Technical University of Crete
Electronic and Computer Engineering

"Affective analysis and Modeling of Spoken Dialogue Transcripts"

by

Elisavet Palogiannidi

thesis committee

Thesis supervisor:	Associate Professor Polychronis Koutsakis
Committee member:	Associate Professor Alexandros Potamianos
Committee member:	Associate Professor Aikaterini Mania

A thesis submitted for the degree of:
Master of Science

July 2016

Abstract

At this thesis we propose affective models for the emotional analysis of lexical units in various granularity levels. We propose and evaluate the use of an affective-semantic model to expand the affective lexica of German, Greek, English, Spanish and Portuguese. Motivated by the assumption that semantic similarity implies affective similarity, we use word level semantic similarity scores as semantic features to estimate their corresponding affective scores. Various context-based semantic similarity metrics are investigated using contextual features that include both words and character n-grams. The model produces continuous affective ratings in three dimensions (valence, arousal and dominance) for all five languages, achieving consistent performance. We achieve classification accuracy (valence polarity task) between 85% and 91% for all five languages. For morphologically rich languages the proposed use of character n-grams is shown to improve performance. Moreover, we created the first Greek affective lexicon, translating the words of the English affective lexicon ANEW and assigning them to native speakers for affective annotation. It contains human ratings for the three continuous affective dimensions of valence, arousal and dominance for 1034 words.

Motivated by recent advances in the area of Compositional Distributional Semantic Models (CDSMs), we propose a compositional approach for estimating continuous affective ratings for adjective-noun (AN) and noun-noun (NN) pairs. The ratings are computed for the three basic dimensions of continuous affective spaces, namely, valence, arousal and dominance. We propose that similarly to the semantic modification that underlies CDSMs, affective modification may occur within the framework of affective spaces, especially when the constituent words of the linguistic structures under investigation form modifier-head pairs (e.g., AN and NN). The affective content of the entire structure is determined from the interaction between the respective constituents, i.e., the affect conveyed by the head is altered by the modifier. In addition, we investigate the fusion of the proposed model with the semantic-affective model proposed in literature applied both at word- and phrase-level. The automatically computed affective ratings were evaluated against human ratings in terms of correlation. The most accurate estimates are achieved via fusion and absolute performance improvement up to 5% and 4% is reported for NN and AN, respectively.

We also investigate text based models for the affective analysis of sentences that are mainly based on affective features. We investigate various datasets including news headlines, movie subtitles, Twitter status updates and spoken dialogue transcriptions and the best (state-of-the-art) performance was obtained for Twitter (0.82 average recall over positive and negative classes).

Περίληψη

Σε αυτή την εργασία προτείνουμε υπολογιστικά μοντέλα για τη συναισθηματική ανάλυση λεξικών μονάδων διαφόρων επιπέδων, ξεκινώντας από λέξεις και καταλήγοντας σε προτάσεις. Προτείνουμε και αξιολογούμε τη χρήση ενός συναισθηματικού-σημασιολογικού μοντέλου που στοχεύει στην επέκταση συναισθηματικών λεξικών διαφόρων γλωσσών όπως Γερμανικά, Ελληνικά, Αγγλικά, Ισπανικά και Πορτογαλικά. Εμπνευσμένοι από την υπόθεση ότι η σημασιολογική ομοιότητα μπορεί να μετατραπεί σε συναισθηματική ομοιότητα, χρησιμοποιούμε σημασιολογικές αποστάσεις ανάμεσα σε λέξεις ως χαρακτηριστικά, προκειμένου να εκτιμήσουμε τις συναισθηματικές ετικέτες των λέξεων. Ερευνούμε διάφορες σημασιολογικές μετρικές που βασίζονται στα συμφραζόμενα, χρησιμοποιώντας χαρακτηριστικά που βασίζονται σε διαφορετικά είδη συμφραζομένων όπως λέξεις ή συνεχόμενους χαρακτήρες. Το μοντέλο παράγει συναισθηματικές ετικέτες σε τρεις συνεχείς συναισθηματικές διαστάσεις (valence, arousal, dominance) σε καθεμία από τις πέντε γλώσσες, πετυχαίνοντας υψηλή απόδοση πάντα. Συγκεκριμένα η απόδοση ταξινόμησης που πετυχαίνουμε κυμαίνεται μεταξυ 85% και 90%. Η χρήση των συνεχόμενων χαρακτήρων ως χαρακτηριστικά συμφραζομένων φαίνεται πως ωφέλησε τις μορφολογικά πλούσιες γλώσσες. Προκειμένου να εφαρμόσουμε το μοντέλο στην Ελληνική γλώσσα χρειάστηκε να δημιουργήσουμε το δικό μας συναισθηματικό λεξικό, μεταφράζοντας τις λέξεις του αντίστοιχου Αγγλικού. Αυτό το λεξικό είναι το πρώτο που δημιουργήθηκε για τα Ελληνικά και περιλαμβάνει ετικέτες για τις τρεις συναισθηματικές διαστάσεις που αναφέραμε για 1034 λέξεις.

Εμπνευσμένοι από τα πρόσφατα πλεονεκτήματα που αποδείχτηκε πως έχουν τα κατανεμημένα σημασιολογικά μοντέλα σύνθεσης (Compositional Distributional Semantic Models (CDSMs) προτείνουμε ένα μοντέλο σύνθεσης για την εκτίμηση συναισθηματικών ετικετών για τις τρεις διαστάσεις, ζευγαριών λέξεων που αποτελούνται από ένα επίθετο που ακολουθείται από ένα ουσιαστικό (ΕΟ) ή από ουσιαστικό που ακολουθείται από ένα άλλο ουσιαστικό (ΟΟ). Προτείνουμε πως όμοια με τη σημασιολογική τροποποίηση που παρατηρείται στα CDSMs είναι δυνατό να συμβεί και συναισθηματική τροποποίηση, ειδικά όταν οι λέξεις που σχηματίζουν τα ζευγάρια που εξετάζουμε είναι σχηματίζουν δομές τροποποιητή-κεφαλής. Το συναισθηματικό περιεχόμενο του ζευγαριού καθορίζεται από την αλληλεπίδραση μεταξύ των δυο λέξεων, δηλαδή το συναισθηματικό περιεχόμενο της κεφαλής διαφοροποιείται κατά τρόπο που καθορίζεται από τον τροποποιητή. Επιπροσθέτως, εξετάζουμε το συνδυασμό του προτεινόμενου μοντέλου με τα σημασιολογικά-συναισθηματικά μοντέλα που προτείνονται στη βιβλιογραφία για λέξεις και ζευγάρια λέξεων. Οι ετικέτες που υπολογίζονται από αυτό το μοντέλο αξιολογούνται με βάση αντίστοιχες που έχουν προκύψει από ανθρώπους και η απόδοση εκτιμάται μετρώντας τη μεταξύ τους συσχέτιση. Οι πιο ακριβείς εκτιμήσεις προκύπτουν συνδυάζοντας τα διάφορα μοντέλα πετυχαίνοντας απόλυτη βελτίωση έως και 5% για ΟΟ ζευγάρια ή 4% για ΕΟ ζευγάρια.

Ερευνούμε επίσης μοντέλα για τη συναισθηματική ανάλυση προτάσεων που βασίζονται κατά κύριο λόγο σε συναισθηματικά χαρακτηριστικά. Χρησιμοποιούμε διάφορα σύνολα δε-

δομένων που καλύπτουν αρκετές περιπτώσεις γραπτού λόγου όπως τίτλοι νέων, υπότιτλοι ταινιών, ενημερώσεις καταστάσεων κοινωνικών δικτύων και καταγραφές ομιλούμενου λόγου. Η καλύτερη απόδοση state-of-the-art επιτεύχθηκε για τη συναισθηματική ανάλυση ενημερώσεων καταστάσεων στο κοινωνικό δίκτυο twitter (0.82 μέσο recall μεταξύ της θετικής και της αρνητικής ομάδας).

Acknowledgments

I would like to express my sincere gratitude, for the second time, to my advisor Alexandros Potamianos for the guidance, the continuous support and the immense knowledge he disseminated to me. I can not imagine a better mentor for me. His continuous guidance did not let any working day to pass without me learning something new. Moreover I would like to deeply thank Polychronis Koustakis for his endless help and support during the last years and to Aikaterini Mania for her participation to my committee.

I would also like to thank all the group members in NTUA and TUC and all my colleagues for the stimulating discussions and for working together. Special thanks go to Dr. Elias Iosif, Athanasia Kolovou and Fenia Christopoulou. I want to thank Nikolaos Malandrakis for sharing with me part of his work and for the many discussions about affective computing.

I gratefully acknowledge the financial support I received from the PortDial project supported by the EU Seventh Framework Programme (FP7), grant number 296170, the SpeDial project supported by the EU Seventh Framework Programme (FP7), grant number 611396 and the BabyRobot project supported by the EU Horizon 2020 Programme, grant number: 687831.

Nothing would be easy if I hadn't by my side some great people, my best friends. I want to thank and to express my endless love to Georgia and Eleni and to my sister Anastasia for the beautiful moments we have shared together. Of course, I would like to thank my parents for their endless love and support and for believing in me.

Although this thesis is related with the emotional analysis of text, I can't find any words that compound my feelings about my life companion Konstantinos. Thank you, for being always by my side, making my daily life beautiful and helping me overcoming the problems and becoming a better person.

"You took my hand and you showed me the way.
Thank you!"

To Tsampika Karakiza

Contents

List of Figures

List of Tables

List of abbreviations

NLP	Natural Language Processing
DSM	Distributional Semantic Models
CDSMs	Compositional Distributional Semantic Models
VSM	Vector Space Model
SAM	Semantic-Affective Model
NRE	Name Entity Recognition
LSE	Least Squares Estimation
RR	Ridge Regression

Chapter 1

Introduction

*"Human behavior flows from three main sources: desire, **emotion**, and knowledge."*
Plato

1.1 The role of emotions in our lives

Emotions play an important role in the way humans think and behave in daily life [42, 44, 62]. Emotions comprise of three critical components, the subjective that allows humans to experience emotion, the physiological that is responsible for the body reactions and the expressive component that handles how humans behave in response to emotion. Emotions can motivate humans to take action and help them service, avoid danger and thrive. They are highly associated with the decision making process [17, 179], since a complete cognitive analysis can be difficult due to uncertainty and ambiguity. Moreover, emotional expressions allow people to understand each other, and as result they contribute to healthy relationships.

1.1.1 The experience of emotion

Emotions can be short-lived or long-lasting depending on importance and rumination of each event [199]. Based on the temporal aspect *emotion* is indicated by very short term, *mood* by long term and *personality* by very long term. Emotions are distinguished into primary (fast response to a situation, e.g., fear, anger, sadness and happiness) or secondary (caused by primary emotions or after more complex chains of thought). Primary emotions may disappear as fast as they appear and they can also be felt as secondary emotions. Secondary emotions can be experienced even after weeks or months from the end of an event, each time we think about it. According to [18], secondary emotions are more complex than primary ones and their appraisal depends much more on the situational context and human's memory than that of primary emotions (secondary emotions are more dependent on human's cognitive reasoning abilities). According to [42], a "thought process" is involved in the elicitation of the secondary emotions, i.e., human brain evaluates the actual stimulus against previously acquired experiences and online generated expectations.

Emotions are mostly determined by one of the oldest parts of our brain, the limbic system, which is shown in Fig. 1.1 (a). Amygdala is one of the most important components

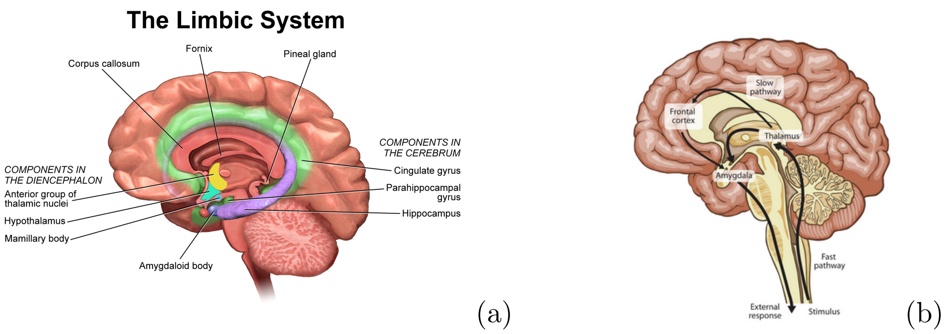


Figure 1.1: (a) The limbic system and (b) the fast and the slow path of emotion experience

of limbic system, which is responsible for the experience of emotions. Amygdala has nerves that send incoming messages from all of our senses, and from organs, throughout our body. Incoming from various stimuli messages reach to the amygdala either from the fast or from the slow pathway. Response to primary emotions, e.g., fear, is determined by the fast pathway through the limbic system. When we feel that we are in danger, thalamus activates and sends an immediate message to the amygdala. Secondary emotions are mainly determined by the slow pathway through the frontal lobes in the cortex. When we feel a secondary emotion, e.g., jealousy, information travels from thalamus to the frontal lobes for cognitive analysis and then finally to amygdala. The fast and the slow pathways are shown in Fig. 1.1 (b). Amygdala have also primary roles in formation and storage of memories associated with emotional events. Thus, our experiences may be interpreted in order to create more complex emotional experiences [109].

Emotions can be viewed from various perspectives such as Darwinian, in which emotions serve evolutionary function (instinct), Cognitive/Appraisal, in which emotions rise through complex cognitive evaluation of world and multiple "modules" represent different layers of perception. However, Neurobiology shows evidence of both (fast-slow (pessimistic-cognitive)) [174]. [174] claims that emotions are emergent processes and they require dynamic computational architecture. They also present a brief survey of emotion theories. [175] describes three theoretical traditions of emotion, i.e., *basic emotion theories* (an event triggers a specific affect programme corresponding to one of the primary emotions and producing characteristic expression patterns and physiological response configurations), *constructivist emotion theories* (although emotion is built upon one or more biological behaviour systems, the functional significance of emotion is to be found primarily within the sociocultural system) and *appraisal theories of emotion* (our interpretation of a situation causes an emotional response that is based on that interpretation). In the absence of physiological arousal, humans decide what to feel after interpreting what has just happened. The two important things in this process are: whether the event is interpreted as good or bad and what is believed to be the cause of the event. In primary appraisal, we consider how the situation affects our personal well-being. In secondary appraisal we consider how we might cope with the situation [56, 182]. According to Conceptual Act Theory [16,94], a typical emotion experience starts with the human experiencing a stimulus and then collecting information about the context as well as visceral signals of valence (e.g. pleasurable feelings) and arousal (e.g. heart rate, clammy skin).

Emotions can still be perceived in text and can be elicited by its content and form [154]. Readers experience emotion by placing themselves in the position of characters of a narrative and imagining their own emotional reaction.

1.2 Affective computing

Emotions are expressed and understood by humans, so the question is how can emotions be detected by machines. Computational models that are able to understand not only what people say but also *how* the information is being conveyed may help a machine to pass the "Turing Test" [194]. The study and development of systems that are able to perceive and express emotions is broadly known with the term *Affective computing*. *Affect* is used to describe topics such as feelings, emotion, mood, however it is commonly used interchangeably with emotion. Positive encompasses all good emotions, while negative encompasses all bad emotions. Feelings are sensations that have been checked against previous experiences, they are personal and biographical and they can be expressed through speech, facial and body gestures. Emotions are short term feelings that are elicited by any event, while mood is a longer feeling state, and it is not always easy to identify its cause.

Emotion can be provoked from various stimuli, in all modalities (speech, facial and body gestures, text), and can be associated with numerous temporal aspects. The aforementioned aspects of emotion motivated the development of a wide range of computational tools such as opinion mining and sentiment analysis systems on text modality [145]. In order to increase the accuracy and reliability of estimations, a plethora of computational systems that aim to the estimate emotion exploiting more than one modality have been developed [30, 86, 146, 157, 212].

In this thesis, we focus on affective computational systems that aim to predict emotion from text modality. Affective analysis of text has become a very popular task due to the fact that during the last decades, people select to exchange short textual messages for their daily communication with each other. The majority of affective text applications are related to social networks [32, 41, 45, 128, 167, 169, 170].

1.2.1 Emotion Representation

The key characteristic of each affective computational system is the emotional representation that it follows. Emotions can be conceptualized either in *discrete* or in *dimensional* view as depicted in Fig. 1.2.

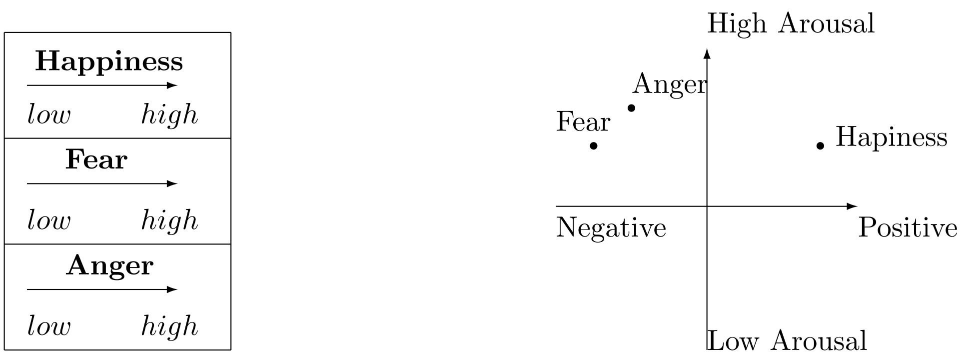


Figure 1.2: Example of Discrete Vs Dimensional (Continuous) Emotion Representation

Dimensional emotion representation is expressed in the seventeenth century by Rene Descartes. The term "passion" is used to express emotions and he claims that complex passions can be constructed from six simple passions, i.e., *wonder, love, hatred, desire, joy, sadness* [48]. The theory of discrete emotions is also adopted by Silvan Tomkins and, specifically, he insists that there are only nine affects, biologically based. Tomkins describes emotions in pairs: the first part of the pair names the mild manifestation and the second the more intense. First he described the basic six emotions, i.e., interest-excitement, enjoyment-joy, surprise-startle, distress-anguish, anger-rage, and fear-terror. Shame-humiliation was evolved later and the final two affects he described are dissmell [1] and disgust [192]. Paul Ekman determined the six basic emotions that are common among people in all cultures. The six basic emotions are: *anger, happiness, surprise, disgust, sadness* and $fear$. Before he published his findings in the early 1970s, it was widely believed that facial expressions and their meanings were specific to each culture [2]. Discrete emotion theories have been criticized on several points like problems in finding correspondences between discrete emotions and brain activity, variability in facial expressions and behaviour and gradations in emotional responses.

Wilhelm Wundt adopted dimensional emotion representation and proposed three principal axes, i.e., pleasure - displeasure, mania - depression, high self-control - "letting-go" [206]. Pleasure - Arousal - Dominance (PAD) model was proposed by Russell and Mehrabian [173]. According to this model, emotional states can be adequately defined by the three aforementioned independent and bipolar dimensions. However, [172] showed that the dimensions that were used for describing affect are interrelated in a systematic fashion. Russell proposed a spatial model (circumplex model) according to which affective concepts fall in a circle that its center is on the intersection of pleasure-displeasure and arousal-sleepiness dimensions. A neurophysiological state characterized along these two dimensions is known as *Core Affect* [171]. Conceptual act model of emotion - a constructive view on experiencing the emotion - was proposed by Lisa Feldman Barrett [16]. It posits that the experience of emotion is analogous to the experience of colour. People experience colours as discrete categories (e.g., blue, red) but the physics of colour is actually continuous, with wavelengths measured in nanometers along a scale from ultraviolet to infrared. A person that experiences a model as "blue" uses her knowledge of colour to give this wavelength a label (in fact, people experience a whole range of wavelengths as "blue"). Similarly with colours, emotions are thought as discrete (e.g., fear, anger, e.t.c.), while core affect is continuous. The conceptual act model suggests that people categorize and label their current feeling of affect, using their knowledge of emotions, just as they experience and label colours. For instance, if someone is experiencing negative affect and sees a snake, she would categorize (and experience) her affective state as "fear," in essence generating an instance of fear. In contrast, a "basic emotions" theorist would say that seeing the snake triggers a dedicated "fear circuit" in the brain.

According to [176] the goal of dimensional affect recognition is to improve the understanding of human affect by modeling affect as a small number of continuously valued continuous time signals (dimensional emotion representation). One of the benefits of dimensional emotion representation is that small differences in affect over time can be encoded. The main dimensions that are used in the dimensional representation of emotion are *valence, arousal (or activation)* and *dominance*. Another dimension - not very

[1] Dissmell is a term coined by Tomkins to differentiate between disgust expressed through the mouth (dislike of taste) and disgust expressed through the nose (dislike of smell).

popular on text - is *surprise*, however affective experiences are best characterized by the dimensions arousal and valence. *Valence* is the subjective feeling of pleasantness or unpleasantness and it ranges from highly positive to highly negative. *Arousal*, that measures the intensity of the emotion, is the subjective state of feeling activated or deactivated and it ranges from calming or soothing to exciting or agitating. *Dominance* represents the controlling and dominant nature of the emotion. For instance while both fear and anger are unpleasant emotions, anger is a dominant emotion, while fear is a submissive emotion. *Surprise* is a brief mental and physiological state, a startle response experienced as the result of an unexpected event and represents the difference between expectations and reality. Surprise can have any valence (positive or negative surprise) and it can occur in varying levels of intensity ranging from very-surprised to little-surprised, however it is not a main dimension for affective analysis of text.

Some common phenomena that appear in text and they are related with surprise, are "Garden path sentences" and "Paraprosdokian". The garden path sentence effect occurs when the sentence has a phrase or word with an ambiguous meaning that the reader interprets in a certain way, and when they read the whole sentence there is a difference in what has been read and what was expected. Garden path sentences illustrate the fact that when humans read, they process language one word at a time. "Garden path" refers to the saying to be led down "the garden path", meaning "to be misled". Garden path sentences are related to Paraprosdokian, according to which the latter part of a sentence or phrase is surprising in a way that causes the reader to reinterpret the first part. It is frequently used for humorous or dramatic effect, sometimes producing an anticlimax. The best paraprosdokians not only change the meaning of an early phrase, they also play on the double meaning of a particular word, creating a form of semantic zeugma or syllepsis [113].

1.3 Thesis Contribution

The work presented in this thesis focuses on the description, development and analysis of affective computational models that estimate emotion on various lexical units. The presented affective models are extended versions of the affective model presented in [107]. The main assumption of this affective model is that "*semantic similarity can be translated into affective similarity*". Specifically, it requires a semantic model that captures how similar each token is to another, and a manually created affective lexicon that is used for seed selection. Usually some hundreds of the affective lexicon entries are used as seeds. The aforementioned model is enhanced in order to achieve more robust performance and it is evaluated on more affective dimensions and languages. Furthermore, we present an innovative compositional model for the affect estimation of word pairs, that is proposed for the first time, since, in literature, compositional models are mainly used in semantic similarity applications. The key characteristic of compositionality is that the (affective) meaning of structures more complex than words is comprised by the interactions of the constituent parts. We also created the first Greek affective lexicon translating a well known English affective lexicon [24] and rating the translated words by native Greek speakers. Affective models that estimate the affective ratings of larger lexical units like sentences, dialogue utterances, news headlines and movie subtitles are also proposed. The system for the affective analysis of tweets was a module of the tweester [142] system that achieved state-of-the-art performance and ranked in the first place of other 19 participants in a worldwide competition.

1.3.1 Challenges

Affective analysis of text is a research area with many challenges. Those challenges make the problem of emotion recognition from text difficult to solve but very interesting. The biggest challenge of affective computational systems is the ambiguity. For example, Fig. 1.3 shows the different emotion that a phrase may convey. It is impossible to detect the different emotions assuming only the lexical information.

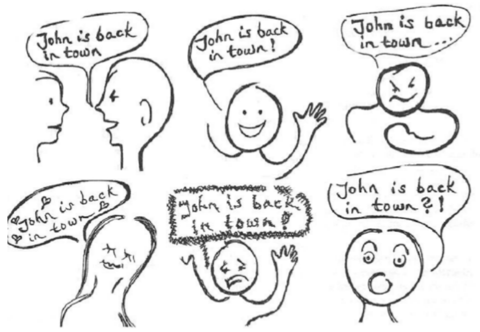


Figure 1.3: The phrase "*John is back in town*" may convey various emotions [9]

In addition, the key characteristics of affective analysis of text are problems related to Natural Language Processing (NLP), to the type of the text or the speaker/writer. The solution of NLP problems like Name Entity Recognition (NRE), Anaphora Resolution and Part of Speech Tagging could help to the accurate estimation of emotion from text. Depending on the speaker/writer, phrases may be sarcastic, ironic or humorous and the detection of such phrases introduces extra challenges that could be faced with affective models that are adapted to the behaviour of each speaker. Another challenge is that the type of the text pieces that are analyzed, are not always formal. For example social media texts like twitter statuses (tweets) are written in informal language and they are usually characterized by abbreviations, lack of capitals, poor spelling, poor punctuation, poor grammar.

Despite the challenges the affective models that are presented in this thesis are competitive and achieve state-of-the-art results for multiple languages and text types (e.g., words, tweets).

1.3.2 Publications

This work resulted in the following publications so far:

- Elisavet Palogiannidi, Elias Iosif, Polychronis Koutsakis and Alexandros Potamianos, "*Valence, Arousal and Dominance Estimation for English, German, Greek, Portuguese and Spanish Lexica using Semantic Models*", in Proceedings of Interspeech, September 2015 [139].

- Elisavet Palogiannidi, Elias Iosif, Polychronis Koutsakis and Alexandros Potamianos "*Affective lexicon creation for the Greek language*", in Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC) 2016 [143].

- Elisavet Palogiannidi, Polychronis Koutsakis and Alexandros Potamianos, "*A semantic-affective compositional approach for the affective labelling of adjective-noun and* *noun-noun pairs*", in Proceedings of WASSA 2016 [140].

- Elisavet Palogiannidi, Athanasia Kolovou, Fenia Christopoulou, Filippos Kokkinos, Elias Iosif, Nikolaos Malandrakis, Harris Papageorgiou , Shrikanth Narayanan and Alexandros Potamianos, "*Tweester: Sentiment analysis in twitter using semantic-affective model adaptation*", in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval) 2016 [142].

- Jose Lopes, Arodami Chorianopoulou, Elisavet Palogiannidi, Helena Moniz, Alberto Abad, Katerina Louka, Elias Iosif and Aleandros Potamianos "*The SpeDial Datasets: Datasets for Spoken Dialogue Systems Analytics*", in Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC) 2016 [99].

- Spiros Georgiladakis, Georgia Athanasopoulou, Raveesh Meena, Jose Lopes, Arodami Chorianopoulou, Elisavet Palogiannidi, Elias Iosif, Gabriel Skantze and Alexandros Potamianos "*Root Cause Analysis of Miscommunication Hotspots in Spoken Dialogue Systems*", in Proceedings of Interspeech 2016 (to appear) [57].

1.3.3 Thesis organization

The thesis is organized as follows: In Chapter 2 we present the related background with numerous applications of affective text analysis. In Chapter 3 we describe the main affective model with its enhancements that is used for the affective ratings of words. The compositional approach for the affect estimation of word pairs is shown in Chapter 4 and the affective models for the estimation of affect in sentence level are analysed in Chapter 5. We close this thesis with the discussion of conclusions and ideas for future research.

Chapter 2

Background

> “*The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions.*”
> Marvin Minsky

The work presented in this thesis deals with the estimation of affective ratings of textual pieces. However, the underlying assumption of the affective computational models we use is that “*semantic similarity can be translated into affective similarity*”. In other words, important requirement in order to understand *how* “something” is said is to understand first *what* “something” means. Thus, semantic models, i.e., models that define how similar is each word to another, are an integral part of our affective computational models. Such models are broadly known with the term Distributional Semantic Models (DSMs) and they are described in the first part of this chapter. The second part of this chapter deals with some of the related works and applications that have been reported in literature. Works of interest cover the automatic estimation of word, phrase and sentence level affective ratings. We also give a brief idea about other than emotion affective dimensions, like personality and humour or irony. We continue reporting a set of interesting affective text approaches and we close this chapter by listing the features, used in the computational models for affective analysis of text.

2.1 Distributional Semantic Models

In the early 50s Harris set the question “*Does language have a distributional structure?*” [65]. He argues that each language can be structured in respect to various independent features and he discusses how a language can be described in a distributional structure, i.e., in terms of the occurrence of parts relative to other parts. The distribution of each language element can be understood as the sum of its environments, i.e., an array of its co-occurrents (the other elements that co-occur with that element in an utterance). The distinction between distributional structure and meaning is not clear, since the latter is a general characteristic of human activity and not a language property (in contrast to the language, a person's store of meanings changes through the years). However, it is a distributional fact that some elements are similar to others. Thus, the distribution of the language elements (co-occurrence arrays) can be used in order to compute the similarity between them. The aforementioned description can be enclosed to the phrase "*similarity of context implies similarity of meaning*" [65].

A plethora of works is based on Harri's statement and implements DSMs in order to extract semantic information of language. [198] surveys the use of Vector Space Models (VSMs) for semantic processing of text. The most attractive property of VSMs is that they extract knowledge directly from a given corpus and they perform well for the computation of similarity of meaning between words, phrases and documents. VSMs that form word-context matrices are most appropriate for computing similarity between words. [46] discovered that word similarity can be measured comparing row vectors in term-document matrices. This approach evaluated by many works [88, 101, 161, 177] concluding that performance of computing semantic similarity between words increases as the context window shrinks, i.e., the immediate context of a word is very important for determining the meaning of a word. [74] shows how semantic similarity between words can be computed through web harvested data without using human annotated knowledge. They compare context-based with co-occurrence-based metric and they show that the former significantly outperforms the latter in terms of correlation with human judgements. The majority of the semantic similarity metrics like context- and co-occurrence- based use hand-crafted language resources [79, 90, 92, 153]. According to the last years trend instead of the traditional count-based distributional models, neural-network inspired representation models [67, 114, 118, 210] are developed and evaluated on word semantic similarity tasks.

2.1.1 Compositional Distributional Semantic Models

Single-word VSMs perform very well on learning lexical information, but they can't be applied on larger lexical structures like phrases and sentences. Compositional Distributional Semantic Models (CDSMs) focus on larger than words lexical structures and they aim to capture *compositionality*, i.e., the meaning of an expression is determined by the meaning of the constituent words when they are combined with each other according to specific rules. Many works aim to capture compositionality using VSMs [208, 209]. Semantic compositionality allows the construction of complex meanings from simpler elements based on the principle that the meaning of a whole is a function of the meaning of the parts [148]. The key characteristic of compositionality is that the meanings of the constituent parts are combined into a single token [77, 117]. Compositional approaches in vector-based semantics can be modelled by applying a function f that acts on two constituents a, b in order to produce the compositional meaning p. Functions that were investigated by [77, 117] are addition and multiplication. The additive compositional model takes the sum of the two vectors weighted with the appropriate weight matrices A and B respectively ($p = Aa + Bb$) and the multiplicative model is the projection of the ab tensor product using a weight tensor C ($p = Cab$). These composition forms can be also simplified using the additive model with scalars instead of matrices. Similarly the multiplicative approach can be reduced to component-wise multiplication. [15] is based on the compositional model proposed in [117] and focuses on adjectives and nouns. [184] proposes a recursive neural network that learns semantic representations at phrase and sentence level. A semantic compositional approach that employs an activation model is proposed in [58].

DSMs and CDSMs provide us the semantic knowledge about lexical units that is incorporated in the affective computational models. The assumption that "*semantic similarity can be translated into affective similarity*", indicates that it is possible to find a mapping from semantic knowledge to affective knowledge.

2.1.2 From semantic to affective space

[159] argues that the cognitive processes should be incorporated in the computational models in order to predict successfully human behaviour. Conceptual semantic spaces and representation learning can be assumed as cognitive processes of human behaviour. We have already discussed in brief how the semantic space can be modelled. However, the core work of this thesis deals with the affective space, and the next step is to discuss how affect can be modelled in order to proceed to a higher level representation. Semantic-Affective models (SAM) [107, 197] aim to map from semantic to affective space based on the assumption that "*semantic similarity implies affective similarity.* The main concept of these models is depicted in Fig. 2.1

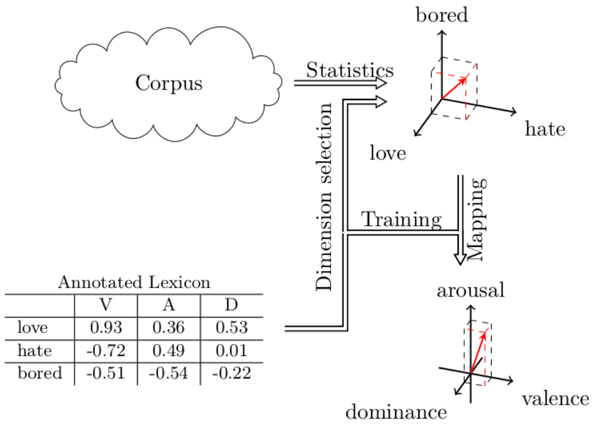


Figure 2.1: Description of how semantic-affective mapping is achieved [105]

As depicted in Fig. 2.1, SAM requires a semantic model that it is built based on the distribution of a corpus, an annotated affective lexicon for selecting the appropriate dimension of seeds and for learning the mapping (weights) from semantic to affective space. In this thesis we follow this cognitive approach of affective analysis of text and we use and enhance the SAM proposed in [107]. According to this model, a semantic space is built following a distributional approach and an external resource with human knowledge (affective lexicon) is used for learning the mapping and finally estimate affective ratings. The model is described in details in Chapter 3.

2.2 Affective Text Applications

The motivation for focussing our efforts on the affective analysis of text is much more than high. Except from the challenges of building models that mimic human behavior in order to understand and feel like humans, this research area is directly connected with a plethora of daily life applications.

During the last years communication through text has rapidly grown. People exchange text messages between each other (SMSs, emails, chat) and express their opinions through micro-blogging (e.g., twitter) or on product review platforms. Communication through texting usually consists of the exchange of short messages and due to its impersonal nature it is preferred, especially in the case of opinion expression. Moreover, text messages is, usually, the only way of communication in social networks, e.g., status updates, comments and chat. Since popularity of social networks has noticeable increased the aforementioned text messages lead to the creation of a huge data pool. Countless text messages are exchanged in daily basis, making their (affective) analysis a hot topic. Some affective analysis tasks are *sentiment analysis* [129,145,203], *aspect-based sentiment analysis* [149,155,156,191], *opinion mining* [29,145], *lexicon expansion* [39,53,80,82,107], *subjectivity identification* [97,144], *sentiment analysis of figurative language* [59,165,166], *summarization* [70,144].

2.2.1 Multilingual affective applications

An open research question is whether affective analysis of text can be handled equally well across different languages. Affective analysis for multiple languages was investigated in [14] for subjectivity detection and in [87] for polarity (valence) prediction. Most computational systems for affective text analysis require affective lexica, i.e., language dependent emotionally annotated resources. Affective Norms for English Language (ANEW) [24] contains 1034 words and each word is assigned with one score per affective dimension (valence, arousal, dominance). ANEW has been translated in other languages, i.e., Spanish [163], European Portuguese [183], Italian [122] and lexica words were annotated by native speakers. Additionally to ANEW, affective lexica with less or more entries can be found for English [187] and other languages such as German [83] and Dutch [123]. Bilingual approaches have been reported as well, e.g., [51] created affective norms from English and Finish nouns. [60] provided an affective lexicon of French attributes and investigated the age influence on the ratings collection process. The words of ANEW were also annotated with respect to discrete (categorical) emotions by [186].

A multilingual approach using machine translation is shown in [13]. [47] translates any other than English language into English and then leverages lexical resources for sentiment analysis available in English. A mutlilingual sentiment analysis method of English, Dutch and French web documents is presented in [20]. [40] estimates the sentiment of English and non-English tweets using emotion tokens like emoticons, repeated punctuation and repeated characters. A probabilistic generative model that aims to capture properties of one language using data from other languages (English, German, Chinese) is described in [22].

In Chapter 3 we show how we created the first Greek affective lexicon, translating the ANEW [143] and how our semantic-affective model can be applied in different languages achieving consistently high performance [139]. We also report the performance of a cross-language model that uses the affective ratings of a language's words for estimating the affective ratings of another language's words.

2.2.2 Sentiment Analysis of Social Media

People usually use social media in order to declare their opinion or how they feel about something, i.e., they express their sentiments. Sentiments are caused by objects and experiences however they may be weak and imperceptible to the conscious mind [43]. Sentiment analysis involves the automatic detection of sentiments embodied in text. Subjectivity and polarity detection are popular sentiment analysis tasks and the proposed approaches are described in [211]. They usually employ lexica and predefined rules to estimate sentiment of small lexical units (words or phrases) that are composed to estimate sentiment of larger lexical units, or they use machine learning techniques to find patterns from the data. Social theories in microblogging are mentioned in [71] and they could be helpful for sentiment analysis. According to "Sentiment Consistency" [3] two messages of the same user are more likely to be consistent than those of two randomly selected messages, and "Emotional Contagion" [68] suggests that two messages of friends are more likely to be similar than those of two randomly selected messages.

The rise of social media such as blogs, social networks reviews and forum discussions, has fuelled interest in sentiment analysis. The public opinions that are shared in social media can be used for decision making and they are very important for businesses, since through opinions they get feedback for their products. Businesses look to automate the process of understanding the conversations and recently the field of sentiment analysis has grown too. A challenge of sentiment analysis in social media is that the expressed opinions are spread over a variety of topics. Due to the explosive growth of social media many sentiment analysis models are applied on social media data like twitter [127, 169, 170], facebook [5, 178], facebook with application to e-learning [134], prooduct/movie review sites [145,191], news and blocks [61] and forums [91,191].

In Chapter 5 we show the model we developed [1] for sentiment analysis in twitter, that ranked first in a wordlide competition [142].

2.2.3 Emotion detection on spoken dialogues

Another interesting application of affective text models is the emotion detection of spoken dialogue transcripts. Such applications can be incorporated in various spoken dialogue systems that humans interact with machines. Emotion recognition from French agent-client dialogues is addressed in [49] using linguistic information. [96] utilizes speech and lexical features in order to predict students' emotions in computer-human spoken tutoring dialogues. A similar application is presented in [54] but for human-human tutoring dialogues. [95] shows how the context of dialogues can be used to improve emotion detection. They use four features to define the lexical context, i.e., the transcriptions of the two previous dialogue turns and the Livenstein distance between the current and the previous two utterances. An application appropriate for call centers is presented in [125]. The issue of detecting emotions on call centers was also addressed in [201] on medical emergency calls in French, incorporating lexical cues, i.e., words and syntax.

In chapter 5 we present a model that detects anger in Greek calls employing affective features derived by text as well as its fusion with the corresponding speech system [99].

2.2.4 Sentiment analysis of News and Movie Subtitles

In the applications described in the previous subsections we have to detect emotion that comes of humans and mirrors the way the feel about an event or the way they react on event. However, non user derived text messages that aim to provoke emotions can also be found (on the web). Some representative examples are news headlines, movie subtitles and stories. [188] describes an affective text task the goal of which is the emotional labelling of news headlines (from New York Times, CNN, BBC News and Google News search engine)

[1] Athanasia Kolovou, Fenia Christopoulou and Filippos Kokkinos are also contributors to this model.

on continuous and discrete scale. The same headlines were used by [189] for the detection of the six basic emotions (anger, disgust, fear, joy, sadness, surprise), by [85] for emotion detection using and SVM approach and by many others, e.g., [33, 34, 78]. Headlines from Times of India were used for disgust, fear, happiness and sadness detection [78].

[147] employs Wordnet [116] to proceed to a scheme for automatic emotional annotations of movie subtitles. In [81] a method for detecting the emotional scenes of movies based on the semantics conveyed by subtitles is described. A challenge that aims to detect emotion in real time conditions is described in [50] and employs subtitle analysis.

In Chapter 5 we show how SAM can be employed for the affective labelling of news headlines and movie subtitles.

2.2.5 Multimodal affective applications

Affective analysis of text can contribute to the development of multimodal affective systems. Affective systems usually require external resources in order to acquire knowledge for each modality. Such resources are available for audio [25], visual [89] and text [24] modalities and they are annotated on the continuous affective dimensions valence, arousal and dominance. Combining systems that analyse affect in different modalities, we usually end up with a more accurate sentiment prediction, as well as we exploit all the available information of online resources such as videos that are used for the expression of sentiments and opinions. A multimodal affective system is proposed in [86] for the task of movie summarization. The work of [124] addresses the issue of multimodal sentiment analysis, experimenting with web videos that come from YouTube. A similar multimodal approach for the sentiment analysis of spanish YouTube videos was proposed by [168]. Many other multimodal works have been proposed for sentiment analysis [152, 157, 158, 204].

In chapter 5 we present a bimodal (text and speech) system for anger detection in Greek human-system calls [99].

2.2.6 Sentiment analysis of political relevant texts

Another interesting application of sentiment analysis concerns on political data. Political data introduce the challenge of the topics variety, and as [195] mentioned, sentiment analysis tasks are harder when the abstraction level of the topics increases. A summary of political texts analysis as well as their challenges on sentiment analysis tasks can be found in [126]. Most times, political relevant text messages are posted on social networks like twitter. A system that analysis tweets with respect to sentiment in real time for the U.S. presidential election is described in [202]. The prediction of the 2009 German elections through twitter sentiment analysis was the goal of [193] and a modeling of the Indial elections is presented in [181]. Political sentiment analysis using tweets is also the main task of [6, 52] and an approach where the Malaisian government social media are analysed with respect to sentiment is shown in [66].

2.3 Other interaction dimensions

Other than emotion interaction dimensions for affective analysis of text include the detection of humour, irony, sarcasm and personality. These dimensions are an integral part of the affective text analysis and even if we don't address any of these dimensions in this word, we present their main idea for completeness reasons.

2.3.1 Personality

[31] summarizes that personality is an affect processing system that describes persistent human behavioural responses to broad classes of environmental stimuli, characterizing a unique individual. It is involved in communication processes and connected to how people interact one another. Personality is also something that changes over time and adapts to the environment. People may also pretend to have different personality traits. The general position of psychologists about these problems is that individuals have some rather fixed core personality traits and other more variable peripheral traits [31]. Computational models for personality detection follow the big-five representation, according to which there are five core personality traits. Each of the five personality traits represents a range between two extremes, i.e., i) extraversion includes characteristics such as excitability, sociability, talkativeness, assertiveness and high amounts of emotional expressiveness, ii) agreeableness includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors, iii) conscientiousness dimension includes high levels of thoughtfulness, with good impulse control and goal-directed behaviours, iv) neuroticism captures the tend to experience emotional instability, anxiety, moodiness, irritability, and sadness, v) openness includes imagination and insight, and those high in this trait also tend to have a broad range of interests. One of the most popular approaches for personality detection through text is the analysis of words [151]. Personality Recognition from Text is very useful in Social Network Analysis and Opinion Mining. Social network data is 1) often not publicly available, 2) unlabeled, 3) very difficult to annotate with personality judgments and 4) in a lot of different languages [31] and this task requires skills from disciplines such as Linguistics, Psychology, Data Mining and Communication Sciences. In [32] some non-trivial problems during Personality recognition from text are summarized, e.g., personality is a fuzzy and subjective notion, annotation of personality in data from social networks, requires judgements of the authors or by native speakers. An approach for adaptive personality recognition on social networks proposed by [31]. [32] presents a system that builds on the fly one personality model for each user in a corpus in an unsupervised way and performs automatic evaluation of the models comparing all of user's posts. They adopted the "Big Five" dimensions and using the linguistic features associated with those dimensions, the system generates one personality model for each user.

2.3.2 Humour, Irony, Sarcasm

Terms such as *humor, irony, sarcasm, funny, laughable, ridiculous* e.t.c., are folk-concepts with fuzzy boundaries if any. Lexicographic studies have shown that the semantic field of what has been broadly defined as "humour" is very rich in closely related, barely distinguishable terms. Irony is seen as distinct from humour, but the same definitional problems exist. Irony however fall under the technical sense of humour. [12]. While there exists humour that is not ironical and ironies that are not perceived as funny, the issue is not as simple as the intersection of two distinct set of facts.

Humour is the tendency of particular cognitive experiences to provoke laughter and provide amusement. The term derives from the humoural medicine of the ancient Greeks, which taught that the balance of fluids in the human body, known as humours, control human health and emotion. It consists of a semantic and a pragmatic facet and it can be used for bonding, releasing tension, attracting a mate, putting a rival in his place, or entertaining a child [72]. Humorous texts maximize the degree of perplexity by profiling a structural ambiguity [166]. Irony can naturally occur in both language and circumstance; one experiences irony when the opposite of an expected situation or idea occurs. In essence, an individual does not need to go out of their way to experience an ironical situation or idea, they can occur naturally. Irony is found in the contrast between expected, or ordinary, outcomes and what actually happens. The greater the distance, the greater the irony. Sarcasm is the use of irony to mock or convey contempt. Because of this, sarcasm tends to find more broad usage than irony. A figure of speech in which the intended meaning is the opposite of that expressed by the words used; usually taking the form of sarcasm or ridicule in which laudatory expressions are used to imply condemnation or contempt. Irony is where "the literal meaning is opposite to the intended"; and sarcasm is "aggressive humour that pokes fun". A list with examples is shown in Fig. 2.2.

Type	**Examples**
Humor	"*A balanced diet is a cookie in each hand.*"
Sarcasm	"*But I shall be sober in the morning, and you will still be ugly*".
Verbal Irony	"*What a nice day*" when it is raining
Situational Irony	A man who is a traffic cop gets his license suspended for unpaid parking tickets
Cosmic Irony	The Titanic was promoted as being 100% unsinkable; but, it sank on its maiden voyage.
Dramatic Irony	In Romeo and Juliet by William Shakespeare Romeo finds Juliet in a drugged stateand he thinks she is dead. He kills himself. When Juliet wakes up she finds Romeo dead and kills herself.
Socratic Irony	A professor never answers questions and does not explain key concepts of the course; however he expects students to come to class after having read their assignment, ready to answer the professor's questions.
Verbal Irony	The day was as normal as a group of seals with wings riding around on unicycles, assuming that you lived someplace where that was very normal.

Figure 2.2: Examples that make the difference between the terms irony, sarcasm and humour more clear.

2.4 Affective text analysis

In the previous section we described some popular affective text applications and dimensions like personality, humour and irony. In this section we discuss about the affective analysis of text in the scope of sentiment analysis and emotion detection. In Fig. 2.3 we show the main categories of the affective computational models. The computational models for affective text analysis are distinguished into three main categories, namely *knowledge based*, *statistical* and *hybrid* techniques. The models of each category are further distinguished based on their granularity level, i.e., affective analysis of *documents*, *sentences*, *phrases*, *words*. Usual documents are product reviews, while sentences can be news headlines, social network statuses (tweets), dialogue utterances or subtitles. Each

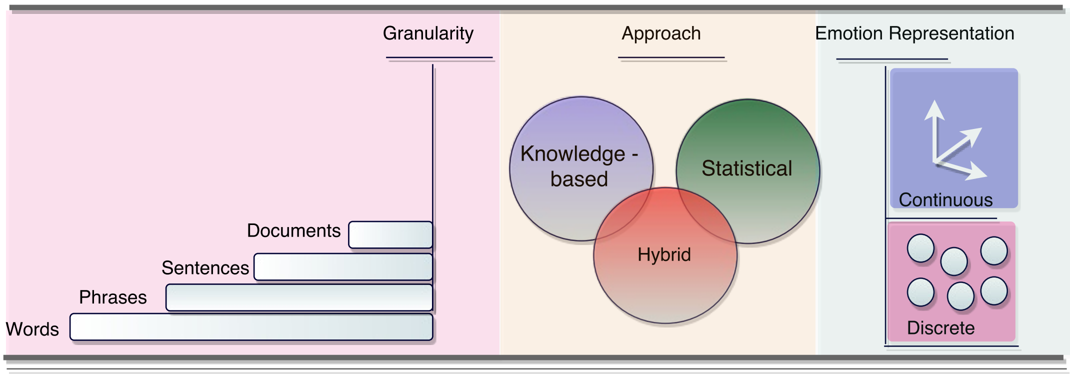


Figure 2.3: Main categories of affective computational models

computational model follows either categorical or continuous emotion representations. Models that follow the former representation focus on extracting a set of emotional labels like *happiness*, *sadness*, *anger* e.t.c., while models that follow the latter representation predict values over emotional continuous dimensions, e,g., *valence*, *arousal*, *dominance*. Sentiment analysis literature summary can be found in [110, 145, 162].

The affective models of higher granularity, usually follow combination techniques in order to estimate the affective rating of a multiword term from the constituent words. It is common to combine only affective bearing words, that are usually extracted from affective lexical resources. Fusion approaches for the extraction of sentence level affective ratings may contain arithmetic methods like *mean*, or syntactic rules, while not many works have address the issue of modeling directly the affective content of n-grams. A sentiment composition approach, i.e., the process that the parts of the multi-word term interact with each other to give an overall sentiment is described in [37]. Sentiment composition may consist of heuristic rules that incorporate negators, or voting based heuristics or heuristics with compositional semantics [120].

2.4.1 Approach Categorization

Each computational model for affective analysis of text belongs to one of the three main categories that were mentioned previously. **Knowledge-based** techniques are very popular and they leverage affective word and multiword resources that have been created either manually or automatically. Such approaches are usually based on the presence of unambiguous affect words (e.g., happy, sad, afraid, bored) and they can be extracted by affective lexica. The major drawback of knowledge based approaches is that they don't take into consideration the structure of large expressions. Moreover, affective lexica are of limited size and as a result many of the mutliword expression words are not contained in them. One more weakness of knowledge based approaches is that they are not language-agnostic, since different knowledge is required for each language. Valid knowledge-based systems provide a comprehensive knowledge base that encompasses as human knowledge as possible. Without reliable resources it is difficult to grasp the semantics related with human behavior and expressed through text. Some keyword and rule based approaches can be found in [7, 34, 98, 111, 130, 132].

Another approach of computational models for affective text analysis is that of **sta-**

tistical modelling. Such approaches employ machine learning algorithms that are fed with training affectively annotated data and classify words of larger lexical expressions to the affective classes. The weakness of statistical methods is that they are semantically weak, i.e., the predictive value of individual lexical or co-occurrence elements is low [28]. Statistical approaches include the building of classifiers [70,136,195,196] or neural network approaches [67, 158, 180, 210]

Hybrid models for affective text analysis combine knowledge-based and statistical techniques. In the following table we categorize literature works based on the approach they follow, the granularity and the emotion representation. Many hybrid approaches for sentence level affective analysis can also be found in the participating teams of the sentiment analysis tasks [128, 155, 156, 169, 170].

Approaches that focus to word granularity can be found in [19, 53, 107, 133, 190, 196]. The issue of emotion detection in phrase level has been tackled in [107,203,207]. Majority of the literature focuses on sentence level emotion prediction [4, 45, 52, 59, 107, 111, 166, 167, 169, 170, 202], while some works scope to all granularity levels [103, 107]. Sentiment analysis of documents was addressed by [21,35,145,205]. Most works aim to the prediction of emotion in continuous represntation, however works that score to discrete emotion prediction can be found as well [9, 186]

[4] describes an unsupervised context-based approach for predicting emotion at sentence level, without the use of any affective lexica or annotated data. [196] introduces an unsupervised algorithm, based on mutual information, for semantic orientation. They define the semantic orientation of a word as the strength of its association with the positive words minus the strength of its association with the negative words. In [121] a quashi-compositional sentiment learning and parsing framework that is well-suited for exhaustive, uniform, and principled sentiment classification across words, phrases and sentences is presented. The proposed method is a hybrid sentiment learning and parsing framework. In [211] sentiment divergence metrics are proposed. First the sentiments associated with individual words are quantified (a word can be either subjective or objective). They use lexicon-based approach and SentiWordNet [53]. The works presented in [127,128,169,170,188] indicate the reaction of systems in more realistic conditions, when there is not fully control of the data, and more specifically they describe tasks for Sentiment analysis on Social Networks and Affect text. [136] experimented on affect sensing using a general purpose dictionary and an affect dictionary and the results showed that the results in each case are similar, for similar number of features. In [135] a comparison of two emotional textual corpora and the features used is presented. The main question in this research is whether the two corpora can be analyzed using the same features or whether the feature set should vary according to some specific properties of a particular corpus.

2.4.2 Annotation and Resources

One bottleneck in the computational systems is the existence of labelled data. It is very important to have resources that can be incorporated in the machine learning algorithms as well as training data. Computational methods are necessary to create or expand an already existing lexicon, creating larger resources such as [53, 190]. However, there are still limitation, e.g. WordNet based efforts can't produce ratings for words not included in WordNet. The disadvantage of manually created lexicons is that they provide a low coverage, since they contain only a few thousand words. Thus computational methods are used to create or expand an already existed lexicon. The assumption for such methods is that semantic similarity can be translated to affective similarity.

Affective datasets have been created in [9, 19, 24, 69, 122, 131, 163, 183, 187, 203] [107] expands such affective lexica and covers a significant fraction of the vocabulary of a language. SentiWordNet and WordNetAffect are examples of large affective resources created through computational models. In the first, [53] annotated automatically all WordNet [116] synsets (sets of synonyms used to represent word senses). The second was developed by [190] who represented the affective meanings by selecting and labeling a subset of WordNet synsets. Computational models for affective text usually incorporate small manually created resources [106], or larger automatically created resources [34]. The availability of multilingual affective resources allows the investigation of the universality of text-based affective models for different languages, enabling the development of cross-language tools.

2.4.3 Features

The big picture of the affective features used for the affective analysis of text is depicted in Fig. 2.4. Specifically, the main categories of affective text features are numerical values, the lexical information itself or non lexical information like punctuation and they can derive from text, can be spoken dialogue specific or incorporate social networks information or they can be extracted from affective resources. Emotion, irony and personality detection cover some very popular affective text tasks. Personality is an affect processing system that describes persistent human behavioural responses to broad classes of environmental stimuli, characterizing a unique individual. It is involved in communication processes and connected to how people interact one another. Personality is also something that changes over time and adapts to the environment. People may also pretend to have different personality traits. The general position of psychologists about these problems is that individuals have some rather fixed core personality traits and other more variable peripheral traits [31]. Humour and irony belong to figurative language and each device exploits different linguistic strategies to be able to produce an effect (e.g., ambiguity and alliteration regarding humour; similes regarding irony). In irony a polarity negation phenomenon occurs [166]. Since irony cuts through different aspects of language (from pronunciation to lexical choice, syntactic structure, semantics and conceptualization), it is unrealistic to seek a general solution just in one single technique or algorithm. Irony courts ambiguity and often exhibits great subtlety and evokes certain types of emotions.

Features that lie in the category of lexical information can be used in any corpus type and for all the affective text tasks. They are more representative since affect in mainly expressed through words (in written language). Such features can be the words of an existing vocabulary (bag of words), the most frequent words, hapax legomena, specific words like singular pronouns or prepositions, negative particles or words that convey positive or negative emotions, long words, i.e., words that contain too many characters or swears [32, 135, 136]. Presence of linguistic indicators such as hedges, modals, certain affixes lie to this category as well [9]. A specific case of lexical information features is the LIWC [151], a tool that proposed for personality detection. This tool is used for the categorization of words. Specifically, the words are classified into categories independently of their emotional content. The underlying idea is that the words people use reflect their feelings. The simplest way to extract these features from a dataset, is to count the words of each category Non lexical information can result to features like the part-of-speech tags,

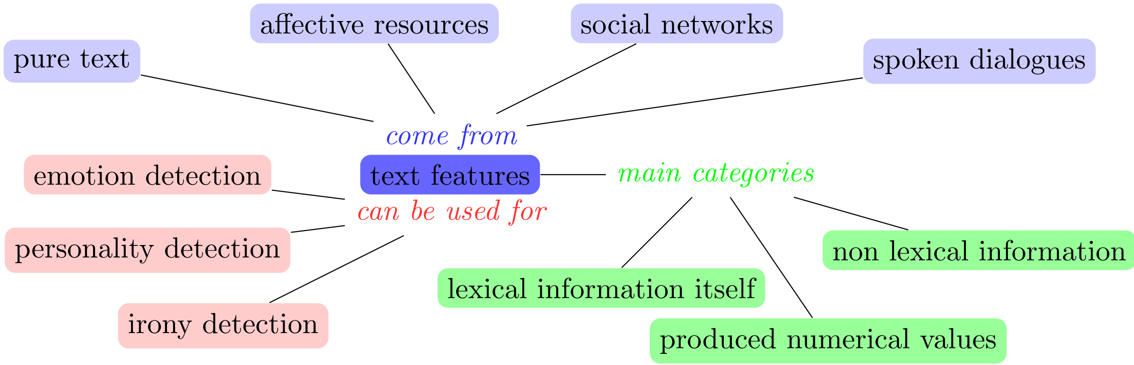


Figure 2.4: The big picture of the features used in affective text computational models.

punctuation, emoticons, urls,reference to other users. Such features are very popular for the task of sentiment analysis of social media and many links to works that use such features can be found in [32, 128, 169, 170].

Affective features can also be numerical values produced by numerous textual characteristics. Numerical features usually include word frequency, word occurrence (based on a dictionary), the number of a word's possible senses [107], affective ratings discrete or continuous, word counts, statistical values like mean, max e.t.c. [107, 135, 136] of affective ratings [103], word counting [102, 151], number of concreteness, familiarity or imaginary words [102]. Semantic features, have also been employed in affective models [107], based on the assumption that "*semantic similarity can be translated into affective similarity*".

Spoken dialogue features can include any of the aforementioned features in each dialogue turn (utterance) as well as dialogue specific features. Dialogue based features include statistics of the previous turn, false starts (e.g., I do-don't), a binary feature whether the previous turn is question or not, the number of barge-in, the turn duration (computed by start and end time if logs are available) [135].

No single feature is particularly ironic, but combinations can be used for irony detection. Contrast features [103] have a high contribution to the detection of ironic textual pieces.

2.5 Summary

In this chapter we introduced to the reader the main research areas and components that are related to this work. We first describe and present literature in the semantic space area and then we continue describing numerous affective text applications. Even if this work focuses only in emotions, we present a brief summary of more interaction dimensions like personality, irony and humour. We close this chapter describing numerous computational modules of affective text analysis systems, like the approach followed, the need of annotated resources and the features.

Chapter 3

Affective labelling of words

"I know nothing in the world that has as much power as a word.
Sometimes I write one, and I look at it, until it begins to shine."
Emily Dickinson

3.1 Introduction

Emotion is mainly conveyed by speech, but it can still be perceived in text and it can be elicited by its content and form [154]. Readers experience emotion by placing themselves in the position of characters of a narrative and imagining their own emotional reaction [84]. Computational tools of affective text analysis can be applied on various applications depending on tasks that they are based on. Affective text tasks are, mainly, categorized based on the emotional representation i.e., discrete vs. continuous emotional representation and the scope of analysis, i.e., predicting emotion on words, phrases, sentences, paragraphs or documents. The approaches that are followed in this thesis mainly concern affective text computational systems that rely on the exploitation of affective lexica. Affective lexica consist of word entries (usually about 1K) of a target language, that are annotated with respect to emotional dimensions, usually valence, arousal and dominance. The structure of the affective lexica is shown in the following figure.

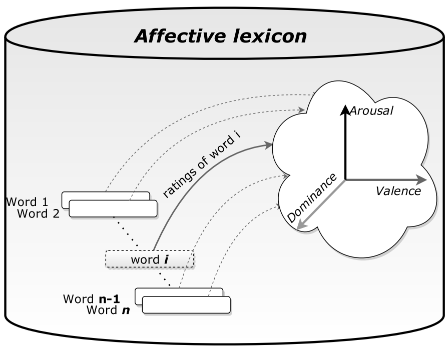


Figure 3.1: Affective lexicon structure.

Since the size of such manually created resources is limited, the development of models that automatically expand them is very important. Such models are used in order to label the affective content of words. Words are the smallest lexical units that may convey emotions and the building blocks for larger lexical units like phrases or sentences, so the estimation of their affective content is a crucial task. In this chapter we analyse a semantic-affective model that automatically expands small affective lexica, in order to achieve good vocabulary coverage.

As denoted by the name, the semantic-affective model leverages both semantic and affective characteristics of words in order to label words with respect to an affective dimension. The semantic-affective model relies on the assumption that "*semantic similarity implies affective similarity*". First, a semantic model is built and then affective ratings are estimated for unknown tokens exploiting the affective ratings of semantically similar words. In this chapter, we are going to explain how the semantic model is built and how the semantic features are incorporated into the semantic-affective model, and how an affective lexicon can be created for a new language. The semantic-affective model is evaluated for multiple languages and affective dimensions. Moreover many semantic-affective model variations are being evaluated and the derived affective ratings are compared and discussed.

3.1.1 Task description

The task we focus on is the affective labelling of words, i.e., to characterize the affective meaning that a word conveys, with respect to the three main affective dimensions valence (V), arousal (A) and dominance (D). Thus, for each word w three continues affective ratings are being estimated, i.e., $V(w)$, $A(w)$, $D(w)$. The approach we follow assumes that there are two spaces, the semantic space, in which all the words live in and they are connected with each other based on their semantic meanings and the affective space, in which words are placed in the space based on their affective meaning. An appropriate mapping is required in order to go from the semantic to the affective space, i.e., to use semantic models in order to estimate the affective ratings of words.

The semantic-affective model finds the appropriate mapping, based on a set of parameters that are going to be explained in the following section. The philosophy of the semantic-affective model that has been created for the described task is depicted through a toy example in the following figure.

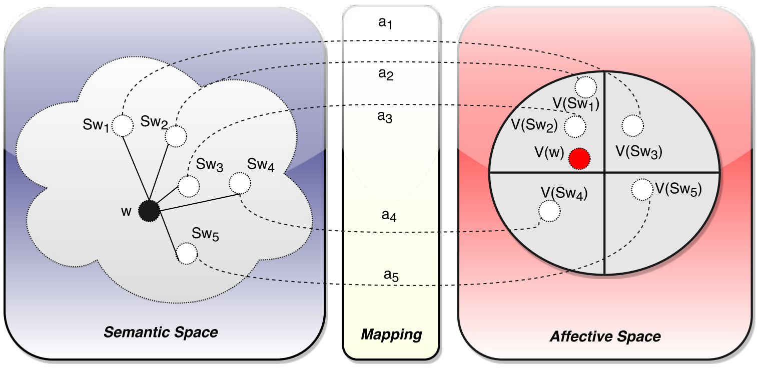


Figure 3.2: Toy example of semantic-affective mapping.

In the toy example shown in the Figure 3.2, we assume a semantic space that contains the word w (black ball) and its semantic distances from five seed words Sw_i and a 1D affective space (Valence space) that contains the valence ratings of the seed words. The valence of the word w (red ball) is estimated combining the affective ratings of the seeds with the semantic distances from w (semantic similarities) and the appropriate mapping (a_i). The affective ratings of the seeds have been extracted from the manually annotated affective lexica. In the next section we describe the manual creation process of an affective lexicon.

3.2 Greek affective lexicon creation

As we have already mention, the affective lexica are necessary in order to bootstrap the process of the affective lexicon expansion (affective labelling of words). Thus, in order to label the words of a given language an affective lexicon in that language is required. One of the most popular affective lexica is the English ANEW [24], that contains 1034 words and their valence, arousal and dominance ratings. Affective lexica were also provided for other languages like Spanish [163], European Portuguese [183], Italian [122], German [83] and Dutch [123]. No affective lexicon was provided for the Greek language, so we created the first Greek affective lexicon based on the English ANEW.

We suggest that the words of an already existing affective lexicon can be transferred to the target language and we translated the words of the English ANEW into Greek. Then the words were shared to the target language's native speakers in order to collect the affective ratings. Each word was manually annotated by 20 participants on average with respect to valence, arousal and dominance (V,A,D). The affective lexicon creation process is summarized in Figure 3.3.

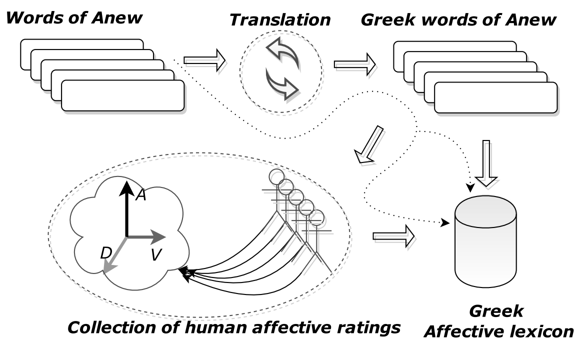
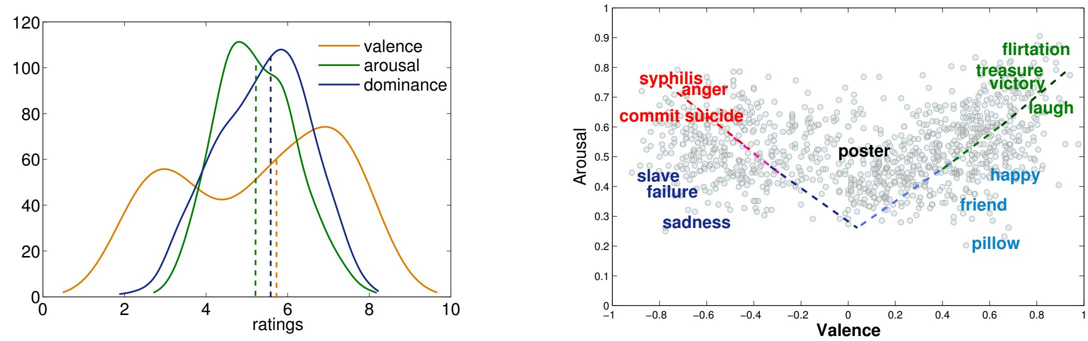


Figure 3.3: Affective lexicon creation process.

At the end of the creation process we have the Greek affective lexicon that consists of 1034 words annotated on V-A-D. The main creation processes are: (1) translation of ANEW words from English to Greek and (2) the collection of human affective ratings. Translation of the ANEW words to Greek was split into two subprocesses: translation proposals, in which two translators were translating the English words into Greek words and translation judgements, in which a third translator selected one of the translation proposals in the case of disagreement. In this way we managed to address the word sense ambiguity issues between the two languages. Collection of human affective ratings took place in a classroom of 105 engineering students (87 males, 18 females), native Greek



(a) Distribution of the affective ratings collected.

(b) Valence-arousal distribution.

Figure 3.4: Distribution of the affective ratings of the Greek ANEW.

speakers, aged from 19 to 30. Each word was rated with respect to (V,A,D) using the 9-point SAM scale ([23]). Each participant was given a sheet with approximately 200 words, a sheet with instructions and the SAM pictures for providing their ratings by circling the corresponding image. In order to avoid context influence the word order was randomized and the participants had one hour in their disposal to provide the affective ratings.

Affective annotation is a very subjective task and two (or more) people may provide very different ratings for the same word. We assume that if many people repeat the annotation for a specific word, the expected value of the ratings will approximate the true affective score of this word. For this reason each word was rated by 20 participants. Regarding the ratings that were collected, no extreme annotator biases were noticed and hence, none of the annotators were excluded from the process. Fleiss' kappa was applied in order to measure the annotators' agreement for each word. The average Fleiss' kappa over all the words indicates fair agreement for valence ($\kappa = 0.29$) and and less so for dominance ($\kappa = 0.23$) and arousal ($\kappa = 0.20$). The aforementioned κ values were estimated on 9-point scale. For 3-point scale the κ values are higher indicating good agreement for valence (κ $= 0.68$) and moderate agreement for arousal ($\kappa = 0.46$) and dominance ($\kappa = 0.49$).

The distributions of the ratings collected for each affective dimension are shown in Figure 3.4a, where dashed lines denote the median values. Comparing the ratings of the three dimensions, we observe the valence distribution is clearly bimodal, which is not the case for the other two distributions. The concentration is higher on words with high valence. This result was also observed by [122]. In Figure 3.4b, the valence-arousal distribution for the words of the Greek affective lexicon is depicted. Each point corresponds to a word, and some words (translated into English) are shown. The distribution appears to follow the well-known from literature V-shape. The interpretation of V-shape is that negative and positive words generally tend to have high arousal, while words with neutral valence have neutral arousal as well.

In Figure 3.5, we show the distributions of the mean values of valence and arousal for the four languages in which ANEW is available. Greek, English [24], Spanish [163] and European Portuguese [183] contain the same information, i.e., the 1034 English words have been translated to the respective languages.

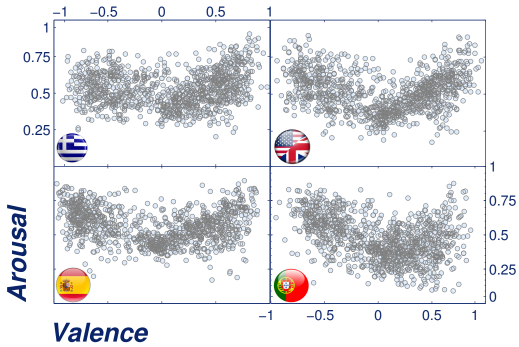


Figure 3.5: Valence-arousal distributions across languages.

As we observe, the distributions are consistent for the four languages, with the differentiation that the V-shape in Greek is not as clear as in the other languages for words with negative valence.

3.3 Semantic Affective model

Now, that the philosophy of the semantic-affective model and the affective lexicon have been explained, we are going to describe in more details how the semantic-affective model works. As shown in Figure 3.2 in order to estimate the affective meaning of a word we should first built a semantic model and then estimate the appropriate mapping that is required for the transformation from the semantic to affective space. A small manually annotated affective lexicon is required for bootstrapping the process, and specifically for selecting the seed words. The selected seeds with their affective ratings are used in order to train the semantic-affective mapping. The affective rating of unknown words is then estimated based on the assumption that words that are semantically related are also affectively related, leading to an expanded affective lexicon.

According to [107], the semantic-affective model estimates affective ratings as a weighted linear combination of semantic similarities between the unknown and the seed words, as follows:

$$\hat{v}(w_j) = \alpha_0 + \sum_{i=1}^{N} \alpha_i v(w_i) S(w_j, w_i), \tag{3.1}$$

where w_j is the unknown word, $w_{1..N}$ are the seed words, $v(w_i)$, α_i are the affective rating and the weight corresponding to the word w_i and $S(\cdot)$ is the semantic similarity metric between two words. The weights α_i are learned automatically. Criterion for selecting the seeds is the affective score of each word, according to the observation in [107] that good seed words have high absolute affective scores. An important advantage of the proposed affective model is the fact that it requires only a small affective lexicon in order to estimate affective ratings for any number of unknown words. Next we detail how the seed word weights α_i and the semantic similarity metrics are estimated.

3.3.1 Semantic model

The semantic model expresses how similar is each word to another. It is built based on a semantic similarity metric $S(\cdot)$ that is used in (3.1) in order to estimate the semantic similarity between two words. $S(\cdot)$ can be implemented within the framework of distributional semantic models (i.e., coprus-based models) that adopt the distributional hypothesis of meaning, i.e., *"similarity of context implies similarity of meaning"*, [65]. Having a vocabulary and a corpus in a target language, we can create a contextual feature vector for each vocabulary entry w_i, as follows. Lexical features are extracted after centering a contextual window of size $2H+1$ words on each instance of the target word w_i in the corpus. Then, a feature vector x_i is formulated by extracting the words in distance H from the window center.

In addition to the context-based metrics where second order word co-occurrences are used, the similarity of words can be also estimated by considering their first-order co-occurrence statistics. The underlying assumption is that the co-occurrence of words within a specified context serves as indicator for their semantic relatedness. A widely-used co-occurrence metric is Google-based semantic relatedness (G) and proposed in [63]. The word co-occurrence was regarded at the sentential level. A comparison of both types of metrics is presented in [75] for the task of similarity computation between nouns. The semantic similarity between two words w_i, w_j can be computed as the cosine of their respective feature vectors:

$$Q^H(w_i, w_j) = \frac{x_i . x_j}{||x_i||\ ||x_j||}. \tag{3.2}$$

The aforementioned feature extraction typically deals with words. However, it can be modified by extracting the character n-grams from the words that are captured by the applied contextual window. Usually, the selection of n depends on the desired amount of lexical information that n-grams carry. The elements of the feature vectors are set according to two schemes: 1) a weighting scheme based on the frequency of the features, 2) a binary scheme (B), where each element is set to one if the corresponding feature frequency is at least one and zero otherwise [74]. The weighting scheme we use is based on the point-wise mutual information (PMI). The PMI between a word w_i and the k–th feature of its vector x_i, f_i^k, is computed as shown in (3.3) [38].

$$\text{PMI}(w_i, f_i^k) = -\log \frac{\hat{p}(w_i, f_i^k)}{\hat{p}(w_i)\hat{p}(f_i^k)}, \tag{3.3}$$

where $\hat{p}(w_i)$ and $\hat{p}(f_i^k)$ are the occurrence probabilities of w_i and f_i^k, respectively, while the probability of their co-occurrence (within the H window size) is denoted by $\hat{p}(w_i, f_i^k)$. The corpus-based frequencies of lexical items (words or character n-grams) were used in order to compute the probabilities, according to maximum likelihood estimation. The scores derived using PMI lie in the $[-\infty, +\infty]$ interval. In particular, we use the positive point-wise mutual information (PPMI), in order to bound the computed scores within the $[0, +\inf]$ interval. PPMI is a special case of PMI according to which the negative PMI scores are set to zero, based on the assumption that the contextual features that exhibit negative PMI do not contribute to the estimation of similarity much [27].

3.3.2 Semantic-affective mapping

With the term semantic-affective mapping we mean the weights α_i of (3.1). These weights are used because not all seed words are equally salient for the estimation of affective ratings. The weights are automatically learned by solving the following linear system that consists of K linear equations with N unknown variables.

$$\begin{bmatrix} 1 & S(w_1,w_1)v(w_1) & \dots & S(w_1,w_N)v(w_N) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S(w_K,w_1)v(w_1) & \dots & S(w_K,w_N)v(w_N) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \begin{bmatrix} 1 \\ v(w_1) \\ \vdots \\ v(w_K) \end{bmatrix} \tag{3.4}$$

In (3.4) N is the number of the seed words, K is the number of the training words and α_i are the weights that are learned and assigned to each seed. The system consists of K linear equations with $N+1$ unknown variables including the bias. The optimal values of these variables can be estimated using Least Squares Estimation (LSE) or regularized LSE, i.e., ridge regression. The weights that were learned are then incorporated in (3.1) with the semantic similarities and the affective ratings of the seeds in order to estimate the affective ratings of unknown words. Another way to write the linear system of (3.4) is the following:

$$\boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{y}, \tag{3.5}$$

where $\boldsymbol{X}$ is a $K \times (N+1)$ matrix containing K training samples and $N+1$ features for each sample, $\boldsymbol{\alpha}$ is a $(N+1) \times 1$ vector including the α_i weights, while $\boldsymbol{y}$ is a $(K+1) \times 1$ vector containing the known affective ratings of the training words. According to LSE, the weights can be estimated as follows:

$$\hat{\boldsymbol{\alpha}} = \underset{\alpha}{\operatorname{argmin}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}||^2 \tag{3.6}$$

LSE may yield weights with large variance, so we investigate the use of Ridge Regression for alleviating this problem. Ridge Regression (RR) uses the estimator shown in (3.7), incorporating a regularization factor, λ, which forces the weights to shrink toward zero:

$$\hat{\boldsymbol{\alpha'}} = \underset{\alpha'}{\operatorname{argmin}} \big[||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha'}||^2 + \lambda||\boldsymbol{\alpha'}|| \big], \tag{3.7}$$

where α' is the weights vector. The λ values should be greater than zero, while for $\lambda = 0$ the LSE and RR estimators are identical [108].

3.4 Experimental Procedure and Results

3.4.1 Task, Corpora and evaluation metrics

The main task of the experimental procedure is the affective labelling of words using the semantic-affective model. Three affective dimensions, i.e., valence, arousal and dominance (V, A, D) and five different languages namely English, German, Greek, Portuguese and Spanish are being investigated.

A corpus per target language was harvested for the computation of semantic similarities. The corpora we used were created using web data as follows. The process starts with the definition of a vocabulary for each language: 135K, 332K, 407K, 125K, 187K entries for English, German, Greek, Portuguese and Spanish, respectively. For each word of the vocabulary a web search query was formulated and the snippets of 1K top-ranked documents were downloaded and aggregated. The derived corpora were used for the extraction of the feature vectors that were used for building the semantic model.

The affective lexica of each language were used for training and testing. All but German affective lexica consist of 1034 words annotated on valence, arousal and dominance. The German affective lexicon consists of 2902 words rated on valence and arousal. To simplify performance comparisons we selected a subset with 1034 words and similar ratings distributions with the rest of the lexica. The affective lexicon expansion is evaluated on each language's affective lexicon, applying 10-fold cross validation. In each fold 10% of the affective lexicon words is used as test (unknown words) and 90% is used as train. The model performance is captured through binary classification accuracy (positive vs. negative values) and Pearson correlation between the automatically estimated and the manually annotated valence ratings.

3.4.2 Semantic-affective model parameters

Within the experimental procedure we aim to investigate the impact of the semantic-affective model parameters to the accurate estimation of the affective labelling of words. The different parameters that are used in the experimental procedure are summarized in the following table.

Languages	English, German, Greek, Portuguese, Spanish
Affective dimensions	Valence, Arousal, Dominance
Semantic Model	Context vs. Co-occurrence similarity metrics
	Words vs. character n-grams contextual features
	Binary vs. frequency based weighting schemes
Affective model	Number of Seeds
	LSE vs. RR semantic-affective mapping
Evaluation Metrics	Pearson Correlation, Classification Accuracy

Table 3.1: Semantic-affective model parameters

The semantic-affective model is language and affective dimension independent. Thus, given that the required resources are available it can be applied to any language and affective dimension. The number of seeds incorporated in the affective lexicon expansion algorithm as well as the machine learning scheme that is adopted for the estimation of the semantic-affective mapping are major parameters for the estimation of the affective ratings of words.

Regarding the semantic features we investigate both context and co-occurrence semantic similarity metrics. Since context based semantic similarities have been proved to be more robust [] we employ them in order to conduct a deeper investigation of the semantic features. The contextual window size is set to $H = 1$ and variations of the contextual features and weighting schemes are used in order to create the similarity metrics described below. The metrics employ two types of contextual features, namely, words (W) and character n-grams (ngram). In the former case the contextual feature vectors of each word have been formulated assuming as context the words that occur in distance of one word. In the latter case, the contextual features are some of the characters of the words that occur in distance one word. Two types of weighting schemes, a binary and a frequency based, are also employed in order to set the elements of the contextual feature vectors as they were described previously. Early fusion schemes were also investigated, combining feature vectors that consist of different size n-grams, or of n-grams and words. In Table

3.2 we list the contextual features and the weighting schemes that are employed by each context-based similarity metric.

Similarity metric	*Contextual features*		*Weigh. scheme*	
	Words	*Character n-grams*	*PPMI*	*Binary*
W-B	✓	×	×	✓
W-PPMI	✓	×	✓	×
4gram-B	×	$n = 4$	×	✓
4gram-PPMI	×	$n = 4$	✓	×
2/3/4/gram-B	×	$n = 2, 3, 4$	×	✓
W+4gram-PPMI	✓	$n = 4$	✓	×

Table 3.2: Contextual features and weighting schemes for the semantic similarity metrics.

3.4.3 Results

We start our experimental procedure by comparing two languages, and two weighting schemes. In the figures below we show the correlation and the classification accuracy of the affective lexicon expansion algorithm (Valence) for Greek and English languages as a function of the seed words. The semantic model has been created using context based semantic similarities with word contextual features and either binary or PPMI-based weighting scheme.

In Figure 3.6(a,b) we show the classification accuracy (high vs. low values) and the Pearson Correlation for Valence, using binary and PPMI based weighting schemes. Both for English and Greek PPMI clearly outperforms the simpler binary feature vector scheme. The performance is similar for the two evaluation metrics, so in the next figure we show the Pearson correlation for the the two languages and the three affective dimensions V-A-D using the best weighting scheme. As observed by 3.6(c) the performance of the affective model is high, especially for valence and dominance[1], while the performance achieved for arousal is relatively poor. The results are consistent with respect to the V-A-D dimensions for both languages, although results for Greek are consistently lower than for English for all three dimensions[2]. The affective model appears to be robust and high performing when at least 100 seeds are used reaching highest performance for 500-600 seeds.

We investigate more the results shown in Figure 3.6(c) and we show the classification accuracy and the Pearson Correlation for valence, arousal and dominance and the five available languages, using context based semantic similarities and PPMI weighting scheme. The performance for each affective dimension and language as a function of the seed words, using the W-PPMI semantic similarity metric is shown in Figure 3.7. The affective model appears to be robust and well-performing when at least 100 seeds are used. The highest performance is reached for 500-600 seeds, for all the affective dimensions. Performance for valence (Figure 3.7(a), 3.7(d)) is consistent across all languages while the highest scores are obtained for English and Portuguese. Performance is consistent

[1] Valence and Dominance are highly correlated (0.86, 0.84 in Greek and English respectively).

[2] This may be attributed to the differences on syntax and morphology that exist between the two languages.

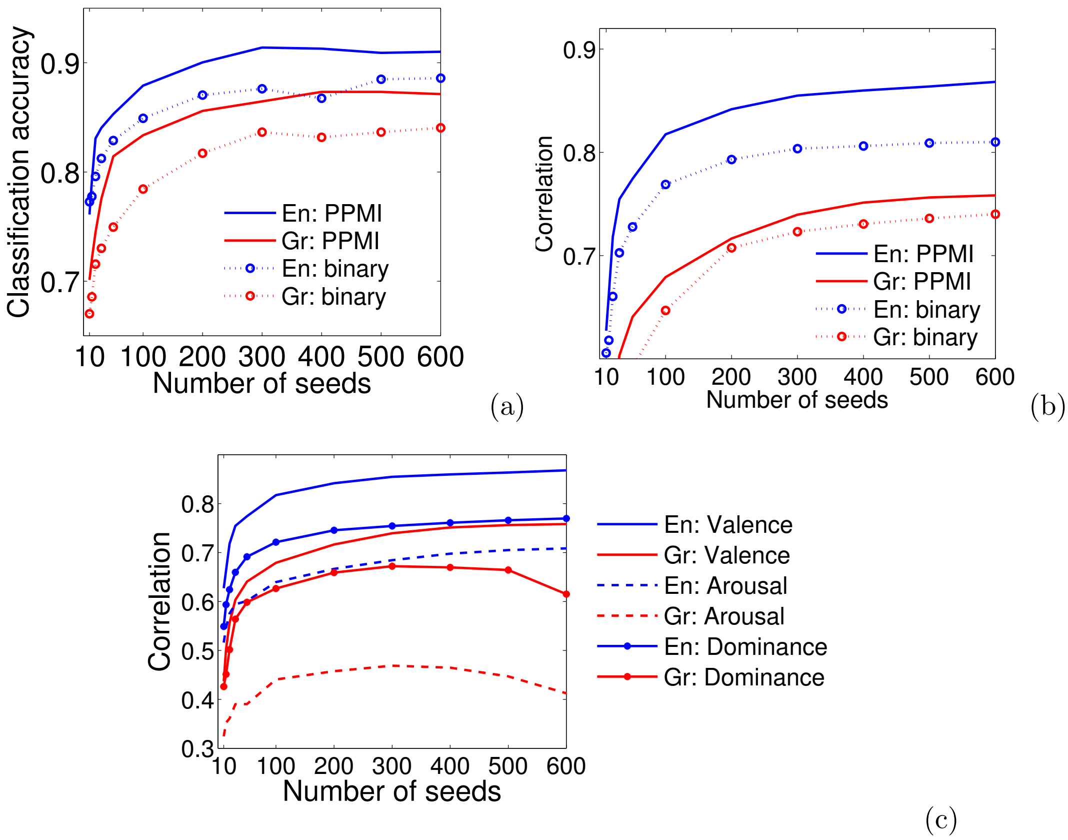


Figure 3.6: Valence classification accuracy (a) (high vs. low values), Pearson correlation (b) valence, arousal and dominance Pearson correlation (c) of PPMI weighting schemes for English (En) and Greek (Gr).

also for arousal (Figure 3.7(b), 3.7(e)) and dominance (Figure 3.7(c), 3.7(f))[3]. Among the three affective dimensions, the highest performance is achieved for valence, while the poorest performance is achieved for arousal[4]. The reasons for the differences in the results across languages are not easily interpreted, however they could be attributed to language morphology differences.

Up to now, we estimate the semantic-affective mapping adopting LSE and we have used only context-based semantic similarities. Thus, we investigate model's performance when co-occurrence-based similarity metrics are used. Regarding the semantic-affective mapping we employ both LSE and RR and we show the obtained performance in terms of correlation in the figures below. When the cooccurrence-based metric is combined with LSE we observe that the model lacks robustness, and performance drops especially for large number of seeds. However, when RR is used performance is robust and also the correlation curve is quite close to the one observed when context-based similarities are used. Correlation of Greek valence ratings when co-occurrence based semantic similarities are used in combination with RR is much higher than when using them with LSE. Still, it is much lower than the correlation when context-based similarities are used. Language

[3]Dominance ratings are not available for German.

[4]Valence and dominance are highly correlated in all four languages.

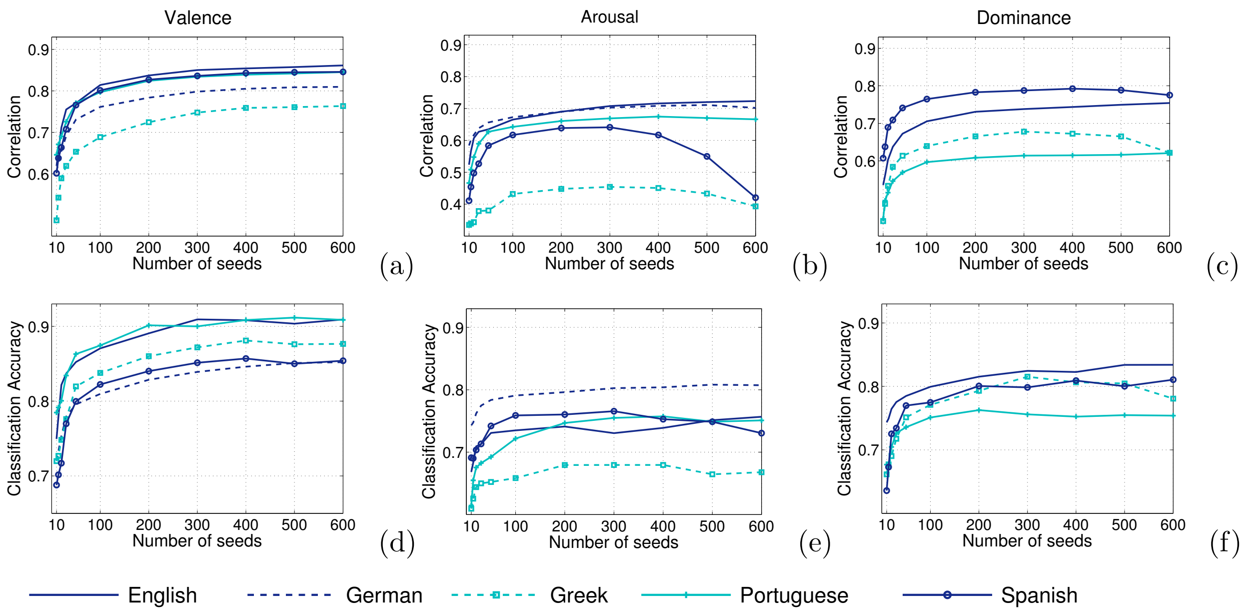


Figure 3.7: Performance of the affective-semantic model in terms of correlation (a, b, c) and classification accuracy (d, e, f) for valence (a, d), arousal (b, e) and dominance (c, f) across all five languages, using W-PPMI for similarity computation.

morphology and characteristics may be responsible for the performance differences between the two languages.

However, as seen in Figure 3.7, the performance of the model is not always robust for a large number of seeds, even if context-based semantic similarity metrics are used. We investigated whether the low performance of arousal for Spanish and Greek can be attributed to the weights estimation method. For this purpose, we use RR with $\lambda = 0.05$ (derived after tuning the parameter for both languages on arousal, W-PPMI and 600 seeds). The classification accuracy and the correlation of the arousal ratings for Spanish and Greek (the two languages with the lowest performance in arousal) using LSE and RR for 10 up to 900 seeds are shown in Figure 3.9(left) and Figure 3.9(right), respectively. It is observed that, as the number of seeds increases the arousal model that uses weights estimated with RR becomes superior to the model that uses weights estimated with LSE. The model that employs RR is robust to a large number of seeds compared to LSE. Additionally, we observed that the best performance achieved using 600 seeds, W-PPMI and LSE can be improved (up to 0.5-1%) for almost all languages when 900 seeds and RR are used.

Finally, we investigate the influence of the semantic models when different contextual features are used. We create various similarity metrics based on the contextual features and the weighting schemes, and we use the semantic affective model with 600 seeds and LSE for the semantic-affective mapping estimation. We report the Pearson Correlation and the classification accuracy for the task of valence ratings estimation. The similarity metrics and the obtained performance are reported in the following table.

The correlation and classification accuracy of the affective (valence) lexicon expansion across the different languages is shown in Table 3.3. In addition to the reported similarity metrics, W-normalized PPMI was also investigated (with no significant perfor-

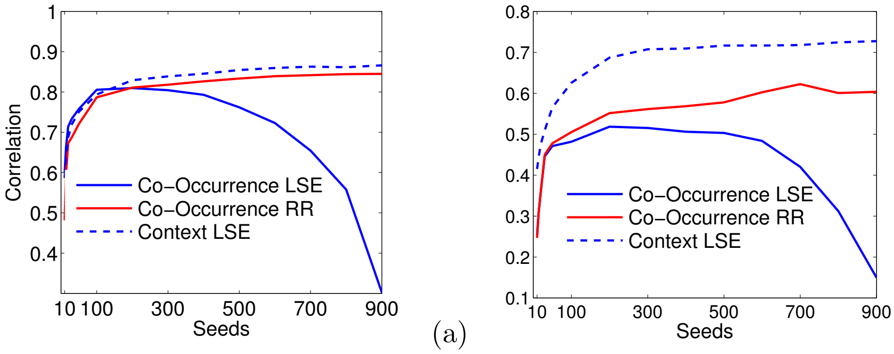


Figure 3.8: Correlation of automatically estimated and human collected valence ratings for English (a) and Greek (b) using different semantic similarity and weights estimation methods

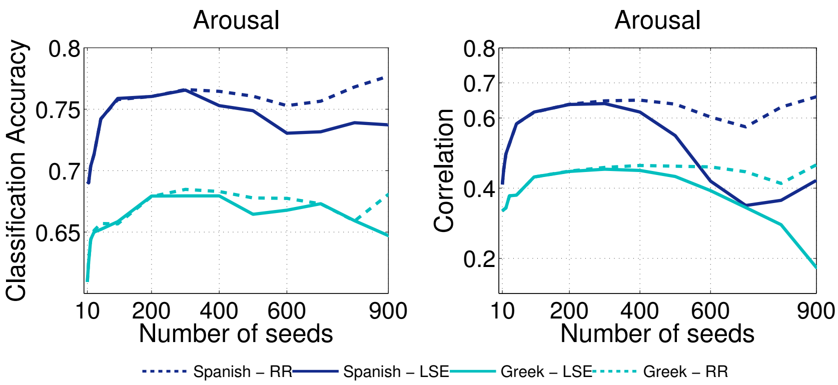


Figure 3.9: Classification accuracy (left) and correlation (right) for arousal using LSE and RR for the estimation of weights.

mance improvement), as well as 2grams-B and 3grams-B (achieving lower performance than 4grams-B). Although the performance is consistently high for all languages and semantic similarity metrics, minor differences exist between them. The PPMI weighting scheme is superior to binary in all cases, especially for German. The highest performance is achieved for English and European Portuguese when the feature vector consists of the contextual words. The use of context character n-grams yields a slight improvement to performance especially for morphologically rich languages when the binary scheme is used. The early fusion scheme that uses both the contextual words and 4-grams achieves the highest performance for almost all languages.

3.5 Cross language affective labelling

Up to now we have shown that given that manually annotated affective lexica are available for a language, they can be used for bootstrapping the semantic-affective model of (3.1). As we mentioned in the previous section, if no affective lexicon is available for a language the semantic-affective model cannot be applied for that language and this is the main

Similarity Metric \ Language	English		Greek		Spanish		Portuguese		German	
	PC	CA(%)	PC	CA(%)	PC	CA(%)	PC	CA(%)	PC	CA(%)
W-B	0.80	86.9	0.74	84.3	0.84	85.9	0.82	89.3	0.68	77.1
W-PPMI	**0.86**	**90.9**	0.76	87.6	0.84	85.3	**0.84**	**90.8**	**0.80**	**85.2**
4gram-B	0.82	87.8	0.77	87.8	0.84	86.4	0.80	87.6	0.78	82.3
4gram-PPMI	0.84	89.8	**0.78**	**87.5**	**0.85**	**87.7**	0.82	87.4	**0.80**	82.6
2/3/4gram-B	0.82	88.1	0.75	87.6	0.83	86.6	0.80	86.4	0.78	82.2
W+4gram-PPMI	**0.85**	**90.5**	**0.79**	**87.2**	**0.85**	**87.9**	**0.83**	89.3	**0.80**	83.0

Table 3.3: Classification Accuracy (CA) and Pearson Correlation (PC) between the manually rated and the automatically estimated valence scores for 600 seeds, across the five languages using various semantic similarity metrics.

reason why we created the Greek affective lexicon. However, we claim that we should be able to estimate the affective labels of words for languages that no affective lexica are available. This claim is due to the fact that creating an affective lexicon is a time consuming and should be done by native speakers of each language.

3.5.1 Experiments and Results

In order to estimate the affective ratings of words for a language without a manually annotated affective lexicon, we create a cross language affective lexicon as follows. We translate the words of another language affective lexicon (source language) in the language of interest (target language) and we use the affective ratings of source language's lexicon. The motivation behind this approach is that words across languages have the same emotional content and thus the affective ratings can be transferred. This approach can be more generic if we create new affective ratings, combining more than one source languages as follows:

$$\hat{\upsilon}_c(w_t) = \frac{1}{K}\sum_{i=1}^{K} \upsilon(w_{s_i}), \tag{3.8}$$

where w_t is the word of the target language and $\hat{\upsilon}_c(w_t)$ is the affective rating of this word estimated by the cross language model, K is the number of the source languages used, w_{s_i} is the word of the source language i and $\upsilon(w_{s_i})$ is the affective rating of the word (extracted from the affective lexica). Having created the cross language affective lexica we evaluate them estimating affective ratings using (3.1) and we show the performance as a function of the seed words. Classification accuracy for the polarity detection task (positive vs. negative valence) is shown in Fig. 3.10 , where Portuguese is the target language and English, Greek, and Spanish have been used as source languages.

The cross language model achieves consistent performance with the classic semantic-affective model, however we can notice differences across the languages. For example cross language models always outperform the Spanish semantic-affective model. This could be an indication that the Spanish affective ratings are problematic. In contrast, non cross language model can outperform the Greek semantic-affective model and a major difference in classification accuracy is noticed. This could be an indication that the Greek words are very different from the rest languages' works, e.g., they have multiple senses and meanings

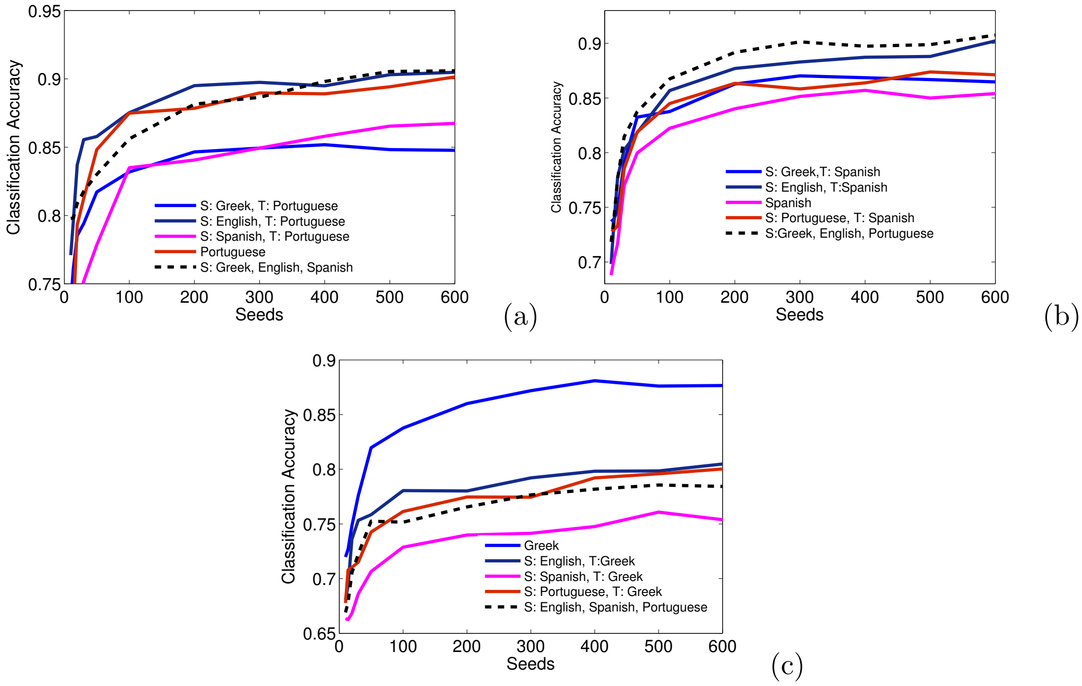


Figure 3.10: Valence cross language classification accuracy for Portuguese (a), Spanish (b) and Greek (c)

and the affective ratings of the other languages are not representative, however further analysis and investigation of this phenomenon is required. Regarding Portuguese both of the aforementioned phenomena are noticed.

3.6 Summary

In this chapter we introduced the problem of the affective estimation of textual units and we undertook the task of affective labelling of words. We analysed how the semantic-affective model works and what are the external required resources. We created the first affective lexicon for the Greek language. The ratings we collected were consistent with other languages' affective lexica. We tested our affective model both on English and Greek for the three affective dimensions and we achieved consistent results. We also developed a cross language model for the case that no affective lexicon is provided for a language.

We expanded the affective lexica of five languages, namely, English, German, Greek, Portuguese and Spanish for the three affective dimensions valence, arousal and dominance utilizing semantic models. Our approach was found to be applicable across all languages and affective dimensions. Minor differences in performance could be attributed to the linguistic properties of each language e.g., morphology. We investigated various parameters for the context-based computation of semantic similarity observing that the character 4-grams (extracted from the contextual words) are salient features for this task. We showed that the performance of the affective model depends on the weights estimation method, especially for a large number of seeds. Specifically, the employment of a more robust approach for the estimation of weights such as ridge regression leads to small performance improvements. Highest performance improvement is achieved when ridge regression is used with co-occurrence based semantic similarities.

Chapter 4

Compositional models for affective labelling of word pairs

News Baby pandas ! ***Baby pandas!*** *Baby pandas!* (positive emotion)
Bathing mom awakes to find ***baby dead****.* (negative emotion)

Real news headlines about babies from news headlines dataset of [188].
Their emotion changes depending on the word that follows the word **baby**.

4.1 Introduction

In the previous chapter we described how word-level affective lexica can be created automatically created. Their creation is based on word-level semantic representations and they are the building blocks for models of larger lexical units e.g., phrases and sentences, following the principle of semantic compositionality [150]. Word-level semantic representations, that are required for the affective lexica creation, constitute the core aspect of DSMs typically constructed from co-occurrence statistics of word tuples. Word semantics have been represented efficiently with vector space models, as shown in [10, 107], however representing the semantics of lexical structures larger than words is not trivial [15]. The reason is that the meaning of complex structures derives from various compositional phenomena [150].

Semantic compositionality allows the construction of complex meanings from simpler elements based on the principle that the meaning of a whole is a function of the meaning of the parts [148]. The key characteristic of compositionality is that the meanings of the constituent parts are combined into a single token [77, 117]. Compositional approaches in vector-based semantics can be modelled by applying a function f that acts on two constituents a, b in order to produce the compositional meaning p. Functions that were investigated by [77, 117] are addition and multiplication. The additive compositional model takes the sum of the two vectors weighted with the appropriate weight matrices A and B respectively ($p = Aa + Bb$) and the multiplicative model is the projection of the ab tensor product using a weight tensor C ($p = Cab$). These composition forms can be also simplified using the additive model with scalars instead of matrices. Similarly the multiplicative approach can be reduced to component-wise multiplication. Motivated by compositionality modeling [15] proposed an approach that focuses on adjective-noun composition according to which nouns are represented as vectors and adjective as functions.

4.2 Compositional Application

We propose that similarly to the semantic modification that underlies Compositional Distributional Semantic Models (CDSMs), affective modification may occur within the framework of affective spaces. In order to validate the assumption that compositional phenomena appear when at least two words coexist, we propose models that estimate continuous affective ratings (valence, arousal, dominance) of adjective-nouns (AN) and noun-nouns (NN) word pairs.

We assume that the affective rating of each word pair may derive either with a compositional or with a non-compositional manner. However, we don't know apriori the compositionality degree of each word pair and in order to handle this uncertainty we propose both compositional and non-compositional (semantic-affective) models. The semantic-affective models are motivated by the assumption that "*semantic similarity implies affective similarity*", while the proposed compositional model is motivated by the CDSM proposed by [15], that focuses on adjective-noun composition and represents adjectives as functions and nouns as vectors. Then, the affective ratings estimated by the compositional model are compared and combined with the ones obtained by the non-compositional semantic-affective models. Similar fusion schemes to [58] that aim to capture the compositionality degree of each word pair are investigated.

Each word of the word pair is assigned a specific role, i.e., the first word is called *modifier* and the second word *head*. The names of the words derived by the assumption that modifier alters the affective meaning of the head. A word pair (also referred as test pair) p is defined as:

$$p = m.h, \tag{4.1}$$

where m is the modifier and h is the head.

4.3 Proposed compositional approach

The assumption of our affective compositional model is that the *modifier* modifies the *head*'s affective rating in order to estimate the word pair's affective rating. Modifier's impact is defined by a weight coefficients matrix or weight scalars depending on the dimensions of the compositional model. The weight coefficients are learned in a distributional approach, according to which K pairs are extracted from a large corpus and serve as the training set of each modifier. The training pairs contain the same modifier with the current test pair and different head, i.e.,

$$p' = m.*, \tag{4.2}$$

where p' is the training word pair, m is the modifier and $*$ indicates any head except from the test pair's p head. For each test pair p, K training pairs p' are extracted.

The semantic-affective model of (3.1) is then employed to estimate the affective ratings of the training heads and the training pairs. The estimated affective ratings are incorporated in an LSE formulation where the ratings of the training pairs formulate the dependent variable and the ratings of the training heads formulate the independent variable. The proposed compositional approach is depicted in Figure 4.1. Training is performed sepa-

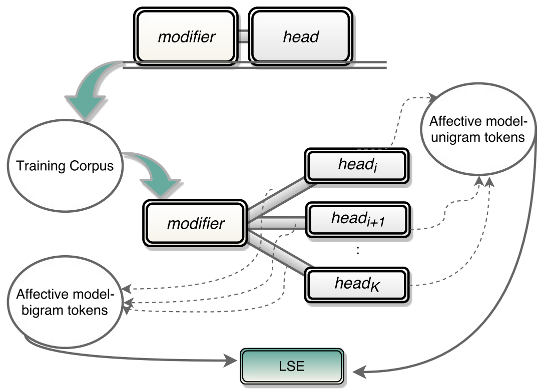


Figure 4.1: Compositional system overview.

rately for each modifier and the impact of each modifier is estimated through supervised learning, e.g., LSE. Finally, the continuous affective ratings of the word pair are estimated via an additive model. This model takes directly into consideration only the head, while the modifier's impact is captured by the weight matrix. The general compositional model is shown below:

$$\hat{\upsilon}_c(p) = \vec{\beta} + W\hat{\upsilon}(h), \tag{4.3}$$

where $\hat{\upsilon}_c(p)$ is the compositional affective rating of the word pair p, $\vec{\beta}$ is the bias vector, W is the coefficients matrix and $\hat{\upsilon}(h)$ is the affective rating of the head, estimated via (3.1). The compositional model may not always be the most appropriate for the estimation of a word pair's affective ratings. Thus, in order to measure the compositionality degree of each word pair the Mean Squared Error (MSE) of the model is measured. Specifically, the distance between the compositional affective ratings and the affective ratings is estimated via (3.1) and averaged over all training pairs:

$$MSE(p) = \frac{1}{K}\sum_{j=1}^{K}(\hat{\upsilon}(p'_j) - \hat{\upsilon}_c(p'_j))^2, \tag{4.4}$$

where $MSE(p)$ is the MSE estimated for each pair p, K is the number of the training pairs p', $\hat{\upsilon}(p'_j)$ is the affective rating of the training word pair p'_j estimated using (3.1) for bigram tokens and $\hat{\upsilon}_c(p'_j)$ is the corresponding affective rating estimated using the compositional model described in (4.3).

4.3.1 3D compositional model

Here we assume that three affective dimensions (valence, arousal, dominance) contribute to the affective content of the word pairs. The 3D compositional model (com3D) is a special case of the general compositional model of (4.3), where $W \in \mathbb{R}^{3\times3}$ and contains the weight coefficients for the three affective dimensions and the bias vector $\vec{\beta} \in\ mathbbR^{3\times1}$ and contains the bias of each affective dimension. The coefficient matrix W and the bias vector $\vec{\beta}$ are estimated using LSE. Specifically, for each test pair p with K training pairs p' K linear equations with 4 unknown variables (3 for the affective dimensions and 1 for the bias) have to be solved. The linear system is shown in 4.5.

$$\begin{bmatrix} 1 & \hat{v}(h_1) & \hat{a}(h_1) & \hat{d}(h_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \hat{v}(h_K) & \hat{a}(h_K) & \hat{d}(h_K) \end{bmatrix} \begin{bmatrix} w_{10} & w_{20} & w_{30} \\ w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \end{bmatrix} = \begin{bmatrix} \hat{v}(m.h_1) & \hat{a}(m.h_1) & \hat{d}(m.h_1) \\ \vdots & \vdots & \vdots \\ \hat{v}(m.h_K) & \hat{a}(m.h_K) & \hat{d}(m.h_K) \end{bmatrix} \tag{4.5}$$

The columns of the weight matrix correspond to valence (w_{1*}), arousal (w_{2*}) and dominance (w_{3*}), while the first row is the bias. With $\hat{v}(\cdot)$, $\hat{a}(\cdot)$, $\hat{d}(\cdot)$ we denote the valence, arousal and dominance affective ratings respectively.

4.3.2 1D compositional model

The 1D compositional model (com1D) is a special case of the 3D compositional model and follows the assumption that each affective dimension is independent of the other affective dimensions. The dimensionality of the model is reduced and thus the behaviour of the modifier, i.e., the coefficient matrix W and the bias vector $\vec{\beta}$ are substituted by two scalars. The general compositional model of (4.3) is transformed to the 1D compositional model using scalar variables. These two scalar coefficients are estimated similarly to the 3D model.

4.4 Fusion of affective models

The same word may have different contribution to the affective content of a phrase depending on the context. For example, when the word “dog” appears in the word pair “happy dog” the conveyed affect is positive, which it is reversed when it appears in the word pair “dead dog”. The accuracy of the proposed affective models depends on the compositionality degree that characterize the pairs of interest. By combining different models we aim to achieve more accurate affective scores. In the next subsection we briefly describe the non compositional models that are used for fusion.

4.4.1 Unigram and Bigram Affective Models

Except from the compositional model we propose two semantic-affective models in order to estimate affective ratings in a non-compositional manner. The unigram and the bigram affective models are based on the semantic affective model that described in 3.1 of Section 3.3. The unigram affective model (U) is based on the assumption that the two words that constitute the word pair contribute equally to its affective content and it is defined as follows:

$$\hat{v}_U(p) = \frac{\hat{v}(w_1) + \hat{v}(w_2)}{2}, \tag{4.6}$$

where p is the word pair, and $\hat{v}(w_1)$ and $\hat{v}(w_2)$ are the affective ratings that are estimated using (3.1). The bigram semantic-affective model (B) handles each word pair as a single token, i.e., $\hat{v}_B(p) = \hat{v}(w_1 w_2)$, where p is the word pair and $\hat{v}_B(p)$ is the affective rating of the word pair that it has been estimated using (3.1) for the bigram lexical token $t = w_1 w_2$.

4.4.2 Fusion schemes

The first fusion scheme combines the affective ratings estimated by the compositional and semantic-affective models. The underlying assumption is that all models contribute equally to the affective meaning of a word pair, that is a word pair exhibits both compositional and non-compositional aspects. This scheme is based on the averaging of affective ratings defined as follows:

$$\Phi_{avg}(p) = \frac{1}{M}\sum_{i=1}^{M}\hat{\upsilon}_i(p), \tag{4.7}$$

where $\Phi_{avg}(p)$ is the fused affective rating of the word pair p, M denotes the number of fused models and $\hat{\upsilon}_i(p)$ stands for the estimated affective rating of word pair p, i.e., $\hat{\upsilon}_1(p)$ = $\hat{\upsilon}_U(p)$, $\hat{\upsilon}_2(p)$ = $\hat{\upsilon}_B(p)$, $\hat{\upsilon}_3(p) = \hat{\upsilon}_{c1}(p)$, $\hat{\upsilon}_4(p) = \hat{\upsilon}_{c3}(p)$. A weighted variant of (4.7) was also investigated, as follows:

$$\Phi^{w}_{avg}(p) = \frac{1}{\sum_{i=1}^{M} w_i}\sum_{i=1}^{M} w_i\hat{\upsilon}_i(p), \tag{4.8}$$

where $\Phi^{w}_{avg}(p)$ is the fused affective rating of the word pair p, $\hat{\upsilon}_(p)$ stands for the affective rating of word pair p estimated by the affective models, as explained in (4.7) and w_i are weights estimated via linear regression. Motivated by the fusion scheme proposed in [58], we propose the use of MSE weight appropriately the compositional and the semantic-affective models. MSE is estimated during training phase as shown in (4.4) and the weight parameter is defined as $\lambda'(p) = \frac{0.5}{1+e^{-MSE(p)}}$, where $MSE(p)$ is the MSE measured for each word pair p and $\lambda'(p)$ is estimated for each word pair p based on the compositional model. (4.4.2) is applied both on 1D and 3D compositional models and the derived $\lambda'(p)$ are averaged in order to formulate the parameter $\lambda(p)$ that is used in the fusion scheme as follows:

$$\Phi^{\mathrm{MSE}}_{avg}(p) = \lambda(p)(w_1\hat{\upsilon}_B(p) + w_2\hat{\upsilon}_U(p)) + (1 - \lambda(p)(w_3\hat{\upsilon}_{c1}(p) + w_4\hat{\upsilon}_{c3}(p)), \tag{4.9}$$

where $w_1, ...w_4$ are the weights that correspond to each affective model and estimated through LSE and $\sum_{i=1}^{4} w_i = 1$, $\hat{\upsilon}_*(p)$ is the affective rating of each word pair derived from each affective model, and $\lambda(p)$ is meant for weighting the contribution of compositional and non-compositional models.

4.5 Experimental Procedure and Results

For evaluating the proposed semantic-affective models we use two word pairs datasets, one consisting of AN and one consisting of NN. The word pairs of the evaluation datasets were extracted from movie reviews by [185] as follows. Each movie review was first split into the constituent sentences and then into the constituent phrases. The derived sentences and phrases were annotated with respect to their polarity using crowdsourcing. We kept only the word pairs that have an adjective or noun as their first word, and their second word is a noun. The created dataset consists of 1009 AN and 357 NN pairs.

The proposed models estimate the affective ratings for each affective dimension in a continuous scale in [-1,1], however we only report results for valence. The semantic-affective model shown in 3.1 was applied for the unigram (U) and bigram (B) models as defined in Section 4.4.1. The parameters of the semantic-affective model are set as follows: $N = 600$ seeds, for $S(\cdot)$ a context-based metric of semantic similarity was applied with window size equal to one, while the extracted features were weighted according to positive pointwise mutual information [38]. LSE was applied for estimating the weights α of 3.1. The compositional model requires a large corpus for extracting the training pairs of each modifier [73]. We use a web harvested corpus that was created posing one query that was formulated for each vocabulary word on search engines and downloading and aggregating the snippets of 1K top-ranked document. For each modifier all word pairs with the same modifier are extracted creating hundreds of training pairs.

We investigate both weighted and unweighted average schemes while a compositionality criterion based on the compositional models' MSE is also incorporated. In (4.7) the average of all affective models is estimated, while a weighted version of this scheme is shown in (4.8). LSE was adopted in order to estimate the weights that capture each model's contribution. $\Phi^{\mathrm{MSE}}_{avg}$ was implemented in a two-fold cross validation scenario. Moreover, we introduce a compositionality criterion based on the MSE that was measured during the compositional training process. Then the weighted average of the compositional and the non-compositional semantic-affective models were estimated as shown in (4.9). The weights w_i were estimated through LSE and two-fold cross validation. The parameter $\lambda(p)$ used in (14) was computed as the average of the corresponding parameters that are estimated for the 1D and the 3D compositional models.

4.5.1 Results

We compare the valence ratings that were automatically estimated against the human valence ratings that were collected via crowdsoursing. We report evaluation results based on three evaluation metrics, namely, Pearson correlation coefficient between the estimated and the human valence ratings, binary classification accuracy (positive vs. negative valence ratings) and F-measure.

Aff. Model	**Correlation**		**Acc. (%)**		**F-measure**	
	NN	**AN**	**NN**	**AN**	**NN**	**AN**
Chance	-	-	76.4	74.1	-	-
U	0.581	0.573	84.3	80.0	0.903	0.874
B	0.507	0.451	76.8	74.6	0.841	0.824
com1D	0.523	0.552	79.2	77.2	0.880	0.866
com3D	0.538	0.552	79.8	78.1	0.883	0.870
Φ_{avg}	0.624	0.609	**86.2**	80.9	**0.916**	**0.882**
Φ^{w}_{avg}	**0.630**	0.608	85.7	**81.3**	**0.911**	**0.882**
$\Phi^{\mathrm{MSE}}_{avg}$	0.624	**0.613**	85.5	80.9	**0.912**	**0.883**

Table 4.1: Performance of affective models for valence estimation.

The evaluation results are reported in Table 4.1 for several semantic-affective models. Regarding individual models, the highest performance is achieved by the unigram model for both AN and NN. This may be attributed to the very good performance at the unigram model as reported in [107, 139]. However the fact that when moving from words to word pairs the performance drops by about 11% is a strong indicator of the need a compositional modeling. As expected, the accuracy of compositional models is between the accuracy of two semantic-affective models. Similar results have been also obtained for compositional models in the semantic space [58]. Using fusion schemes that combine the compositional with the semantic-affective models we can achieve the best performance exceeding the performance of all individual models. Simple fusion schemes such as average of the affective ratings derived from the individual models can increase the performance of the best model up to 5% in terms of correlation. Similar performance increase is observed for the rest of the evaluation metrics as well.

4.6 Summary

We proposed a compositional model for estimating continuous affective ratings of AN and NN structures consisting of words that formulate modifier-head pairs. The composition was motivated by the affective interaction of modifier and head words, while it was implemented as a affine operation in the continuous affective space. The compositional models were compared and fused with two semantic-affective models defined at the unigram and bigram level. The best performance overall was achieved by the fusion-based approach suggesting that there is not a single model that works for all word pairs, i.e., the degree and type of compositionality is different for each word pair.

Chapter 5

Affective labelling of sentences

"Language is the expression of ideas
by means of speech-sounds combined into words.
Words are combined into sentences,
this combination answering to that of ideas into thoughts."
Henry Sweet

5.1 Introduction

Up to now we have presented models able to estimate the affective ratings of words small phrases. However, most times people express their opinions, or talk creating sentences and utterances. Affective labelling of sentences is a challenging task, since each sentence has different syntactic structure and writing style. Also, ironic and humorous sentences contain language phenomena that should be modelled appropriately. Moreover, sentences that come from different resources may have totally different form and style, making their affective labelling almost a different task. For example tweets and news headlines are both written sentences but with different form and style. Similarly, movie subtitles and spoken dialogue utterances are both transcriptions of dialogues but with very different form and style.

5.2 News headlines

News headlines are emotion conveying sentences and they are short, i.e., approximately 6-7 words on average. Headlines language is compressed and usually short and rhyming words are preferred. Many times news headlines are incomplete sentences and the syntactic may be not formal, e.g., headlines with noun phrases and no verbs, headlines that contain only nouns. Headlines serve the role of offering the reader the chance to choose and to stimulate the interest of the reader for the content of the article.

We used the news headlines dataset provided in [188]. This dataset contains a test dataset (1000 news headlines) and a trial dataset (250 news headlines). The news headlines are annotated on polarity level (positive - negative)[1]. The test dataset consists of 2815 unigrams and 4873 bigrms.

Two different approaches were used for estimating the polarity of each headline. The first one is based on the aggregated fusion of the affective ratings of unigrams and bigrams while the second builds classifiers using affective and lexicosyntactic features. We use the semantic - affective model described in 3.1 in order to estimate the valence of each unigram and bigram in the headline. Then, we fused the affective ratings of the n-grams following three different schemes, as proposed in [107], i.e., (a) average, (b) weighted average, (c) maximum.

The idea of the average fusion scheme is that the affective rating of a sentence is the average of its components. We average affective ratings for the three aforementioned cases of contribution, thus:

$$\upsilon_a(h) = b_0 + b_1 \frac{1}{N} \sum_{i=1}^{N} \upsilon(w_i), \tag{5.1}$$

where $\upsilon_a(h)$ is the valence of the news headline, N is the number of n-grams in the headline, $\upsilon(w_i)$ is the valence of the token w_i and b_0, b_1 are trainable weighs (estimated using the trial dataset).

However, since not all words contribute equally to the affective strength of a sentence, a weighted average fusion scheme is also proposed and terms with higher absolute valence are weighted more.

$$\upsilon_w(h) = b_0 + b_1 \frac{1}{\sum_{i=1}^{N} |\upsilon(w_i)|} \sum_{i=1}^{N} \upsilon(w_i)^2 \cdot sign(\upsilon(w_i)), \tag{5.2}$$

where $\upsilon_w(h)$ is the valence of the news headline estimated with the weighted average scheme, N is the number of n-grams in the headline, $\upsilon(w_i)$ is the valence of the token w_i, $sign(\cdot)$ is the sign of the token's valence and b_0, b_1 are trainable weighs (estimated using the trial dataset).

According to the maximum absolute value fusion model, the highest absolute valence dominates the affective meaning of the headline.

$$\upsilon_m(h) = b_0 + b_1 \underset{i}{max}(\upsilon(w_z)) \tag{5.3}$$

$$z = \underset{i}{argmax}(|\upsilon(w_i)|), \tag{5.4}$$

where $\upsilon_m(h)$ is the valence of the news headline estimated with the maximum absolute valence fusion scheme, z is the token with the maximum absolute valence, $\upsilon(w_z)$ is the valence of the token w_z and b_0, b_1 are trainable weighs (estimated using the trial dataset).

For each fusion scheme we assume the three different cases. In the first one instead of keeping all the words of the news headline we can apply a POS-tag filtering in order to keep only the content words, i.e., adverbs, noun, adjective, verbs (ANAV), or only the unigrams or only the bigrams. The results obtained with the three fusion schemes are shown in the next table.

[1] The dataset is also annotated on the six basic emotions, but we don't propose a model for predicting discrete emotions

Affective Model	Classification Accuracy (%)		
Chance	52.6		
	ANAV	1-gram	2-gram
Average	**72.4**	70.9	58.8
Weighted Average	71.6	**73.1**	**59.5**
Maximum absolute valence	67	66.4	53.4

Table 5.1: Classification accuracy on news headlines using affective ratings fusion schemes

No major performance differences are observed between the fusion schemes, however the weighted schemes clearly outperform he maximum absolute valence scheme. Lowest performance is observed when the news headline's valence is estimated assuming only the bigrams and this is expectable, since the semantic affective model works better for words as it is designed (seed lexica contain words).

Regarding the second approach that requires the building of a classifier, both affective and lexicosyntactic features are used. Lexiconsyntactic features include i) part of speech tags (POS), i.e., the number of nouns, verbs, adverbs, adjectives, prepositions, ii) POS pronouns, i.e., the number of first and second person pronouns and the total number of pronouns, iii) word based, i.e., the number of words, the words in upper case, the number of long words (more than six characters), iv) punctuation, i.e., the number of the question marks, commas, periods and exclamation marks v) other features that contains the number of self references, the number of negative particles (e.g., n't), the number of interjections (e.g., Uh-huh) and the number of articles.

Regarding the affective features we did not use the semantic affective model, but instead we use the conjunction of two manually annotated affective lexica, i.e the anew [24] and the general inquirer [187]. The motivation behind this was to keep the "strong" affect bearing words. Then we use as features the number of the words with positive valence, the number of the words with negative valence and their normalized distance, i.e., $\frac{\#ofpositivewords - \#ofnegativewords}{\#ofwords}$.[2]

We experimented training various machine learning classifiers using the aforementioned features. In the following table we show the results obtained training a Naive Bayes, a Naive Bayes tree, a Support Vector Machine and a Random Forest Classifier. We apply 10-fold cross validation and we show the binary classification accuracy obtained for each classifier and feature set

Feature set	Classification Accuracy (%)			
Chance	52.6			
	Naive Bayes	Naive Bayes tree	SMO	Random Forest
Unigram Valence	72.7	70.3	69.6	69.7
Lexicosyntactic	53.7	53.3	54.4	53.4
Lexicosyntactic + affective	60.5	59.6	62.3	58.9

Table 5.2: lassification accuracy on news headlines building classifiers with lexicosyntactic and affective features

[2]Some of the features may seem not very appropriate for this dataset, however this model was also applied to other datasets, e.g., call center dialogues, stories and product reviews.

As expected, lexicon syntactic features are weak and do not contribute to a more accurate affective rating estimation. Even when early fusion scheme occurs creating feature vectors with lexicosyntactic and affective features, performance stays low. Not significant differences among the different classifiers were observed, indicating that classifiers based on feature independence like Naive Bayes are suitable for this task.

The first fusion scheme that is based on the valence ratings of the headline's words is more appropriate for the valence estimation of the whole headline. The headlines usually contain affect conveying words, thus if a model that estimates the valence of the constituent words accurately is available the headline's valence will be accurate too.

5.3 Movie Subtitles

Movie subtitles are spoken dialogue utterances that have been transferred from speech to text modality through transcribing process. Movie subtitles are transcriptions that potentially convey emotions, however the challenge is that the subtitles come from an audio visual environment that is lost in transcription.

The subtitles were taken from 12 30-minute movie parts that were selected from [104]. These movies are ten winners of the Academy Award for best picture for the years 1998-2007 and two award winning animation films, namely; "Shakespeare in Love", "American Beauty", "Gladiator", "A Beautiful Mind", "Chicago", "The Lord of the Rings: The Return of the King", "Million Dollar Baby", "Crash", "The Departed", "No Country for Old Men", "Ratatouille" and "Finding Nemo". Using the Academy Award winners list is one way of ensuring the high quality of the movies by a well-acknowledged criterion. One expected effect of this perceived quality is the higher correlation between intended and expected emotion [3]; a high quality movie is expected to be successful in creating the desired emotional experience.

5.3.1 Crowdsourcing for annotating subtitles on valence

No annotations were provided with the subtitles, thus we employed crowdsourcing in order to annotate the subtitles on valence. We create three crowdsourcing tasks in order to collect valence ratings for the subtitles of the movies described in the previous section. The tasks were uploaded on Crowdflower crowdsourcing platform [1]. Crowdflower is a platform that provides Europeans access to the many crowdsourcing channels from all over the world, such as Amazom Mechanical Turk. More about crowdsourcing and crowdflower can be found in [138, 141].

In order to annotate the subtitles we created a crowdsourcing task that the annotators had to give a valence rating for each subtitle [4]. The 12 movies contain 4295 subtitles that were merged by a preprocessing step in order to create utterances with complete meanings, forming 3286 subtitles. In order to ensure that we collect good quality annotations we

[3] Intended emotion describes the emotional response that the movie attempts to evoke in its viewers, experienced emotion describes the emotion a user actually feels when watching the movie, while expected emotion is the expected value of experienced emotion in a population.

[4] We also created two more tasks that with (i) annotation of subtitle part level and (ii) annotation on subtitle pair level. The collected ratings were used in order to find out whether there is any correlation between the valence of subtitle parts. One finding was that we observed correlation between the subtitle and the second subtitle part for task(ii). This could be an indicator that the last part of the sentence is more important for the valence of the sentence

incorporate in our experimental procedure Crowdflower's mechanism for quality control Test Questions, a.k.a. Gold Data. Test Questions are units that are pre-labelled with known answers that are randomly inserted throughout the task. They ensure that only the contributors demonstrating competency in the task are allowed to submit annotations. The cost of collecting 19,125 valence annotations in total was approximately $37, i.e., for each subtitle we collected five annotations. The average of these annotations was the final annotation of the subtitle.

5.3.2 Experimental procedure

For the automatically valence estimation of each subtitle we used the semantic affective model of 3.1. We apply the leave-one-movie-out technique, i.e. the subtitles of a given movie are the test set and the subtitles of the rest movies are the training set of the given movie. This technique is repeated for all the movies, and the results presented are the average results of the 12 movies. We must note that this technique is necessary only for predicting two coefficients, using linear regression that rescale the valence score predicted as follows:

$$v'(s) = \beta_0 + \beta_1 v(s), \tag{5.5}$$

where $v'(s)$ is the final valence of the subtitle, $v(s)$ is the valence of the subtitle that derives after deriving the valence of the constituent words and β_0, β_1 are the weights that were learned on the other movies' subtitles.

In the next figures we show the average classification accuracy and correlation as a function of the seed words that were used for the semantic affective model (3.1).

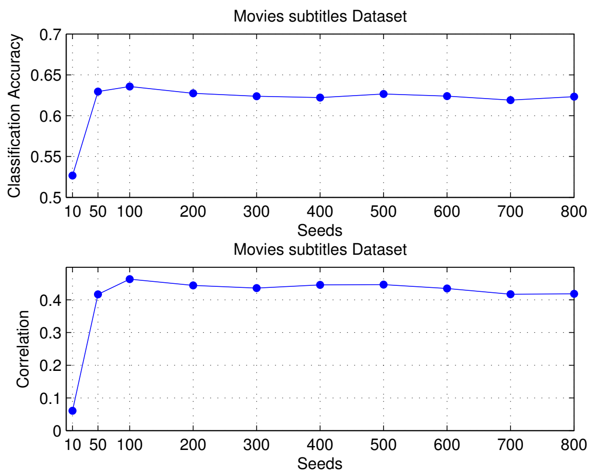


Figure 5.1

The baseline classification accuracy equals to 52.7% and it is yielded by assigning the test samples to the most populous class. Regarding the classification task, we observe that our fusion scheme achieves higher accuracy (64% approximately) compared to the baseline, which is shown to be robust as the number of seeds increases. A moderate correlation (0.44) is obtained for at least 100 seeds, which drops slightly for greater number of seed words.

5.4 Twitter

Nowadays, the usage of social networks such as twitter dominates the daily communication of hundreds of millions of people around the world. People often share opinions and express their feelings about various topics through social networks. Tasks such as opinion mining and sentiment analysis [145] have become very popular since they can capture a large portion of the public opinion. Sentiment analysis of tweets is very relevant across various domains such as commerce [76], disaster management [200] or health [36]. Sentiment analysis of tweets is especially challenging task due to the terse and informal writing style, the semantic diverse of content, as well as the often "unconventional" grammar and orthography. Many computational systems like those submitted to SemEval 2015 task 10 [128, 169], incorpotate bag-of-words models with twitter specific features like hashtags and emoticons [26, 45]. Word embeddings, obtained from large amounts of tweets are used under the scope of an unsupervised approach for sentiment analysis [11]. Additionally, deep learning models have recently become very popular for twitter sentiment analysis [180]. Topic modeling approaches for sentiment analysis can also be found in literature, e.g. [8, 93, 100, 112, 160]

5.4.1 Sentiment Analysis in twitter

Sentiment analysis in twitter is a very popular task that has attracted the interest of many teams all over the world [127,128,169,170]. We developed a system for participating in [128][5] called tweester [142]. This system combines in a probabilistic way various systems like neural networks, semantic-affective models and topic modelling approaches. Each model is used to predict a tweet's sentiment (positive, negative or neutral) and a late fusion scheme is adopted for the final decision. The developed system participated in two subtasks of [128], namely "tweet polarity classification" and "tweet polarity classification given a topic" and it won the latter. In the next section we describe the baseline system that is based on the semantic - affective model while the description of the rest of the systems and their fusion can be found in [142]

Semantic - affective based system

Semantic - affective based system is an enhancement of the system submitted by [103] to [170]. The major changes include the different manipulation of hasgatags, muliword expressions and the affective features as well as the incorporation of new features. A plethora of features is extracted, the majority of which are lexicon-based. Feature extraction occurs in tweet, suffix and prefix level. Assume the following tweet: *"Lol Red Sox just slid through 3rd base #out"* (tweet level). A window applied at the beginning estimates the prefix e.g. *"Lol Red Sox"* (prefix level) and a window at the end of the tweet estimates the suffix *"3rd base #out"* (suffix level). A summary of the baseline system is shown in the Fig. 5.2. As shown in the figure, the analysis of the tweet starts with hashtag preprocessing, then POS tagging and finally feature extraction for the tweet, suffix and prefix level.

Based on the assumption that hashtags in different positions in a tweet may have different semantic interpretation, the tweets are transformed as follows: if a hashtag occurs

[5]This work was a collaboration with the authors of [142]. Athanasia Kolovou developed the CNN system, Fenia Christopoulou the topic modeling based system, Filippos Kokkinos developed the fusion of the different models and the rest members had advising roles.

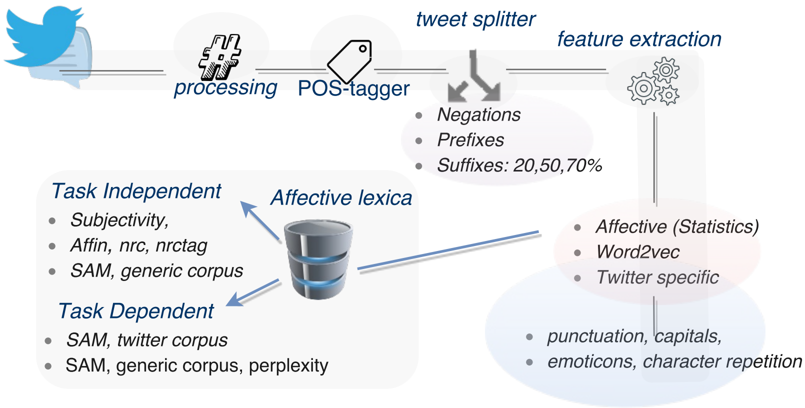


Figure 5.2: Semantic - affective based system overview

at the end of the tweet it is assumed to convey semantic iformation. Otherwise the hashtag is treated as a word or possibly a union of words. In the later case only the corresponding word is kept (e.g. “#moon is so beautiful tonight” → “moon is so beautiful tonight” but “What a beautiful night under the moonlight #romantic” stays as is). A hashtag that contains a union of words has to be expanded, i.e., hashtag has to be expanded into its constituent words (e.g., #Hockeyisback → Hockey is back). Hashtags were expanded using Viterbi Algorithm [55] with n-gram datasets that consist of n-grams and their frequency. The frequency and the relative frequency of hashtags to be expanded were also used as features, as well as the indicators (binary features) that a tweet contains hashtags that require expansion. POS-tagging / Tokenization was then performed using the ARK NLP tweeter tagger [137], a Twitter-specific tagger. A tweet contain single tokens like words or emoticons and punctuations or multiword expressions. Multiword expressions (MWE) are non-compositional expressions that are processed as a single token. They were detected using the Gensim library [164] and they were added to the affective lexica.

Some parts of the tweets may be crucial for the correct understanding of its affective meaning. We assume that such parts, are prefix, the suffix and the negated parts. Negations were detected using the list proposed by Christopher Potts’ tutorial. When a negation token is detected, the tokens that follow are marked as negated until a punctuation mark is reached. Then, feature extraction is applied on the negated part of the tweet. Windows are used in order to split a tweet into prefix and suffix, an approach to estimate the cognitive dissonance phenomenon that is associated with sarcasm, irony and humor [166]. The suffix is estimated by applying windows that keep the 20%, 50% and 70% of tokens that occur at the end of the tweet. Then, feature extraction is repeated for the tweets’ suffixes and prefixes.

Majority of the features are affective features that derive from the semantic-affective model proposed by [107]. The semantic-affective model (SAM) of 3.1 was employed for the estimation of the affective ratings of tweet’s tokens. The dimensionality of the affective features was reduced by keeping only the polarity features (instead of using additional affective dimensions like arousal and dominance). As shown in Fig. 5.2, the affective lexica can be task independent or task dependent. In the former case we have third party lexica, i.e., affin [133] contains discrete polarity ratings in the range $[-5, 5]$, nrc, nrctag [119] contain continuous polarity ratings for tokens, generated from a collection of tweets that had a positive or a negative word hashtag, and the affective lexicon created using a general-purpose corpus of 116M sentences. Task dependent affective lexica were created though SAM using a twitter corpus of 115M tweets, or adapting the general corpus to twitter building a language model on twitter data and using perplexity filtering to keep the more relevant sentences. The affective features extracted were the statistics of the tokens grouped be the POS tag, i.e., length (cardinality), min, max, max amplitude, sum, average, range (max minus min), standard deviation and variance.

In addition to using semantic similarity features we use semantic representation features, i.e., word embeddings that are utilized for the semantic similarity estimation, i.e., the $S(\cdot)$ that appears in 3.1. Word embedding features were derived using word2vec [115], representing each word as a 300-D vector. Since tweet-level features are required, for each tweet a 300-D vector is generated by averaging the corresponding vectors of the constituent words [6].

Character features include frequencies and relative frequencies of selected characters. The selected characters are the capitalized letters, the punctuation marks, the emoticons as well as repetition characters, i.e., at least three same successive characters in a word. Subjectivity features were also extracted based on the subjectivity lexicon of [203]. Specifically, the frequency and the relative frequency of the strong positive/negative and weak positive/negative words were used as features.

A two-stage feature selection was applied, the first one took place on each feature set separately and the second to the combined feature set of the first stage. The final features are used for training a Naive Bayes tree classifier. Both feature selection [7] and the classification have been implemented using Weka [64] and trained on twitter data provided by [128].

The system classifies each tweet either to three classes (positive, negative, neutral) or to two classes (positive, negative). We measure the performance of the system by computing macroaverage recall, i.e., the average recall for the positive and the negative class as shown below:

$$\rho^{PN} = \frac{1}{2}(\rho^{P} + \rho^{N}) = \frac{1}{2}(\frac{PP}{PP + NP} + \frac{NN}{NN + PN}), \tag{5.6}$$

where P, N is the positive and the negative class respectively, ρ indicates the recall, PP the number of correctly predicted positive tweets, NP the number of positive tweets that predicted as negative, NN the number of correctly predicted negative tweets, PN the number of negative tweets that predicted as positive. In the table below we show the macroaverage recall for a twitter dataset that consists of 10K tweets. Tweester is the system developed for the sentiment analysis task, that consists of seven subsystems. Semantic - Affective based (SAMb) is the system we described and CNN [8] is one of tweester's the subsystems.

[6]These features were extracted by Athanasia Kolovou
[7]Best First algorithm was used for feature selection.
[8]CNN was developed by Athanasia Kolovou

System	ρ^{PN}
SAMb	0.821
CNN	0.752
SAMb + CNN	0.827
tweester	0.797

Table 5.3: Macroaverage recall of 10K tweets using the semantic affective based model and combining it with other systems.

Semantic affective based system proved to be the most robust system achieving the highest performance among the others and higher performance than the submitted system, indicating that the semantic affective model is robust and it can be applied successfully for the sentence level affective labelling.

5.5 Spoken Dialogues

During the last decades human - machine communication is used for many daily life application. Dialogues between humans and machines are short, since the dialogue is usually driven by the machine, which urges humans to answer to specific questions. As part of this work we proceed to the affective analysis of Greek and English call center dialogues. Specifically, we aim to classify the utterances of the English dialogues into positive and negative and the utterances of the Greek dialogues into angry and not angry. We focus our efforts on the detection of discrete emotions like anger since for real life application it is important to being able to understand when the interlocutor has been upset, or frustrated.

One usual issue of real call center data is that they are imbalanced as regards their emotional classes and specifically the neutral or the positive class dominates. We used the transcriptions of 1362 English audio calls [9] (342 negative and 1020 positive). One of the problems of this dataset is that almost the half of the sentences are single words (644) and especially “yes” or “no”. The main problem in real call center data is that humans can tell the same lexical information but with totally different way. Say we have utterances with the lexical information “no”. Given that the annotations are based on the speech, the utterances may have both “negative” and “non-negative” annotation labels, but the affective models for text will always classify utterances with the same lexical information always to the same class.

In order to predict the emotional class of spoken dialogue utterances accurately enough, we need to have dialogues with more rich lexical information. So we use a Greek dataset that consists of 200 call center human-machine dialogues for movie ticketing information [99]. The audio recordings and the corresponding transcriptions were available for each dialogue. We used the transcriptions of this dataset for anger detection (“angry” vs. “not angry”) on utterance level. The anger annotations were used both for training and evaluation purposes. The anger detection text system was developed adopting a leave-one-dialogue-out scheme, and building a Random Forest classifier with statistics of words’ dominance ratings as features. The text system performance was evaluated the measuring unweighed average recall and the classification accuracy and it was also fused with speech

[9] Each call contains a user utterance.

system as described in [99]. Results are shown in table 5.4 and as we observe the text based system outperforms the chance system but the performance increases more when it is combined with the speech based system.

	UAR	Accuracy (%)
Chance	0.5	59.5
Text	0.61	59
Fusion with speech	0.67	68

Table 5.4: Performance of anger detection systems.

5.6 Summary

In this chapter we presented a description of the various text datasets that can be used for affective analysis on sentence level. This is a challenging task and its success is highly correlated with the nature of the data. For example detecting emotion from human-machine dialogue is a difficult and challenging task. There are cases that lexical information is not sufficient for emotion detection and speech modality should be incorporated. The next most difficult task is the emotion detection from movie subtitles and this is due to the reason why the lexical information of the movie subtitles co-exist with other modalities (audio and visual) that are lost in the transcriptions. On the other hand, sentences like news headlines and twitter that are written in order to be read by other people seem to be more appropriate for the affective analysis of text.

Chapter 6

Conclusions and Future Work

"*All models are wrong, but some are useful.*"
George E. P. Box

"*Prediction is very difficult, especially about the future.*"
Niels Bohr

6.1 Conclusions

At this work we proposed affective models for emotion detection of various granularity lexical units, we applied our affective models to multiple languages and we experimented with cross-language models that use information of one language in order to predict the affective ratings of words in another language. Within this work we created the first Greek affective lexicon, that contains manually annotated affective ratings for 1034 words. We applied the affective models on multiple continuous affective dimensions and we focused on detecting anger from real call center human-machine dialogues. We proposed an affective model that assumes that compositionality occurs on the affective space instead of the semantic space. Finally, we experimented with numerous written text datasets including news headlines, subtitles, tweets and transcriptions.

For the task of affective labelling of words we employed the semantic affective model of [107]. This model estimates the affective ratings of unknown words computing the linear combination of the affective ratings of N seed words and the semantic similarity between the seeds and the unknown words weighted by weights that capture the importance of the seed words. This model requires an affective lexicon for bootstrapping the process by selecting the seeds. We enhanced this model investigating various parameters for the context-based computation of semantic similarity and we observed that the character 4-grams (extracted from the contextual words) are salient features for the affective labelling of words. We showed that the performance of the affective model depends on the weights estimation method, especially for a large number of seeds. Specifically, the employment of a more robust approach for the estimation of weights, such as ridge regression, leads to noticeable performance improvements. We showed that this model is language independent by applying it to five languages, namely, English, German, Greek, Portuguese and Spanish and achieving consistent performance. Since no affective lexicon was provided for Greek language, we created the first Greek affective lexicon translating the words of English affective lexicon [24] and sharing them to 105 native Greek speakers in order to collect manually annotated affective ratings. Moreover we showed that cross-language affective models can achieve good performance too. The model was applied on three different affective dimensions, i.e., valence, arousal and dominance.

Next, we focused our efforts to the affective labelling of higher granularity level and specifically of word pairs. For this task we followed a compositional approach and specifically we claimed that compositionality occurs in the affective instead of the semantic space. The compositional model is applicable on any affective dimension and affective space, while we present the performance of 3D and 1D models. We restrict the word pairs keeping only AN and NN pairs that we further distinguish into modifier-head structures. The compositional assumption is that the modifier acts upon the head and alternates its affective content, and thus we propose a distributional approach in order to learn modifier's behavior. We show how the compositional models were compared and fused with no compositional models in order to increase performance. Fusion results coincide with the motivation of the fusion suggesting that there is not a single model that works for all word pairs, i.e., the degree and type of compositionality is different for each word pair.

Finally, we conducted a set of experiments for lexical units larger than words and phrases. Specifically, we worked on sentence level for various datasets like news headlines, subtitles, twitter posts and transcriptions from real call center dialogues. We showed that the difficulty of detecting emotion in sentences depends mainly on type of the dataset. For example detecting emotions from transcriptions is a hard problem due to the loss of speech information. The task becomes ever harder in the case that the transcriptions are very short phrases. Thus, the affective model we propose may fail in cases that the lexical information may contain neutral, non emotional words but the speech information is highly emotional. On the other hand, sentences like news headlines and twitter that consist of lexical information seem to be more appropriate for the affective analysis of text.

Overall, we achieved comparable to state-of-the-art performance for the task of emotion detection from text. The enhanced affective model for detecting emotion of words improved the performance of the previous model and it was also applied to more languages. As regards the sentiment analysis of tweets, the performance of the models we described achieved state-of-the art results. However, the main conclusion of this work is that, possibly, composition can take place in the affective space. Compositional models in the literature take place in the semantic space and within this work we showed that affective compositional models can achieve comparable to the state-of-the-art performance.

6.2 Future Work

Further extensions of this work include an even more depth analysis of the semantic-affective model, i.e., to investigate and find out the reasons behind the minor performance changes to the different languages. The most possible reason for these differences are due to the morphology of each language, so we are going to investigate the role of morphology in the computation of semantic models. Also, we aim to verify the universality of our findings by experimenting with more languages. The ideal plan would be to cover multiple languages, e.g., both morphological rich and poor or under-resourced languages. The model could also be extended to more affective dimensions given that affective ratings are available. Regarding the cross language models, there is plenty of work in order to investigate the most efficient source-target language pairs.

Regarding the compositional aspects of emotions, we set the basis for a plethora of applications, since we showed that composition may also occur at the affective space. Before applying affective compositional models to larger lexical units, we should improve the models for the word pairs. Specifically, we should identify parameters that control the degree of compositionality, and learn automatically the parameters used on the fusion scheme. A further extension of the model includes the ambiguous interaction between the modifier and the head, i.e., to assume that the head also alternates the affective content of the modifier. Moreover, we believe that incorporating compositional semantics in the semantic-affective model that is used in the training of the affective compositional model will boost performance, so our future plans will be on this direction. Then, we are going to experiment with this model in various domains and languages as well as with the adaptation of the distributional training process, i.e., we claim that finding ways to adapt the training process to the pairs of interest is going to increase performance. Further extension of the affective compositional model would be its generalization, aiming to the affective labelling of more complex lexical structures. In more complex lexical structures we should first detect the compositional structures and then fuse them with the appropriate way. Our long-term goal is to formulate a generic framework for integrating the compositional and non-compositional aspects of semantic and affective spaces bringing together theories from the areas of cognitive science psycholinguistics and data-driven computational models.

Regarding the affective analysis of text in sentence granularity level we are going to focus our efforts on the sentiment analysis in social media, starting from Twitter. The proposed models can be applied to a variety of tasks including personality detection. Investigating other than emotion interaction dimensions like humour, irony and personality would be a an interesting future task.

Bibliography

[1] Crowdflower crowdsourcing service. `http://crowdflower.com/`.

[2] Ekman's basic emotions. `http://education-portal.com/academy/lesson/ekmans-six-basic-emotions-list-definitions-quiz.html#lesson`.

[3] Abelson, R. Whatever became of consistency theory? *Personality and Social Psychology Bulletin* (1983).

[4] Agrawal, A., and An, A. Unsupervised emotion detection from text using semantic and syntactic relations. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE*, vol. 1, pp. 346–353.

[5] Ahkter, J. K., and Soria, S. Sentiment analysis: Facebook status messages. *Unpublished master's thesis, Stanford, CA* (2010).

[6] Ahmed, S., Jaidka, K., and Skoric, M. Tweets and votes: A four-country comparison of volumetric and sentiment analysis approaches. In *Proc. of 10th International AAAI Conference on Web and Social Media* (2016).

[7] Al Masum, S., Prendinger, H., and Ishizuka, M. Emotion sensitive news agent: An approach towards user centric emotion sensing from the news. In *Proc. of Web Intelligence, IEEE/WIC/ACM* (2007), pp. 614–620.

[8] Alam, M. H., Ryu, W., and Lee, S. Joint Multi-grain Topic Sentiment: Modeling Semantic Aspects for Online Reviews. *Information Sciences 339* (2016), 206–223.

[9] Alm, E. C. O. *Affect in text and speech.* ProQuest, 2008.

[10] Androutsopoulos, I., and Malakasiotis, P. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research* (2010), 135–187.

[11] Astudillo, R., Amir, S., Ling, W., Martins, B., Silva, M. J., and Trancoso, I. INESC-ID: Sentiment Analysis without Hand-Coded Features or Linguistic Resources using Embedding Subspaces. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (2015), Association for Computational Linguistics, pp. 652–656.

[12] Attardo, S. Humor and irony in interaction: From mode adoption to failure of detection. *EMERGING COMMUNICATION 3* (2002), 159–180.

[13] Balahur, A., and Turchi, M. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language 28*, 1 (2014), 56–75.

[14] Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. Multilingual subjectivity analysis using machine translation. In *Proc. of Empirical Methods in Natural Language Processing* (2008), pp. 127–135.

[15] Baroni, M., and Zamparelli., R. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP)* (2010).

[16] Barrett, L. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review 10*, 1 (2006), 20–46.

[17] Bechara, A., Damasio, H., and Damasio, A. Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex 10*, 3 (2000), 295–307.

[18] Becker-Asano, C., and Wachsmuth, I. Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems 20*, 1 (2010), 32–49.

[19] Bobicev, V., Maxim, V., Prodan, T., Burciu, N., and Angheluş, V. Emotions in words: Developing a multilingual WordNet-Affect. In *Computational Linguistics and Intelligent Text Processing* (2010), pp. 375–384.

[20] Boiy, E., and Moens, M.-F. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval 12*, 5 (2009), 526–558.

[21] Boldrini, E., Balahur, A., Martínez-Barco, P., and Montoyo, A. Using emotiblog to annotate and analyse subjectivity in the new textual genres. *Data Mining and Knowledge Discovery 25*, 3 (2012), 603–634.

[22] Boyd-Graber, J., and Resnik, P. Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation. In *Proc. of Empirical Methods in Natural Language Processing* (2010), pp. 45–55.

[23] Bradley, M., and Lang, P. Measuring emotion: The self-assessment manikin and the semantic differential. *Behavior Therapy and Experimental Psychiatry 25*, 1 (1994), 49–59.

[24] Bradley, M., and Lang, P. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings, technical report c-1. Tech. rep., The Center for Research in Psychophysiology, University of Florida, 1999.

[25] Bradley, M., and Lang, P. The international affective digitized sounds (IADS)[: stimuli, instruction manual and affective ratings]. Tech. rep., NIMH Center for the Study of Emotion and Attention, 1999.

[26] Büchner, M., and Stein, B. Webis: An Ensemble for Twitter Sentiment Detection. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (2015), Association for Computational Linguistics, pp. 582–589.

[27] Bullinaria, J. A., and Levy, J. P. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods 39*, 3 (2007), 510–526.

[28] Cambria, E. Affective computing and sentiment analysis. *IEEE Intelligent Systems 31*, 2 (2016), 102–107.

[29] Cambria, E., Schuller, B., Xia, Y., and Havasi, C. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 2 (2013), 15–21.

[30] Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., and Karpouzis, K. Multimodal emotion recognition from expressive faces, body gestures and speech. In *Artificial intelligence and innovations 2007: From theory to applications*. Springer, 2007, pp. 375–388.

[31] Celli, F. *Adaptive Personality Recogntion from Text*. PhD thesis, University of Trento, 2012.

[32] Celli, F. Unsupervised personality recognition for social network sites. In *Proc. of International Conference on Digital Society (ICDS)* (2012), pp. 59–62.

[33] Chaffar, S., and Inkpen, D. Using a heterogeneous dataset for emotion analysis in text. In *Advances in Artificial Intelligence*. 2011, pp. 62–67.

[34] Chaumartin, F. Upar7: A knowledge-based system for headline sentiment tagging. In *Proc. of the 4th International Workshop on Semantic Evaluations* (2007), pp. 422–425.

[35] Chenlo, J. M., Hogenboom, A., and Losada, D. E. Sentiment-based ranking of blog posts using rhetorical structure theory. In *International Conference on Application of Natural Language to Information Systems* (2013), Springer, pp. 13–24.

[36] Chew, C., and Eysenbach, G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS ONE 5*, 11 (2010), 1–13.

[37] Choi, Y., and Cardie, C. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proc. of Empirical Methods in Natural Language Processing* (2008), pp. 793–801.

[38] Church, K. W., and Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics 16*, 1 (1990), 22–29.

[39] Cruz, F. L., Troyano, J. A., Ortega, F. J., and Enríquez, F. Automatic expansion of feature-level opinion lexicons. In *Proc. of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (2011), pp. 125–131.

[40] Cui, A., Zhang, M., Liu, Y., and Ma, S. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Information retrieval technology*. 2011, pp. 238–249.

[41] Dai, W., Han, D., Dai, Y., and Xu, D. Emotion recognition and affective computing on vocal social media. *Information & Management 52*, 7 (2015), 777–788.

[42] Damasio, A. Descartes' error: Emotion, reason and the human mind. *Putnam, New York* (1994), 195–201.

[43] Damasio, A. *Descartes' error: Emotion, reason, and the human brain.* Penguin. com, 2005.

[44] Darwin, C., Ekman, P., and Prodger, P. *The expression of the emotions in man and animals.* Oxford University Press, USA, 1998.

[45] Davidov, D., Tsur, O., and Rappoport, A. Enhanced sentiment learning using twitter hashtags and smileys. In *Proc. of Coling 2010:* (2010), pp. 241–249.

[46] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American society for information science 41*, 6 (1990), 391.

[47] Denecke, K. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on* (2008), IEEE, pp. 507–512.

[48] Descartes, R. *The passions of the soul.* Hackett Publishing, 1989.

[49] Devillers, L., Lamel, L., and Vasilescu, I. Emotion detection in task-oriented spoken dialogues. In *Proc. of ICME* (2003), vol. 3, IEEE, pp. III–549.

[50] Dhall, A., Goecke, R., Joshi, J., Wagner, M., and Gedeon, T. Emotion recognition in the wild challenge 2013. In *Proc. of the 15th ACM on International conference on multimodal interaction* (2013), pp. 509–516.

[51] Eilola, T. M., and Havelka, J. Affective norms for 210 British English and Finnish nouns. *Behavior Research Methods 42*, 1 (2010), 134–140.

[52] Elghazaly, T., Mahmoud, A., and Hefny, H. Political sentiment analysis using twitter data. In *Proc. ofthe International Conference on Internet of things and Cloud Computing* (2016), p. 11.

[53] Esuli, A., and Sebastiani, F. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proc. of LREC* (2006), pp. 417–422.

[54] Forbes-Riley, K., and Litman, D. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proc. of HLT-NAACL* (2004), pp. 201–208.

[55] Forney, J., and David, G. The Viterbi algorithm. *IEEE 61*, 3 (1973), 268–278.

[56] Frijda, N. H. The emotions: Studies in emotion and social interaction. *Paris: Maison de Sciences de l'Homme* (1986).

[57] Georgiladakis, S., Athanasopoulou, G., Meena, R., Lopes, J., Chorianopoulou, A., Palogiannidi, E., Iosif, E., G., S., and Potamianos, A. Root cause analysis of miscommunication hotspots in spoken dialogue systems. In *Proc. of Interspeech* (2016 (*to appear*)).

[58] Georgiladakis, S., E., I., and A., P. Fusion of compositional network-based and lexical function distributional semantic models. In *Proc. of CMCL* (2015), pp. 39–47.

[59] Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., and Reyes, A. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval)* (2015), pp. 470–478.

[60] Gilet, A., Grühn, D., Studer, J., and Labouvie-Vief, G. Valence, arousal, and imagery ratings for 835 French attributes by young, middle-aged, and older adults: The French Emotional Evaluation List (FEEL). *European Review of Applied Psychology 62*, 3 (2012), 173–181.

[61] Godbole, N., Srinivasaiah, M., and Skiena, S. Large-scale sentiment analysis for news and blogs. *ICWSM 7*, 21 (2007), 219–222.

[62] Goleman, D. *Emotional intelligence*. Bantam, 2006.

[63] Gracia, J., Trillo, R., Espinoza, M., and Mena, E. Querying the web: A multiontology disambiguation method. In *Proc. of International Conference on Web Engineering* (2006), pp. 241–248.

[64] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*, 1 (2009), 10–18.

[65] Harris, Z. Distributional structure. *Word 10*, 23 (1954), 146–162.

[66] Hasbullah, S. S., Maynard, D., Chik, R., Mohd, F., and Noor, M. Automated content analysis: A sentiment analysis on malaysian government social media. In *Proc. of the 10th International Conference on Ubiquitous Information Management and Communication* (2016).

[67] Hashimoto, K., Miwa, M., Tsuruoka, Y., and Chikayama, T. Simple customization of recursive neural networks for semantic relation classification. In *EMNLP* (2013), pp. 1372–1376.

[68] Hatfield, E., Cacioppo, J., and Rapson, R. *Emotional contagion*. 1994.

[69] Hatzivassiloglou, V.and McKeown, K. Predicting the semantic orientation of adjectives. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics* (1997), pp. 174–181.

[70] Hu, M., and Liu, B. Mining and summarizing customer reviews. In *Proc. of the 10^{th} international conference on Knowledge discovery and data mining, SIGKDD* (2004), ACM, pp. 168–177.

[71] Hu, X., Tang, L., Tang, J., and Liu, H. Exploiting social relations for sentiment analysis in microblogging. In *Proc. of the sixth ACM international conference on Web search and data mining* (2013), pp. 537–546.

[72] Huron, D. *Sweet anticipation: Music and the psychology of expectation*. The MIT Press, 2006.

[73] Iosif, E., Georgiladakis, S., and Potamianos, A. Cognitively Motivated Distributional Representations of Meaning. In *Proc. of LREC* (2016).

[74] Iosif, E., and Potamianos, A. Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering 20*, 11 (2010), 1637–1647.

[75] Iosif, E., and Potamianos, A. Similarity computation using semantic networks created from web-harvested data. *Natural Language Engineering 21*, 01 (2015), 49–79.

[76] Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology 60*, 11 (2009), 2169–2188.

[77] Jeff, M., and Lapata., M. Vector-based models of semantic composition. In *Proc. of (ACL)* (2008), pp. 236–244.

[78] Jia, Y., Chen, Z., and Yu, S. Reader emotion classification of news headlines. In *Proc. of Natural Language Processing and Knowledge Engineering (NLP-KE)* (2009), IEEE, pp. 1–6.

[79] Jiang, J. J., and Conrath, D. W. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008* (1997).

[80] Kaji, N., and Kitsuregawa, M. Building lexicon for sentiment analysis from massive collection of html documents. In *Proc. of EMNLP-CoNLL* (2007), pp. 1075–1083.

[81] Kalyan, C., and Kim, M. Detecting emotional scenes using semantic analysis on subtitles, 2009.

[82] Kanayama, H., and Nasukawa, T. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing* (2006), pp. 355–363.

[83] Kanske, P., and Kotz, S. Leipzig affective norms for German: A reliability study. *Behavior research methods 42*, 4 (2010), 987–991.

[84] Keltner, D., and Ekman, P. Introduction to expression of emotion. In *R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Eds.), Handbook of Affective Sciences* (2003), Oxford University Press, pp. 411–414.

[85] Kirange, D., and R, D. R. Emotion classification of news headlines using svm. *Asian Journal of Computer Science & Information Technology 2*, 5 (2013).

[86] Koutras, P., Zlatintsi, A., Iosif, E., Katsamanis, A., Maragos, P., and Potamianos, A. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *Proc. of Image Processing (ICIP), 2015 IEEE* (2015), pp. 4361–4365.

[87] Kozareva, Z. Multilingual affect polarity and valence prediction in metaphor-rich texts. In *Proc. of Association for Computational Linguistics* (2013), pp. 682–691.

[88] Landauer, T. K., and Littman, M. L. Fully automatic cross-language document retrieval using latent semantic indexing.

[89] Lang, P., Bradley, M., and Cuthbert, B. International affective picture system (IAPS): Technical manual and affective ratings. Tech. rep., Gainesville, FL: The Center for Research in Psychophysiology, University of Florida, 1999.

[90] Leacock, C., and Chodorow, M. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database 49*, 2 (1998), 265–283.

[91] Li, N., and Wu, D. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems 48*, 2 (2010), 354–368.

[92] Li, Y., Bandar, Z. A., and McLean, D. An approach for measuring semantic similarity between words using multiple information sources. *Knowledge and Data Engineering, IEEE Transactions on 15*, 4 (2003), 871–882.

[93] Lin, C., and He, Y. Joint sentiment/topic model for sentiment analysis. In *Proc. of the 18th Conference on Information and Knowledge Management (CIKM)* (2009), Association for Computing Machinery, pp. 375–384.

[94] Lindquist, K. A., and Barrett, L. Constructing emotion the experience of fear as a conceptual act. *Psychological science 19*, 9 (2008), 898–903.

[95] Liscombe, J., Riccardi, G., and Hakkani-Tur, D. Using context to improve emotion detection in spoken dialog systems.

[96] Litman, D. J., and Forbes-Riley, K. Predicting student emotions in computer-human tutoring dialogues. In *Proc. of the 42nd Annual Meeting on Association for Computational Linguistics* (2004), p. 351.

[97] Liu, B. Sentiment analysis and subjectivity. *Handbook of natural language processing 2* (2010), 627–666.

[98] Liu, H., Lieberman, H., and Selker, T. A model of textual affect sensing using real-world knowledge. In *Proc. of the 8th international conference on Intelligent user interfaces* (2003), ACM, pp. 125–132.

[99] Lopes, J., Chorianopoulou, A., Palogiannidi, E., Moniz, H., Abad, A., Louka, K., Iosif, E., and Potamianos, A. The spedial datasets: Datasets for spoken dialogue systems analytics. In *Proc. of LREC* (2016).

[100] Lu, B., Ott, M., Cardie, C., and Tsou, B. K. Multi-aspect sentiment analysis with topic models. In *Proc. of International Conference on Data Mining Workshops (ICDMW)* (2011), IEEE, pp. 81–88.

[101] Lund, K., and Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers 28*, 2 (1996), 203–208.

[102] Mairesse, F., Walker, M., Mehl, M., and Moore, R. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.(JAIR) 30* (2007), 457–500.

[103] Malandrakis, N., Kazemzadeh, A., Potamianos, A., and Narayanan, S. S. SAIL: A hybrid approach to sentiment analysis. In *Proc. of 7th International Workshop on Semantic Evaluation SemEval* (2013), pp. 438–442.

[104] Malandrakis, N., Potamianos, A., Evangelopoulos, G., and Zlatintsi, A. A supervised approach to movie emotion tracking. In *Proc. of international conference on acoustics, speech and signal processing (ICASSP)* (2011), IEEE, pp. 2376–2379.

[105] Malandrakis, N., Potamianos, A., Hsu, K. J., Babeva, K. N., Feng, M. C., Davison, G. C., and Narayanan, S. Affective language model adaptation via corpus selection. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (2014), IEEE, pp. 4838–4842.

[106] Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. Kernel models for affective lexicon creation. In *INTERSPEECH* (2011), pp. 2977–2980.

[107] Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. S. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing 21*, 11 (2013), 2379–2392.

[108] Marquardt, D., and Snee, R. Ridge regression in practice. *The American Statistician 29*, 1 (1975), 3–20.

[109] McGaugh, J. The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annu. Rev. Neurosci. 27* (2004), 1–28.

[110] Medhat, W., Hassan, A., and Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal 5*, 4 (2014), 1093–1113.

[111] Meena, A., and Prabhakar, T. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *Advances in Information Retrieval.* Springer, 2007, pp. 573–580.

[112] Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proc. of the 16th International Conference on World Wide Web (ICWWW)* (2007), Association for Computing Machinery, pp. 171–180.

[113] Meng, M., and Bader, M. Ungrammaticality detection and garden path strength: Evidence for serial parsing. *Language and Cognitive processes 15*, 6 (2000), 615–666.

[114] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[115] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Proc of. Advances in Neural Information Processing systems (NIPS)* (2013), pp. 3111–3119.

[116] Miller, G. WordNet: a lexical database for english. *Communications of the ACM 38*, 11 (1995), 39–41.

[117] Mitchell, J., and Lapata, M. Composition in distributional models of semantics. *Cognitive Science 34*, 8 (2010), 1388–1429.

[118] Mnih, A., and Kavukcuoglu, K. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems* (2013), pp. 2265–2273.

[119] Mohammad, S. M., Kiritchenko, S., and Zhu, X. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval)* (2013), pp. 321–327.

[120] Moilanen, K., and Pulman, S. Sentiment composition. In *Proc. of RANLP* (2007), vol. 7, pp. 378–382.

[121] Moilanen, K., Pulman, S., and Zhang, Y. Packed feelings and ordered sentiments: Sentiment parsing with quasi-compositional polarity sequencing and compression. In *Proc. of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (2010), pp. 36–43.

[122] Montefinese, M., Ambrosini, E., Fairfield, B., and Mammarella, N. The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior research methods 46*, 3 (2014), 887–903.

[123] Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A., De Schryver, M., De Winne, J., and Brysbaert, M. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior research methods 45*, 1 (2013), 169–177.

[124] Morency, L.-P., Mihalcea, R., and Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proc. of the 13th international conference on multimodal interfaces* (2011), pp. 169–176.

[125] Morrison, D.and Wang, R., and De Silva, L. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication 49*, 2 (2007), 98–112.

[126] Mullen, T., and Malouf, R. A preliminary investigation into sentiment analysis of informal political discourse. In *Proc. of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (2006), pp. 159–162.

[127] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval)* (2013).

[128] Nakov, P., Ritter, A., Rosenthal, S., Stoyanovand, V., and Sebastiani, F. SemEval-2016 task 4: Sentiment Analysis in Twitter. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval)* (2016).

[129] Nasukawa, T., and Yi, J. Sentiment analysis: Capturing favorability using natural language processing. In *Proc. of the 2nd international conference on Knowledge capture* (2003), ACM, pp. 70–77.

[130] Neuman, Y., Kedma, G., Cohen, Y., and Nave, O. Using web-intelligence for excavating the emerging meaning of target-concepts. In *Web Intelligence and Intelligent Agent Technology (WI-IAT)* (2010), vol. 1, IEEE, pp. 22–25.

[131] Neviarouskaya, A., Prendinger, H., and Ishizuka, M. Sentiful: Generating a reliable lexicon for sentiment analysis. In *Affective Computing and Intelligent Interaction and Workshops, 3rd International Conference on* (2009), IEEE, pp. 1–6.

[132] Neviarouskaya, A., Prendinger, H., and Ishizuka, M. Recognition of affect, judgment, and appreciation in text. In *Proc. of the 23rd International Conference on Computational Linguistics* (2010), Association for Computational Linguistics, pp. 806–814.

[133] Nielsen, F. Å. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proc. of the ESWC Workshop on Making Sense of Microposts* (2011), pp. 93–98.

[134] Ortigosa, A., Martín, J., and Carro, R. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior 31* (2014), 527–541.

[135] Osherenko, A. Affect sensing using lexical means: Comparison of a corpus with movie reviews and a corpus with natural language dialogues. In *The Workshop Programme Corpora for Research on Emotion and Affect Tuesday 23 rd May 2006* (2006).

[136] Osherenko, A., and André, E. Lexical affect sensing: Are affect dictionaries necessary to analyze affect? In *Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 230–241.

[137] Owoputi, O., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. of North American Chapter of the Association for Computational Linguistics (NAACL)* (2013).

[138] Palogiannidi, E. Using crowdsourcing for grammar induction with application to spoken dialogue systems. In *School of ECE, Technical University of Crete* (2013).

[139] Palogiannidi, E., Iosif, E., Koutsakis, P., and Potamianos, A. Valence, Arousal and Dominance Estimation for English, German, Greek, Portuguese and Spanish Lexica using Semantic Models. In *Proc. of Interspeech* (2015), pp. 1527–1531.

[140] Palogiannidi, E., Iosif, E., Koutsakis, P., and Potamianos, A. A semanticaffective compositional approach for the affective labelling of adjective-noun and noun-noun pairs. In *of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)* (2016).

[141] Palogiannidi, E., Klasinas, I., Potamianos, A., and Iosif, E. Spoken dialogue grammar induction from crowdsourced data. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014), IEEE, pp. 3211–3215.

[142] Palogiannidi, E., Kolovou, A., Christopoulou, F., Iosif, E., Malandrakis, N., Papageorgiou, H., Narayanan, S., and Potamianos, A. Tweester at SemEval 2016: Sentiment Analysis in Twitter using Semantic-Affective Model Adaptation. In *of the 10th International Workshop on Semantic Evaluation (SemEval)* (2016).

[143] Palogiannidi, E., Koutsakis, P., Iosif, E., and Potamianos, A. Affective lexicon creation for the Greek language. In *Proc. of LREC* (2016).

[144] Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of the 42nd annual meeting on Association for Computational Linguistics* (2004), p. 271.

[145] Pang, B., and Lee, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval 2*, 1-2 (2008), 1–135.

[146] Pantic, M., and Rothkrantz, L. J. Toward an affect-sensitive multimodal human-computer interaction. 1370–1390.

[147] Park, S.-B., Yoo, E., Kim, H., and Jo, G.-S. Automatic emotion annotation of movie dialogue using wordnet. In *Proc. of Intelligent Information and Database Systems*. 2011, pp. 130–139.

[148] Partee, B. Lexical semantics and compositionality. *An invitation to cognitive science: Language 1* (1995), 311–360.

[149] Pavlopoulos, I. Aspect based sentiment analysis. *Athens University of Economics and Business* (2014).

[150] Pelletier, F. J. The principle of semantic compositionality. *Journal of Topoi 13*, 1 (1994), 11–24.

[151] Pennebaker, J. W. The secret life of pronouns. *New Scientist 211*, 2828 (2011), 42–45.

[152] Pérez-Rosas, V., Mihalcea, R., and Morency, L.-P. Utterance-level multimodal sentiment analysis. In *Proc. of ACL* (2013), pp. 973–982.

[153] Petrakis, E. G., Varelas, G., Hliaoutakis, A., and Raftopoulou, P. X-similarity: computing semantic similarity between concepts from different ontologies. *JDIM 4*, 4 (2006), 233–237.

[154] Picard, R. *Affective computing*. MIT press, 2000.

[155] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval)* (2015), pp. 486–495.

[156] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval)* (2014), pp. 27–35.

[157] Poria, S., Cambria, ., Howard, N., Huang, G.-B., and Hussain, A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing 174* (2016), 50–59.

[158] Poria, S., Cambria, E., and Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proc. of EMNLP* (2015), pp. 2539–2544.

[159] Potamianos, A. Cognitive multimodal processing: from signal to behavior. In *Proc. of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges* (2014), pp. 27–34.

[160] Rao, Y. Contextual Sentiment Topic Model for Adaptive Social Emotion Classification. *IEEE Intelligent Systems 31*, 1 (2016), 41–47.

[161] Rapp, R. Word sense discovery based on sense descriptor dissimilarity. In *Proc. of the ninth machine translation summit* (2003), pp. 315–322.

[162] Ravi, K., and Ravi, V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems 89* (2015), 14–46.

[163] Redondo, J., Fraga, I., Padrón, I., and Comesaña, M. The Spanish adaptation of ANEW (Affective Norms for English Words). *Journal of Behavior Research Methods 39*, 3 (2007), 600–605.

[164] Řehůřek, R., and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the LREC Workshop on New Challenges for NLP Frameworks* (2010), pp. 45–50.

[165] Rentoumi, V., Giannakopoulos, G., Karkaletsis, V., and Vouros, G. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proc. of RANLP* (2009), pp. 370–375.

[166] Reyes, A., Rosso, P., and Buscaldi, D. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering 74* (2012), 1–12.

[167] Reyes, A.and Rosso, P., and Veale, T. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* (2013), 1–30.

[168] Rosas, V., Mihalcea, R., and Morency, L.-P. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 3 (2013), 38–45.

[169] Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., and Stoyanov, V. Semeval-2015 Task 10: Sentiment Analysis in Twitter. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval)* (2015), pp. 451–463.

[170] Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval)* (2014), pp. 73–80.

[171] Russell, J., and Barrett, L. F. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology 76*, 5 (1999), 805.

[172] Russell, J. A. A circumplex model of affect. *Journal of Personality and Social Psycology 39*, 6 (1980), 1161–1178.

[173] Russell, J. A., and Mehrabian, A. Evidence for a three-factor theory of emotions. *Journal of research in Personality 11*, 3 (1977), 273–294.

[174] Scherer, K. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences 364*, 1535 (2009), 3459–3474.

[175] Scherer, K. R. Emotion theories and concepts (psychological perspectives).

[176] Schuller, B., Valster, M., Eyben, F., Cowie, R., and Pantic, M. Avec 2012: the continuous audio/visual emotion challenge. In *Proc. of the 14th ACM international conference on Multimodal interaction* (2012), ACM, pp. 449–456.

[177] Schütze, H. Automatic word sense discrimination. *Computational linguistics 24*, 1 (1998), 97–123.

[178] Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., and Ungar, L. Towards assessing changes in degree of depression through facebook. In *Proc. of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (2014), pp. 118–125.

[179] Schwarz, N. Emotion, cognition, and decision making. *Cognition & Emotion 14*, 4 (2000), 433–440.

[180] Severyn, A., and Moschitti, A. UNITN: Training deep convolutional neural network for Twitter sentiment classification. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (2015), Association for Computational Linguistics, Association for Computational Linguistics, pp. 464–469.

[181] Singhal, K., Agrawal, B., and Mittal, N. Modeling indian general elections: sentiment analysis of political twitter data. In *Information Systems Design and Intelligent Applications*. 2015, pp. 469–477.

[182] Smith, C. A., and Lazarus, R. S. Emotion and adaptation.

[183] Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., and Frade, C. S. The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Journal of Behavior research methods 44*, 1 (2012), 256–269.

[184] Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. Semantic compositionality through recursive matrix-vector spaces. In *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2012), pp. 1201–1211.

[185] Socher, R., Perelygin, A., J.W., Chuang, J., Manning, C., A.N., and Potts, C. Parsing With Compositional Vector Grammars. In *EMNLP* (2013).

[186] Stevenson, R., Mikels, J., and James, T. Characterization of the affective norms for English words by discrete emotional categories. *Behavior Research Methods 39*, 4 (2007), 1020–1024.

[187] Stone, P., Dunphy, D. C., and Smith, M. S. The General Inquirer: A Computer Approach to Content Analysis.

[188] Strapparava, C., and Mihalcea, R. SemEval-2007 Task 14: Affective text. In *Proc. of International Workshop on Semantic Evaluation (SemEval)* (2007), pp. 70–74.

[189] Strapparava, C., and Mihalcea, R. Learning to identify emotions in text. In *Proc. of the 2008 ACM symposium on Applied computing* (2008), pp. 1556–1560.

[190] Strapparava, C., and Valitutti, A. WordNetAffect: an affective extension of WordNet. In *Proc. of LREC* (2004), pp. 1083–1086.

[191] Thet, T. T., Na, J.-C., and Khoo, C. S. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science* (2010).

[192] Tomkins, S. S. *Affect imagery consciousness: Volume I: The positive affects*, vol. 1. 1962.

[193] Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proc. of ICWSM 10* (2010), 178–185.

[194] Turing, A. M. Computing machinery and intelligence. *Mind 59*, 236 (1950), 433–460.

[195] Turney, P. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th annual meeting on association for computational linguistics* (2002), pp. 417–424.

[196] Turney, P., and Littman, M. Unsupervised learning of semantic orientation from a hundred-billion-word corpus.

[197] Turney, P., and Littman, M. L. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS) 21*, 4 (2003), 315–346.

[198] Turney, P. D., and Pantel, P. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research 37*, 1 (2010), 141–188.

[199] Verduyn, P.and Lavrijsen, S. Which emotions last longest and why: The role of event importance and rumination. *Motivation and Emotion 39*, 1 (2015), 119–127.

[200] Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A., and Anderson, K. M. Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In *Proc.*

of the 5th International Conference on Web and Social Media (ICWSM) (2011), American Association for Artificial Intelligence.

[201] Vidrascu, L., and Devillers, L. Detection of real-life emotions in call centers. In *Proc. of INTERSPEECH* (2005), pp. 1841–1844.

[202] Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proc. of ACL System Demonstrations* (2012), pp. 115–120.

[203] Wilson, T., Wiebe, J., and Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)* (2005), pp. 347–354.

[204] Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L.-P. Youtube movie reviews: Sentiment analysis in an audio-visual context. *Intelligent Systems, IEEE 28*, 3 (2013), 46–53.

[205] Wu, Q., and Tan, S. A two-stage framework for cross-domain sentiment classification. *Expert Systems with Applications 38*, 11 (2011), 14269–14275.

[206] Wundt, W. M. *Principles of physiological psychology.* 1969.

[207] Xu, T., Peng, Q., and Cheng, Y. Identifying the semantic orientation of terms using s-hal for sentiment analysis. *Knowledge-Based Systems 35* (2012), 279–289.

[208] Yessenalina, A., and Cardie, C. Compositional matrix-space models for sentiment analysis. In *Proc. of the Conference on Empirical Methods in Natural Language Processing* (2011), pp. 172–182.

[209] Zanzotto, F. M., Korkontzelos, I., Fallucchi, F., and Manandhar, S. Estimating linear models for compositional distributional semantics. In *Proc. of the 23rd International Conference on Computational Linguistics* (2010), pp. 1263–1271.

[210] Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. Relation classification via convolutional deep neural network. In *Proc. of COLING* (2014), pp. 2335–2344.

[211] Zhang, Z.and Li, X., and Chen, Y. Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews. *ACM Transactions on Management Information Systems (TMIS) 3*, 1 (2012), 5.

[212] Zlatintsi, A., Iosif, E., Maragos, P., and Potamianos, A. Audio salient event detection and summarization using audio and text modalities. In *Proc. of Signal Processing Conference (EUSIPCO), 23rd European* (2015), IEEE, pp. 2311–2315.