman corpus. En particular, la métrica empleada es Fleiss (1971) kappa. Por otro lado, también evaluamos el tiempo necesario por cada anotador para llevar a cabo cada tarea. Para ello, se utiliza el registro de tiempo proporcionado por la plataforma de anotación.

El estudio realizado permite verificar la adecuación y fiabilidad de nuestra metodología en la anotación de emociones en texto escrito y nos permite obtener las siguientes conclusiones principales:

1. Se demuestran los beneficios de los procesos de pre-annotación en el etiquetado de emociones, ya que los resultados en tiempo de anotación muestran una ganancia de cerca de un 20% cuando se aplica el proceso de pre-annotación *supervisado* (Pre-ML) con respecto a no utilizar la pre-annotación (No-pre). Además, los experimentos realizados muestran que todas las tareas alcanzan un "acuerdo sustancial" y, por tanto, el proceso de pre-annotación no reduce el rendimiento del anotador (IAA).
2. Con respecto a la evaluación intrínseca, las ganancias obtenidas por el método de pre-annotación *supervisada* en términos de tiempo con respecto al proceso de pre-annotación *no supervisada* permiten concluir que el uso de un proceso de pre-annotación preciso proporciona beneficios relevantes en la tarea de etiquetado de emociones. En consecuencia, los sistemas existentes de detección de emociones desarrollados hasta el momento podrían emplearse para pre-annotar nuevos datos.
3. Las mejoras alcanzadas por el Anotador 3 (el anotador con peor rendimiento) en términos de tiempo y acuerdo demuestran la usabilidad de nuestra metodología cuando los anotadores no son buenos, ya que sus mejores resultados se han obtenido cuando se emplea un proceso de pre-annotación.

B.3 Conclusiones y trabajo futuro

Debido a la necesidad de desarrollar nuevas técnicas capaces de etiquetar eficientemente grandes cantidades de datos con emociones, en cualquier género textual y con sólidos estándares de fiabilidad, este trabajo se ha centrado en uno de los desafíos más importantes del RE en texto: el desarrollo de técnicas para la anotación de corpus con emociones. Nuestra principal motivación fue las dificultades asociadas con el desarrollo de este tipo de recursos demostradas por las investigaciones más relevantes llevadas a cabo hasta el momento. Es cierto que los problemas de creación de corpus relacionados con el tiempo y coste de su desarrollo son compartidos por otras tareas de PLN. Sin embargo, en RE textual, estos problemas son más desafiantes debido a que la detección de emoción en texto puede ser difícil incluso para los humanos, incrementando el tiempo y el coste de desarrollo, así como presentando problemas para obtener Acuerdo entre anotadores (IAA).

Con este contexto, esta tesis abordó la tarea de anotación de emociones en texto, proporcionando técnicas/metodologías automáticas y semiautomáticas con la intención de contribuir a abordarla eficientemente. Nos centramos en la anotación de emociones en texto escrito en Inglés para cualquier género textual, a nivel de oración y empleando un conjunto de categorías emocionales como etiquetas.

A continuación se exponen las principales conclusiones y contribuciones que aporta esta tesis que se pueden resumir en los siguientes puntos:

- **Análisis del estado de la cuestión con especial énfasis en la creación de recursos lingüísticos para la tarea de RE en texto**

Este análisis nos permitió verificar que los modelos de emoción basados en categorías (en particular, las emociones básicas propuestas por [Ekman \(1992\)](#)) son las más populares entre los enfoques computacionales, ya que la mayoría de ellos emplea este conjunto de emociones debido a su simplicidad al abordar el análisis desde el punto de vista humano y computacional. Además, el análisis cronológico de estos recursos nos permite observar una tendencia en la aplicación de técnicas semiautomáticas para la anotación de emociones. Esto se debe a dos hechos principales: las desventajas de la anotación manual y el crecimiento exponencial de la cantidad de información subjetiva en la Web

2.0 (blogs, redes sociales, servicios de microblogging, etc.).

- **Investigación en técnicas de anotación (Bootstrapping basado en [IL](#) y pre-anotación) para el etiquetado de emociones**

Esta investigación estudia diferentes métodos eficientes en términos de tiempo y costes para construir recursos. Teniendo en cuenta los problemas de anotación de emoción a la hora de abordar la tarea de manera eficiente y con alta fiabilidad, exploramos técnicas de anotación alternativas empleadas en otras disciplinas de PLN con el fin de mejorar la tarea de anotación de emociones. Estos métodos han demostrado su usabilidad y aplicabilidad en otras tareas de PLN que nos permiten considerarlos adecuados para abordar esta tarea, obteniendo mejoras en su proceso de desarrollo.

- **Propuesta y desarrollo de la técnica [IL](#) para anotación de emociones en texto**

Se presenta una técnica de bootstrapping basada en [IL](#), una técnica no supervisada que crea clasificadores a partir de datos no etiquetados. Esta es una de las características más atractivas para la anotación de emociones porque permite construir corpus emocionales donde se minimiza la influencia de anotadores humanos. Además, su simplicidad y flexibilidad para aplicarlo con otras categorías emocionales o géneros lo convierten en una técnica atractiva a considerar cuando el número de recursos etiquetados son escasos o demasiado costosos de desarrollar en grandes cantidades.

- **Propuesta y desarrollo de procesos de pre-anotación para abarcar la tarea de anotación de emociones (*EmoLabel*)**

EmoLabel es una metodología semiautomática en la que se lleva a cabo un proceso de pre-anotación con el objetivo de ayudar a los anotadores humanos a decidir cuál es la emoción dominante en cada oración. Si bien es cierto que esta propuesta no es tan eficiente como la técnica [IL](#) en términos de tiempo y coste, ya que requiere la participación de anotadores humanos, consideramos importante explorar técnicas de emoción alternativas en las que participaran humanos. Al fin y al cabo, estamos tratando de detectar emociones humanas. Además, *EmoLabel* proporciona adaptabilidad y versatilidad, permitiendo usar diferentes conjuntos de categorías de emociones, así como determinar el número

de categorías asociadas a cada oración.

- **Evaluación de la técnica IL**

Con el fin de verificar la idoneidad de IL para la anotación de emociones, se llevaron a cabo dos evaluaciones. Por un lado, se construyó un modelo de emoción a partir del corpus etiquetado automáticamente para evaluar la usabilidad de ese corpus. Por otro lado, la calidad de las anotaciones automáticas se evalúa a través de la medida de acuerdo entre el corpus desarrollado con nuestro enfoque (anotación automática) y el *gold standard* de los corpus Aman y Affective Text (anotación manual). Ambas evaluaciones nos permiten verificar la viabilidad del IL como una técnica para la anotación automática de emociones en texto, reduciendo el coste y el tiempo de desarrollo del mismo, ya que ambas evaluaciones obtuvieron resultados alentadores.

- **Evaluación de *EmoLabel***

Con el objetivo de realizar una evaluación en profundidad de *EmoLabel*, se requiere una evaluación intrínseca y extrínseca. El objetivo de la evaluación intrínseca es evaluar cuál es el mejor proceso de pre-anotación que se empleará en la segunda fase de *EmoLabel*. Para lograrlo, la evaluación se lleva a cabo comparando las emociones propuestas por cada método con las anotadas en el *gold standard* de cada uno de los corpus empleados en la evaluación. La evaluación extrínseca tiene como objetivo la evaluación del trabajo de los anotadores humanos en la segunda fase de *EmoLabel*. Con este fin, se lleva a cabo una tarea de anotación manual con tres anotadores. De acuerdo con la evaluación extrínseca, los experimentos realizados muestran los beneficios de los procesos de pre-anotación en el etiquetado de emociones, ya que los resultados en el tiempo de anotación muestran una ganancia de cerca de un 20% cuando se aplica el proceso de pre-anotación (Pre-ML) con respecto a sin pre-anotación (No-pre). Además, los experimentos realizados muestran que todas las tareas alcanzan un "acuerdo sustancial" y, por tanto, el proceso de pre-anotación no reduce el rendimiento del anotador, ni el acuerdo entre ellos (IAA).

Trabajo futuro

Como trabajos futuros de esta tesis, podemos destacar las siguientes líneas de investigación a corto, medio y largo plazo:

- **Mejorar la técnica IL**

Dado que el núcleo de la técnica de bootstrapping basada en IL es la clasificación inicial no supervisada y con el objetivo de reducir las oraciones falsas anotadas en el proceso inicial, una mejora sería explorar métodos alternativos para la creación de las semillas. Por ejemplo, considerar fenómenos como la negación o modificadores del lenguaje, agregar análisis de los emoticonos para enriquecer el proceso, o generar mediante sistemas de *Generación del Lenguaje Natural* (GLN) (del inglés Natural Language Generation, NLG) un conjunto de oraciones simple con contenido emocional (vocabulario emocional implícito) que posteriormente sería enriquecido con oraciones reales por similitud semántica. Además, dado los resultados obtenidos, consideramos interesante la aplicación de esta propuesta en otros géneros de la Web 2.0 como los mensajes de Twitter, publicaciones de Facebook, comentarios de noticias o foros donde hay un alto contenido emocional, ya que estos géneros permiten a las personas publicar mensajes para compartir información, opinión y emociones.

- **Mejorar la metodología *EmoLabel***

La investigación futura en *EmoLabel* se centrará en aprovechar al máximo el gran potencial de la anotación previa para crear grandes cantidades de datos anotados con emociones que permitan aplicar algoritmos de *Aprendizaje automático* (del inglés Machine Learning, ML) y/o *Aprendizaje profundo* (del inglés Deep Learning, DL) con el objetivo de construir sistemas de reconocimiento de emociones precisos. Para lograrlo, desarrollaremos la segunda fase de *EmoLabel* con más datos extraídos de los nuevos géneros de la Web 2.0 en plataformas de crowdsourcing con más anotadores. Además, los resultados logrados por el enfoque *supervisado* en el proceso de pre-anotación son prometedores y abren la posibilidad de reutilizar los modelos emocionales existentes como IBM Tone Analyzer⁹, cuyo sus valores-F están alrededor del

⁹<https://tone-analyzer-demo.ng.bluemix.net/>

60-70%, para pre-annotar nuevos datos.

- **Explorar ambas propuestas en otros idiomas**

Principalmente, el análisis automático de la emoción en el texto hasta ahora se ha centrado en el inglés debido a la falta de recursos emocionales en otros idiomas. Debido a ello y teniendo en cuenta los resultados logrados por nuestras propuestas (Bootstrapping basado en [IL](#) y *EmoLabel*), es de notable interés explorar más a fondo la aplicación de ellos en otros idiomas europeos como el español, el italiano o el holandés, así como otros idiomas asiáticos, como el bangla o el hindi para analizar cómo afectan las influencias culturales en la detección de emociones. Para ello, es importante que el desarrollo de estos recursos se lleve a cabo conjuntamente con personas nativas ya que la relación de una palabra con los conceptos emocionales puede depender de la ideología y, en general, de los aspectos culturales ([Strapparava, 2016](#)).

- **Analizar otras alternativas para la anotación de emociones en texto**

Si bien hemos evaluado dos técnicas efectivas de anotación, no descartamos y podría ser atractiva la evaluación de otras alternativas para la anotación de emociones como *Aprendizaje activo* (del inglés Active Learning, [AL](#)) o la aplicación de principios de diseño de juegos en la tarea. En cuanto a la estrategia de [AL](#), aplicaremos un método que utilice la estimación de confianza de los modelos de clasificación para determinar si una oración debe ser revisada por anotadores humanos o no. Esto nos permitirá reducir el número de oraciones utilizadas en la tarea de anotación manual. Para este fin, podemos usar [PAL](#) ([Skeppstedt et al., 2017](#)), una herramienta para pre-annotación y [AL](#). Acerca de la aplicación de principios de diseño de juegos a la tarea de anotación, la idea es que los anotadores humanos participen en el etiquetado de emociones sin darse cuenta de que están anotando un texto, con el objetivo de no afectar a su interpretación emocional del texto. Para lograrlo, sería interesante crear una aplicación móvil que pregunte al usuario sobre el contenido emocional de sus textos de una manera no intrusiva.

- **Estudiar cuáles son las categorías de emoción más apropiadas para el texto**

Centrándose en los modelos de emociones categóricas, las emociones básicas de Ekman (1992) son el conjunto más popular empleado en los enfoques computacionales. Sin embargo, este modelo de emoción se derivó originalmente de expresiones faciales y fisiológicas y, por lo tanto, no se basa en teorías del lenguaje. Durante el desarrollo de esta tesis, encontramos dificultades para detectar emociones como ASCO, MIEDO o SORPRESA en el texto, como muchos otros investigadores. Por lo tanto, un análisis de cuáles son las emociones expresadas en el texto, como el estudio llevado a cabo por Liew (2015), junto con una definición de un conjunto más representativo de categorías para el análisis textual parece ser prometedor y sería una gran contribución a la comunidad investigadora.

- **Estudiar los beneficios del análisis de emociones en otras disciplinas**

La mejora de los métodos de anotación de emociones nos permitirá construir una gran cantidad de datos con contenido emocional que se utilizarán para mejorar el rendimiento de los algoritmos de *Aprendizaje profundo* (DL), donde se requieren grandes cantidades de datos de *entrenamiento*. Además, la creación de un sistema de reconocimiento de emociones preciso para evaluar y representar las emociones de las personas a partir de sus comentarios en la Web social, junto con la información geográfica y temporal disponible en estos géneros, nos permitirá crear perfiles emocionales de usuario que aportarán beneficios sustanciales a diferentes tareas como la prevención del suicidio, identificación de casos de ciberacoso, o la *educación en línea* (del inglés *e-learning*).

References

- Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, 579–586.
- Aman, S. (2007). *Recognizing emotions in text* (PhD Thesis). University of Ottawa (Canada).
- Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *Text speech and dialogue* (pp. 196–205). Springer.
- Aman, S., & Szpakowicz, S. (2008). Using Roget’s Thesaurus for Fine-grained Emotion Recognition. In *International joint conference on natural language processing* (pp. 296–302).
- Antoine, J.-Y., Villaneau, J., & Lefevre, A. (2014). Weighted Krippendorff’s alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics (eacl 2014)* (pp. 550–559).
- Anusha, V., & Sandhya, B. (2015). A Learning Based Emotion Classifier with Semantic Text Processing. In E.-S. M. El-Alfy, S. M. Thampi, H. Takagi, S. Piramuthu, & T. Hanne (Eds.), *Advances in intelligent informatics* (pp. 371–382). Cham: Springer International Publishing.
- Arnold, M. B. (1960). *Emotion and personality*. New York: Columbia University Press.
- Aubur, D., Armantrout, R., Crystal, D., & Dirda, M. (2004). *Oxford American Writer’s Thesaurus*. Oxford University Press.

- Balabantaray, R. C., Mohammad, M., & Sharma, N. (2012). Multi-Class Twitter Emotion Classification: A New Approach. *International Journal of Applied Information Systems (IJ AIS)*, 4(1), 48–53.
- Banea, C., Chen, D., Mihalcea, R., Cardie, C., & Wiebe, J. (2014). SimCompass: Using Deep Learning Word Embeddings to Assess Cross-level Similarity. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)* (pp. 560–565).
- Banea, C., Mihalcea, R., & Wiebe, J. (2008, may). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In *Proceedings of the sixth international conference on language resources and evaluation (lrec-08)* (pp. 2764–2767). Marrakech, Morocco: European Language Resources Association (ELRA).
- Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the fourth conference on language resources and evaluation (lrec-04)* (pp. 1313–1316). European Language Resources Association.
- Boldrini, E., & Martínez-Barco, P. (2012). *EMOTIBLOG: A model to Learn Subjective Information Detection in the New Textual Genres of the Web 2.0-Multilingual and Multi-Genre Approach-* (PhD Thesis). University of Alicante.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy & Experimental Psychiatry*, 25(1), 49–59.
- Bradley, M. M., & Lang, P. J. (1999a). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings* (Tech. Rep.). The Center for Research in Psychophysiology, University of Florida.
- Bradley, M. M., & Lang, P. J. (1999b). *International affective digitized sounds (IADS): stimuli, instruction manual and affective ratings* (Tech. Rep. No. B-2). Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.
- Bradley, M. M., & Lang, P. J. (2007). *Affective norms for English text (ANET): Affective ratings of text and instruction manual.* (Tech. Rep.).

- The Center for Research in Psychophysiology, University of Florida.
- Brysbaert, M., & New, B. (2009, nov). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.
- Bu, Q., Simperl, E., Zerr, S., & Li, Y. (2016). Using microtasks to crowdsource DBpedia entity classification: A study in workflow design. *Semantic Web Journal*, 1–16.
- Buechel, S., & Hahn, U. (2017). EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics (eacl 2017)* (pp. 578–585).
- Cambria, E., Fu, J., Bisio, F., & Poria, S. (2015). AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis. In *Proc. of the twenty-ninth {aai} conference on artificial intelligence, january 25-30, 2015, austin, texas, {usa.}* (pp. 508–514).
- Cambria, E., Olsher, D., & Rajagopal, D. (2014). SenticNet 3: A Common and Common-sense Knowledge Base for Cognition-driven Sentiment Analysis. In *Proceedings of the twenty-eighth aai conference on artificial intelligence* (pp. 1515–1521). AAAI Press.
- Canales, L., Daelemans, W., Boldrini, E., & Martínez-Barco, P. (2017). Towards the Improvement of Automatic Emotion Pre-annotation with Polarity and Subjective Information. In *Proceedings of the 11th biennial recent advances in natural language processing conference (ranlp 2017)* (p. 7).
- Chaffar, S., & Inkpen, D. (2011). Using a Heterogeneous Dataset for Emotion Analysis in Text. In *Proceedings of the 24th canadian conference on advances in artificial intelligence* (pp. 62–67). Berlin, Heidelberg: Springer-Verlag.
- Cherry, C., Mohammad, S. M., & De Bruijn, B. (2012). Binary Classifiers and Latent Sequence Models for Emotion Detection in Suicide Notes.

- Biomedical informatics insights*, 5(Suppl 1), 147–154.
- Chou, W.-C., Tsai, R. T.-H., Su, Y.-S., Ku, W., Sung, T.-Y., & Hsu, W.-L. (2006). A Semi-automatic Method for Annotating a Biomedical Proposition Bank. In *Proceedings of the workshop on frontiers in linguistically annotated corpora 2006* (pp. 5–12). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Choudhury, M. D., Gamon, M., & Counts, S. (2012). Happy, Nervous or Surprised? Classification of Human Affective States in Social Media. In *Proceedings of the 6th international aaai conference on weblogs and social media*.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Collins, M., & Singer, Y. (1999). Unsupervised Models for Named Entity Classification. In *In proceedings of the joint sigdat conference on empirical methods in natural language processing and very large corpora* (pp. 100–110).
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 18 (1)(1), 32–80.
- Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online Passive-Aggressive Algorithms. *Journal of Machine Learning Research*, 7, 551–585.
- Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In *Advances in information retrieval* (pp. 693–696).
- Darwin, C. (1998). *The Expression of Emotions in Man and Animals*, ed. P. Ekman (Orig. publ ed.; HarperCollins, Ed.). London.
- De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *J. Mach. Learn. Res.*, 13(1), 2063–2067.
- de Albornoz, J., Plaza, L., & Gervás, P. (2012). SentiSense: An easily

- scalable concept-based affective lexicon for Sentiment Analysis. In *The 8th international conference on language resources and evaluation (lrec 2012)*.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Deng, L., & Yu, D. (2014). *Deep Learning: Methods and Applications*. Now Publishers Inc.
- Denis, P., & Sagot, B. (2009). Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort. In *Proceeding of the 23rd pacific asia conference on language, information and computation* (pp. 110—119).
- Desmet, B., & Hoste, V. (2013, nov). Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16), 6351–6358.
- Dinu, G., & Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. *Proceeding of the 3th International Workshop Conference on Learning Representations (ICLR 2015)*.
- Ekman, P. (1971). *Universals and Cultural Differences in Facial Expressions of Emotion* (Vol. 19). University of Nebraska Press.
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 169–200.
- Erk, K., & Pado, S. (2006). Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of {lrec} 2006*. Genoa, Italy.
- Farzindar, A., & Inkpen, D. (2015). *Natural Language Processing for Social Media*. Morgan & Claypool Publishers.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Fort, K., Nazarenko, A., & Rosset, S. (2012). Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis. In *Proceeding of international conference on computational linguistics (coling 2012)* (pp. 895–910).

- Fort, K., & Sagot, B. (2010). Influence of pre-annotation on pos-tagged corpus development. In *Proc. of the fourth linguistic annotation workshop* (pp. 56–63). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Ganchev, K., Pereira, F., Mandel, M., Carroll, S., & White, P. (2007). Semi-automated Named Entity Annotation. In *Proceedings of the linguistic annotation workshop* (pp. 53–56). Stroudsburg, PA, USA: Association for Computational Linguistics.
- García Pablos, A., Cuadros Oller, M., & Rigau Claramunt, G. (2015). *Unsupervised Word Polarity Tagging by Exploiting Continuous Word Representations* (Vol. 55).
- Généreux, M., & Evans, R. (2006). Distinguishing affective states in weblogs. *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs, Stanford, CA*.
- Ghazi, D. (2016). *Identifying Expressions of Emotions and Their Stimuli in Text* (PhD Thesis). University of Ottawa.
- Ghazi, D., Inkpen, D., & Szpakowicz, S. (2010). Hierarchical Versus Flat Classification of Emotions in Text. In *Proceedings of the naacl hlt 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 140–146). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Gill, A. J., Gergle, D., French, R. M., & Oberlander, J. (2008). Emotion rating from short blog texts. In *26th annual sigchi conference on human factors in computing systems, conference proceedings, chi 2008* (pp. 1121–1124).
- Gliozzo, A., & Strapparava, C. (2009). *Semantic Domains in Computational Linguistics*. Springer-Verlag Berlin Heidelberg.
- Gliozzo, A., Strapparava, C., & Dagan, I. D. O. (2009). Improving Text Categorization Bootstrapping via Unsupervised Learning. *ACM Transactions on Speech and Language Processing*, 6(1).
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD*

Explorations, 11(1), 10–18.

- Hasan, M., Rundensteiner, E., & Agu, E. (2014). EMOTEX: Detecting Emotions in Twitter Messages. In *Ase bigdata/socialcom/cybersecurity conference* (pp. 27–31).
- Henriksson, A., Kvist, M., Dalianis, H., & Duneld, M. (2015). Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics*, 57, 333–349.
- Imbir, K. K. (2015, sep). Affective norms for 1,586 polish words (ANPW): Duality-of-mind approach. *Behavior Research Methods*, 47(3), 860–870.
- Imbir, K. K. (2016). Affective Norms for 718 Polish Short Texts (ANPST): Dataset with Affective Ratings for Valence, Arousal, Dominance, Origin, Subjective Significance and Source Dimensions. *Frontiers in Psychology*, 7, 1030.
- Izard, C. E. (1971). *The face of emotion* (New York: Appleton-Century-Crofts., Ed.).
- Jakobson, R. (1960). Linguistics and Poetics. In *Style in language* (pp. 350—377). MIT Press Cambridge, MA, USA ©1997.
- Kaplan, A. M., & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 54(2), 105–113.
- Kennedy, A., & Inkpen, D. (2006). Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22.
- Kenter, T., & de Rijke, M. (2015). Short Text Similarity with Word Embeddings. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 1411–1420). New York, NY, USA: ACM.
- Keshtkar, F., & Inkpen, D. (2010). A Corpus-based Method for Extracting Paraphrases of Emotion Terms. In *Proceedings of the naacl hlt 2010 workshop on computational approaches to analysis and generation of*

- emotion in text* (pp. 35–44). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Kim, S. M. (2011). *Recognising Emotions and Sentiments in Text* (PhD Thesis). University of Sydney.
- Ko, Y., & Seo, J. (2004a). Learning with Unlabeled Data for Text Categorization Using Bootstrapping and Feature Projection Techniques. In *Proceedings of the 42nd annual meeting on association for computational linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Ko, Y., & Seo, J. (2004b). Using the feature projection technique based on a normalized voting method for text classification. *Information Processing & Management*, 40(2), 191–208.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802–5805.
- Krcadinac, U., Pasquier, P., Jovanovic, J., & Devedzic, V. (2013). Synesketch: An Open Source Library for Sentence-Based Emotion Recognition. *IEEE Transactions on Affective Computing*, 4(3), 312–325.
- Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Sage Publications.
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174.
- Lang, P., Bradley, M., & Cuthbert, B. (1999). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*.
- Lapesa, G., & Evert, S. (2014). A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection. *Transactions of the Association for Computational Linguistics (TACL)*, 2, 531–545.
- Levenson, R. W. (2003). Autonomic specificity and emotion. In *Handbook of affective sciences*. (pp. 212–224). New York, NY, US: Oxford University Press.

-
- Li, L., Wang, M., Zhang, L., & Wang, H. (2014). Learning Semantic Similarity for Multi-label Text Categorization. In X. Su & T. He (Eds.), *Proceeding of workshop on chinese lexical semantics* (pp. 260–269). Cham: Springer International Publishing.
- Li, M., Long, Y., Qin, L., & Li, W. (2016, may). Emotion Corpus Construction Based on Selection from Hashtags. In N. C. C. Chair) et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)*. Paris, France: European Language Resources Association (ELRA).
- Liew, J. S. Y. (2015). Discovering Emotions in the Wild: An Inductive Method to Identify Fine-grained Emotion Categories in Tweets. In *Proceedings of the 28th international florida artificial intelligence research society conference (flairs conference)* (pp. 317–322).
- Liew, J. S. Y., Turtle, H. R., & Liddy, E. D. (2016). EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis. In *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)* (pp. 1149–1156).
- Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., . . . Solti, I. (2014). Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association : JAMIA*, 21(3), 406–413.
- Liu, B., Li, X., Lee, W. S., & Yu, P. S. (2004). Text classification by labeling words. In *Proceedings of the conference on natural language processing and information extraction* (pp. 425—430).
- Louviere, J. J. (1991). *Best-worst scaling: A model for the largest difference judgments*.
- Magnini, B., & Cavaglia, G. (2000). Integrating Subject Field Codes into WordNet. In *Lrec* (pp. 1413–1418). European Language Resources Association.
- Manning, C. D., Bauer, J., Finkel, J., Bethard, S. J., Surdeanu, M., &

- McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60).
- Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.
- Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014). SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)* (pp. 1–8).
- Mccallum, A., & Nigam, K. (1999). Text Classification by Bootstrapping with Keywords, EM and Shrinkage. In *Workshop on unsupervised learning in natural language processing (acl)* (pp. 52–58).
- McDonald, R., Crammer, K., & Pereira, F. (2005). Online Large-margin Training of Dependency Parsers. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 91–98). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual. *Current Psychology*, 15(4), 505–525.
- Mihalcea, R., & Strapparava, C. (2012). Lyrics, Music, and Emotions. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 590–599). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the First International Conference on Learning Representations (ICLR 2013)*, 1–12.

- Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), 39–41.
- Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp* (pp. 1003–1011).
- Mishne, G. (2005). Experiments with Mood Classification in Blog Posts. In *Proceedings of the 1st workshop on stylistic analysis of text for information access*.
- Mohammad, S. M. (2012a). # Emotional tweets. In *Proceedings of the first joint conference on lexical . . .* (pp. 246–255). Montréal, Canada: Association for Computational Linguistics.
- Mohammad, S. M. (2012b). 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* (pp. 587–591). Montréal, Canada: Association for Computational Linguistics.
- Mohammad, S. M., & Bravo-Marquez, F. (2017). WASSA-2017 - Shared Task on Emotion Intensity. In *Proceedings of the workshop on computational approaches to subjectivity, sentiment and social media analysis (wassa)*. Copenhagen, Denmark.
- Mohammad, S. M., & Kiritchenko, S. (2015). Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*, 31(2), 301–326.
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a WordEmotion Association Lexicon. *Computational Intelligence*, 29(3), 436–465.
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.
- Munezero, M., Kakkonen, T., Sedano, C. I., Sutinen, E., & Montero, C. S. (2013, sep). EmotionExpert: Facebook game for crowdsourcing annotations for emotion detection. In *2013 ieee international games*

- innovation conference (igic)* (pp. 179–186).
- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2010). AM: Textual Attitude Analysis Model. In *Proceedings of the naacl hlt 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 80–88). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). Affect Analysis Model: Novel Rule-based Approach to Affect Sensing from Text. *Natural Language Engineering*, 17(1), 95–135.
- Nowson, S., Oberlander, J., & Gill, A. (2005). Weblogs, genres and individual differences. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1666–1671).
- Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.
- Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological review*, 97(3), 315.
- Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), 71–106.
- Parrott, W. G. (Ed.). (2001). *Emotions in social psychology: Essential readings*. New York, NY, US: Psychology Press.
- Pearson, K. (1956). *Early Statistical Papers*. University Press.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12, 2825–2830.
- Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count: LIWC2007, Operator's manual*. Austin, TX: LIWC.net.
- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. In *Proceeding of international conference on empirical methods in natural language processing (emnlp)* (Vol. 14, pp. 1532–1543).
- Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., . . . Brew, C. (2012, jan). Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, 5(Suppl 1), 3–16.
- Picard, R. W. (1997). *Affective computing*. MIT Press Cambridge, MA, USA ©1997.
- Platt, J. (1999). Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. In *Proceeding of international conadvances in neural information processing systems* (pp. 557–563).
- Plutchik, R. (1962). *The Emotions*. New York: Random House.
- Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3–33).
- Plutchik, R. (1994). *The psychology and biology of emotion*. New York: Harper Collins.
- Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., & Bandyopadhyay, S. (2013). Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2), 2–9.
- Predoiu, C., Dascalu, M., & Trausan-Matu, S. (2014, sep). Trust and user profiling for refining the prediction of reader’s emotional state induced by news articles. In *2014 roedunet conference 13th edition: Networking in education and research joint event renam 8th conference* (pp. 1–6).
- Preotiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). *Modelling Valence and Arousal in Facebook posts*. NAACL.
- Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics* (pp. 482–491). Stroudsburg, PA, USA: Association for Computational

Linguistics.

- Qadir, A., & Riloff, E. (2013). Bootstrapped Learning of Emotion Hashtags {#}hashtags4you. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, wassa@naacl-hlt 2013, 14 june 2013, atlanta, georgia, {usa}* (pp. 2–11).
- Read, J. (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the acl student research workshop* (pp. 43–48). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Rehbein, I., Ruppenhofer, J., & Sporleder, C. (2009). Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proceedings of the third linguistic annotation workshop* (pp. 19–26). Suntec, Singapore: Association for Computational Linguistics.
- Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In *Proceedings of the seventh conference on natural language learning at hlt-naacl 2003 - volume 4* (pp. 25–32). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. In N. Calzolari et al. (Eds.), *Lrec* (pp. 3806–3813). European Language Resources Association (ELRA).
- Rothe, S., Ebert, S., & Schütze, H. (2016). Ultradense Word Embeddings by Orthogonal Transformation. *arXiv preprint arXiv:1602.07572*.
- Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.
- Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695–729.
- Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6), 1061–1086.

- Sintsova, V., Musat, C.-C., & Pu, P. (2013). Fine-Grained Emotion Recognition in Olympic Tweets Based on Human Computation. *4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.
- Sintsova, V., & Pu, P. (2016). Dystemo: Distant Supervision Method for Multi-Category Emotion Recognition in Tweets. *ACM Trans. Intell. Syst. Technol.*, 8(1), 13:1—13:22.
- Skeppstedt, M., Kvist, M., Nilsson, G. H., & Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, 49, 148–158.
- Skeppstedt, M., Paradis, C., & Kerren, A. (2017). PAL, a tool for pre-annotation and active learning. *Journal for Language Technology and Computational Linguistics*, 31(1), 91–110.
- Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 35th annual sigir conference*.
- Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Staiano, J., & Guerini, M. (2014). Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 427–433). Baltimore, Maryland: Association for Computational Linguistics.
- Stevenson, R. A., Mikels, J. A., & James, T. W. (2007, nov). Characterization of the Affective Norms for English Words by discrete emotional categories. *Behavior Research Methods*, 39(4), 1020–1024.
- Strapparava, C. (2016). Emotions and NLP: Future Directions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, senti-*

References

- ment and social media analysis* (p. 180). Association for Computational Linguistics.
- Strapparava, C., & Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In *Proceedings of the 4th international workshop on semantic evaluations (semeval-2007)* (pp. 70–74).
- Strapparava, C., & Mihalcea, R. (2014). Affect Detection in Texts. In Rafael Calvo, Sidney D’Mello, Jonathan Gratch & A. Kappas (Eds.), *The oxford handbook of affective computing* (p. 40).
- Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: an Affective Extension of WordNet. In *4th international conference on language resources and evaluation* (pp. 1083–1086).
- Suero Montero, C., & Suhonen, J. (2014). Emotion Analysis Meets Learning Analytics: Online Learner Profiling Beyond Numerical Data. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research* (pp. 165–169). New York, NY, USA: ACM.
- Suttles, J., & Ide, N. (2013). Distant Supervision for Emotion Classification with Discrete Binary Values BT - Computational Linguistics and Intelligent Text Processing. In A. Gelbukh (Ed.), *14th international conference on intelligent text processing and computational linguistics (cicling 2013)* (pp. 121–136). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Thayer, R. E. (1989). *The Biopsychology of Mood and Arousal*. Oxford University Press USA.
- Tomkins, S. S. (1984). Affect theory. *Approaches to emotion*, 163, 163–195.
- Vaassen, F. (2014). *Measuring Emotion: Exploring the feasibility of automatically classifying emotional text* (PhD Thesis). University of Antwerp, Antwerp, Belgium.
- Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, 50(3), 289–335.
- Villalón, J., Kearney, P., Calvo, R. A., & Reimann, P. (2008). Glosser:

-
- Enhanced Feedback for Student Writing Tasks. In *Proceeding of the eighth IEEE international conference on advanced learning technologies* (pp. 454–458).
- Wallbott, H., & Scherer, K. R. (1986). How universal and specific is emotional experience? Evidence from 27 countries on five continents. *Social Science Information*, 25(4), 763–795.
- Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing Twitter "Big Data" for Automatic Emotion Identification. In *International conference on social computing (socialcom)*.
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 1191–1207.
- Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form*.
- Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological bulletin*, 98(2), 219–235.
- Whissel, C. (1989). The dictionary of affect in language. In R. Plutchik & H. Kellerman (Eds.), *The measurement of emotions* (pp. 113–131). Academic Press.
- Wundt, W. M. (1905). *Grundzüge der physiologischen Psychologie*. Leipzig: Engelmann.
- Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd annual meeting on association for computational linguistics* (pp. 189–196). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., ... Zhang, X.-J. (2016). Building Chinese Affective Resources in Valence-Arousal Dimensions. In *Hlt-naacl 2016 - proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies*. (pp. 540–545).