UNIVERSITÁ DI PISA
DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE

DISTRIBUTIONAL MODELS OF EMOTIONS FOR SENTIMENT ANALYSIS IN SOCIAL MEDIA

DOCTORAL THESIS

Author
Lucia C. Passaro

Tutor (s)
Prof. Alessandro Lenci
Prof. Francesco Marcelloni
Prof. Gigliola Vaglini

Reviewer (s)
Prof. Malvina Nissim
Prof. Carlo Strapparava

The Coordinator of the PhD Program
Prof. Marco Luise

Pisa, January 2017

XXIX Cycle

Acknowledgements

HEARTFELT thanks go first of all to Prof. Alessandro Lenci, who guided me during my PhD and who has been an amazing advisor in so many aspects of this role. I'm also truly thankful to Prof. Francesco Marcelloni and Prof. Gigliola Vaglini for their teachings, for many helpful discussions and feedback, as well as for their valuable advice.

I would like to thank my colleagues in the CoLing Lab, especially Marco Senaldi and Dr. Gianluca Lebani, who shared with me the ups and downs of my doctoral years, and made my experience fruitful and enjoyable. Among them, Gianluca was a fundamental support, always available to share his experience, and to provide me with many invaluable comments and insightful suggestions at each stage of this research.

This thesis is based upon work supported in part by Regione Toscana who funded the project "SEMantic instruments for PubLIc administrators and CitizEns (SEMPLICE)", and from the Company ETI[3]. Many thanks to my colleagues that, with great enthusiasm, made my PhD an opportunity to gather experience in applying research to industrial needs.

I would like to thank Prof. Malvina Nissim and Prof. Carlo Strapparava, for their helpful comments and inputs: these pages undoubtedly benefited from their reviews.

A special acknowledgement goes to my ever-supporting family and to my life partner Marcello, for being always there for me. They ever encouraged and supported me through thick and thin of these years. Without them, I wouldn't be standing here today.

Last but not least, I'm deeply grateful to all the friends who stood by me through these years, sharing with me both the good and the bad. They all have my sincere gratitude.

Summary

WITH the proliferation of social media, textual emotion analysis is becoming increasingly important. Sentiment Analysis and Emotion Detection can be useful to track several applications. They can be used, for instance, in Customer Relationship Management to track sentiments towards companies and their services, or in Government Intelligence, to collect people's emotions and points of views about government decisions. It is clear that tracking reputation and opinions without appropriate text mining tools is simple infeasible. Most of these tools are based on sentiment and emotion lexicons, in which lemmas are associated with the sentiment and/or emotions they evoke. However, almost all languages but English lack high-coverage inventories of this sort.

This thesis presents several sentiment analysis tasks to illustrate challenges and opportunities in this research area. We review different state-of-the-art methods for sentiment analysis and emotion detection and describe how we modeled a framework to build emotive resources, that can be effectively exploited for text affective computing. One of the main outcome of the work presented in this thesis is ItEM, which is a high-coverage Italian EMotive lexicon created by exploiting distributional methods. It has been built with a three stage process including the collection of a set of highly emotive words, their distributional expansion and the validation of the system. Since corpus-based methods reflect the type of the corpus from which they are build, in order to create a reliable lexicon we collected a new Italian corpus, namely FB-NEWS15. This collection has been created by crawling the Facebook pages of the most important Italian newspapers, which typically include a small number of posts written by the journalists and a very high number of comments inspired by long discussions among readers about such news.

Finally, we describe some experiments on the sentiment polarity classification of tweets. We started from a system based on supervised learning that was originally developed for the Evalita 2014 SENTIment POLarity Classification task [14] and subsequently explored the possibility to enrich this system by exploiting lexical emotive features derived from social media texts.

Sommario

CON lo sviluppo dei social media, lo studio delle emozioni nei testi sta diventando sempre più importante e strumenti come Sentiment Analysis ed Emotion Detection si stanno diffondendo sempre di più per monitorare diverse applicazioni. Per esempio, in programmi di Customer Relationship Management possono essere usati per tenere traccia dei sentimenti dei consumatori nei confronti delle aziende e dei loro servizi, oppure nell'ambito della Government Intelligence possono essere cruciali per collezionare e analizzare punti di vista diversi riguardo le decisioni del governo. Monitorare reputazione e opinioni senza gli strumenti di text mining appropriati può risultare un compito molto difficile, se non impossibile. La maggior parte di questi strumenti si servono di lessici, in cui alle parole è attribuito un certo grado di positività, negatività e/o valori emotivi. Putroppo ancora oggi molte lingue non dispongono di questi lessici.

Questa tesi presenta diversi task di Sentiment Analysis e illustra le potenzialità di quest'area di ricerca mostrando diversi metodi allo stato dell'arte, nonché la creazione di un framework per costruire risorse lessicali emotive che possono essere sfruttate in applicazioni di text affective computing. Uno dei primi risultati di questo lavoro è ItEM, un lessico emotivo per l'italiano ad alta copertura, che è stato creato sfruttando metodi distribuzionali. ItEM è stato creato in un processo a tre fasi che includono la raccolta di insiemi di parole fortemente emotive, la loro espansione distribuzionale e la validazione del sistema. Poiché un lessico creato con metodi corpus-based è fortemente dipendente dal corpus dal quale vengono estratti i profili distribuzionali, dopo la validazione del metodo, è stato creato il corpus FB-NEWS15, composto di testi provenienti dalle pagine Facebook dei maggiori quotitiani italiani, che contengono un gran numero di commenti degli utenti che discutono sulle notizie pubblicate.

Infine si mostrano degli esperimenti di sentiment polarity classification sui tweet. Partendo da un sistema supervisionato sviluppato per la partecipazione a Evalita 2014 SENTIment POLarity Classification task [14], è stata esplorata la possibilità di arricchire il sistema con features lessicali emotive derivanti da testi dei social media.

List of Publications

International Conferences/Workshops with Peer Review

[C1] Boschetti, F., Cimino, A., Dell'Orletta, F., Lebani, G. E., **Passaro, L. C.**, Picchi, P., Venturi G., Montemagni S. and Lenci, A. (2014). Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II. *In Proceedings of the LREC 2014 Workshop LRT4HDA*. (pp. 70-76). Reykjavík (Iceland), May 2014.

[C2] **Passaro, L. C.** and Lenci, A. (2014). "Il Piave mormorava…": Recognizing Locations and other Named Entities in Italian Texts on the Great War. *In Proceedings of CLiC-it 2014.* (pp. 286–290). Pisa (Italy), December 2014.

[C3] **Passaro, L. C.** and Lenci, A. (2015). Extracting terms with EXTra. *In Proceedings of EUROPHRAS 2015.* (pp. 188–196). Málaga (Spain), July 2015.

[C4] Castagnoli, S., Lebani, G. E., Lenci, A., Masini, F., Nissim, M. and **Passaro, L. C.** (2015). POS-patterns or Syntax? Comparing methods for extracting Word Combinations. *In Proceedings of EUROPHRAS 2015.* (pp. 116-128). Málaga (Spain), July 2015.

[C5] **Passaro L. C.**, Pollacci, L. and Lenci, A. (2015). ItEM: A Vector Space Model to Bootstrap an Italian Emotive Lexicon. *In Proceedings of CLiC-it 2015.* (pp. 215–220). Trento (Italy), December 2015.

[C6] Nissim, M., Castagnoli, S., Masini, F., Lebani, G. E., **Passaro, L. C.** and Lenci, A. (2015). Automatic extraction of Word Combinations from corpora: evaluating methods and benchmarks. *In Proceedings of CLiC-it 2015.* (pp. 204-208). Trento (Italy), December 2015.

[C7] **Passaro, L. C.** and Lenci, A. (2016). Evaluating Context Selection Strategies to Build Emotive Vector Space Models. *In Proceedings of LREC 2016.* (pp. 2185-2191). Portorož (Slovenia), May 2016.

[C8] **Passaro, L. C.**, Bondielli, A. and Lenci, A. (2016). FB-NEWS15: A Topic-Annotated Facebook Corpus for Emotion Detection and Sentiment Analysis. *In Proceedings of CLiC-it 2016.* (pp. 228-232). Napoli (Italy), December 2016.

Others

[O1] **Passaro, L. C.**, Lebani, G. E., Pollaci, L., Chersoni, E., and Lenci, A. (2014). The CoLing Lab system for Sentiment Polarity Classification of tweets. *In Proceedings of EVALITA 2014.* (pp. 87-92). Pisa (Italy), December 2014.

[O2] **Passaro, L. C.**, Bondielli, A. and Lenci, A. (2016). Exploiting Emotive Features for the Sentiment Polarity Classification of tweets. *In Proceedings of EVALITA 2016.* (pp. 205-210). Napoli (Italy), December 2016.

Contents

CHAPTER 1

Introduction

The goal of Sentiment Analysis is to enable computers to recognize and express emotions, and as such is part of the so-called Affective Computing research field [157]. Current affect detection systems focus solely on individual modalities or channels, such as face, voice and text [28]. In this work, we focus on text affective computing and on its two basic tasks, i.e. Sentiment Polarity Classification [109, 143, 219] and Emotion Detection [28, 193].

It is a well known fact that the reputation of public entities, such as organizations and public institutions is largely influenced by the behavior of all those posts that, for one reason or another, talk about them. The opinion that stake-holders express about a public entity is central for its reputation. In this scenario, the web is undoubtedly the crucial source of information and public entities are learning how to monitor the web, and in particular the Social Media to learn what the people think and/or say about them.

Tracking reputation without appropriate text mining tools can become very difficult, because company analysts would have to read hundreds of reports, articles and posts and then aggregate subjective information *in some way*. Most of the appropriate computational tools employ lexical resources to extract the sentiment orientation of a piece of text. Sentiment and emotion lexicons, in which lemmas are associated with the emotions they evoke, are valuable resources that can help the development of detection algorithms. However, almost all languages but English lack high-coverage and high-quality emotion inventories of this sort. Building these resources is very costly and requires a lot of manual effort by human annotators. A possible solution, alternative to the costly building of novel resources from scratch, is the transfer of a lexicon from a language to another one, but in our opinion this is a temporary solution because the connotation of lexical items is a cultural phenomenon that may vary greatly between different languages and time spans [38].

This thesis explores the possibilities offered by sentiment and emotion recognition in computing, focusing in particular on the case of the Italian language, due to its lacking of reliable relevant lexical resources.

1.1 Current approaches

Today, most of the Natural Language Processing (NLP) approaches to Sentiment Analysis are based on a binary categorization of emotions, articulated along the key opposition between POSITIVE and NEGATIVE emotions [108, 109, 144, 206]. Typically, these systems ascribe words like "rain" and "betray" to the same emotion class in that they both evoke negative emotions, without further distinguishing between the emotion they evoke, i.e. SADNESS and ANGER respectively.

A new trend is leading to the development of annotated emotive resources such as Wordnet Affect [194], EmoLex [123] and the Hashtag Emotion Lexicon [121]. If on the one hand these emotive lexicons can be exploited to classify texts on the basis of their polarity [30, 123], they cannot be directly compared each other because of the difficulty in defining what an emotion is. In fact, a number of taxonomies exist for emotions, and consequently, the algorithms produce not comparable multi-label classifications.

1.2 Our approach

The goal of this work is to build a NLP system that allows for a flexible emotion classification of texts. We plan to achieve this result by exploiting both distributional methods as well as text categorization and information extraction techniques. We try to develop a dynamic framework that was as economic as possible, in which lexical resources can easily be adapted to many contexts and domains.

One of the main outcome of this work is ItEM, a high-coverage emotion lexicon for Italian, in which each target term is provided with an association score with eight basic emotions including JOY, SADNESS, FEAR, DISGUST, SURPRISE, ANGER, TRUST and ANTICIPATION. ItEM is not a static lexicon, but it also provides a dynamic method to continuously update the emotion value of words, as well as to increment its coverage.

For the languages lacking lexical resources oriented to emotion recognition, the main advantage of using a resource of this kind is that it requires a minimal use of external resources and a little annotation work. Moreover the update of the resource can be mostly automatized and, theoretically, it could be applied to any existing emotion taxonomy.

1.3 Contribution and implementation

The goal of this dissertation is to advance the state of the art of research in the field of Text Affective Computing (i.e. Sentiment Analysis and Emotion Detection). It takes a two-pronged approach: a technological contribution intended to facilitate future research, and experiments which seek to shed light on high-level questions about the properties of models that can lead the development of theoretical work.

The technological contribution for practitioners and researchers is a software which allows experimenters to create and run experiments for the creation of their own lexicons without advanced programming techniques. This software is designed for the experiments often run in computational linguistics, but it can also be used by psycholinguists (who will mostly benefit of the informatic implementation) and computer scientists (who will mostly benefit of the treatment of the linguistic material). The most important contribution of the work is the infrastructure that facilitates the implementation of novel techniques for updating and mantaining the lexical resources at large scale. Building the infrastructure required the implementation of a set of libraries (cf. below) to build and explore automated models of word meaning based on the principles of distributional semantics. The experiments presented in this thesis address questions that assume a distributional semantics framework, but that do not rely neither on a particular theory of emotions nor on a fixed constraint set for developing the vector space models (VSMs).

It is hoped that this software will make it faster and easier to conduct experiments to create customized emotive lexicons as well as encourage researchers and users to increase the reproducibility of their work. In addition, the thesis reports on good and bad practices for the evaluation of the result, that is well known as the bottleneck of sentiment and emotion based projects. An ambitious long-term plan is to create an interface that will enable the real-time creation and update of a lexicon on the basis of a set of parameters including the corpus and the emotion taxonomy.

For the implementation, it has been necessary the creation of a framework for organizing and processing textual data coming from different sources, by taking advantage from both linguistic and information technology skills. Most of the algorithms described in this thesis have been implemented in Python Software Foundation[1]. When talking about NLP, the initial data are not structured texts written in natural language that have to be organized into corpora, or existing corpora that need to be adapted to specific applications.

As for the creation of corpora from scratch, we created two main crawlers interfaced with the two major social networks Facebook and Twitter. Once collected and anonymized, the corpora have been processed using state-of-the-art linguistic pipelines to identify sentences, lemmas and grammatical information. In addition, in order to deal with textual units greater than the single tokens, we created new algorithms [150] aimed at extracting collocations, multi-words terms and idioms starting from statistical and rule-based information (cf. section 3.2). Once the corpora have been collected or reorganized, it has been necessary to create efficient and customizable tools for extracting information from them. In particular, we created a pipeline to create an emotive lexicon starting from a PoS-tagged corpus and a list of seed words highly associated with the emotions belonging to a given taxonomy (cf. section 2.5.1). The pipeline consists of the following steps:

Collecting co-occurrences : The data is collected in a square matrix of co-occurrence counts. Context vectors are defined as the columns of the matrix. Such a matrix of co-occurrence counts (the co-occurrence matrix) is denoted by F (frequency).

[1] Python Language Reference, version 2.7. Available at `http://www.python.org`.

A cell f_{ij} registers the frequency of occurrence of the word i in the context of word j. In this implementation we assume that the context of a word is the list of its surrounding words. The output is a dense matrix of co-occurrence counts, obtained using two main filters as input: the parts of speech (PoS) of interest and the word window (i.e. co-occurrences are collected only for selected PoS occurring in a defined window of words).
parameters: the PoS-tagged corpus in CONLL format [135]; the context window; target and contexts; the parts of speech of interest.

Weighting : Raw frequencies are transformed into significance scores that are more suitable to reflect the importance of the contexts. The measures currently implemented are the Pointwise Mutual Information [31], the Local Mutual Information [57], the Log Likelihood Ratio [50] as well as an identity function weighting the n-grams with their raw frequency.
parameters: the dense matrix registering the co-occurrences; the weighting scheme.

Dimensionality reduction : The dimensionality reduction is accomplished by using Singular Value Decomposition (SVD).
parameters: the (weighted) dense matrix registering the co-occurrence; the number of dimensions.

Compute the centroid vectors : The vectors of the seed words are isolated in the co-occurence matrix and for each (emotion, PoS) a centroid vector is computed as the mean of the vectors belonging to that emotion and PoS.
parameters: a list of seed words encoded in a list of associations (emotion, word-PoS); the (weighted) dense matrix registering the co-occurrence.

Lexicon creation : The lexicon is created by computing a similarity measure between the target words and the centroid vectors, to obtain the degree of association between the word and the emotions in the taxonomy. Implemented similarity measures are the scalar product (dot product) of two vectors, the cosine similarity of two vectors and the euclidean similarity of two vectors.
parameters: the similarity function; the centroid vectors; the vectors of the target words.

1.4 Application areas

Sentiment Analysis and Emotion Detection have great potential if integrated as sub-components of other applications such as Customer Relationship Management, Journalism, Business Intelligence, Government Intelligence and E-commerce:

- In Customer Relationship Management it can be used to track sentiments towards companies and their services, products or others target entities by exploiting textual communication between the company and its customers;
- In Government Intelligence, it can be used to collect people's emotions and points of views about government decisions and initiatives;
- In Competitive Intelligence it could be useful in assessing the limitations of particular products and this information could lead the development of improved products or services [177];

– In the field of Business Intelligence, corporate interest in affective computing is growing, and the companies invest increasing amount of money in marketing strategies to collect and predict the customers' attitudes toward their products and brands to manage their web reputation.

The design of automatic tools able to process sentiments over the web in real time and to create condensed dashboards representing such sentiments and points of view, is actually one of the most important development areas as demonstrated by the exponentially increased number of publications in this field [109, 144].

1.5 Structure of the thesis

This work is organized as follows: Chapter 2 presents a review of different state-of-the-art methods for sentiment analysis and emotion detection. In particular, it focuses on the existing lexicons (section 2.2) and on the techniques aimed at the creation of new sentiment lexicons (section 2.3) using several methods including rule based approaches (section 2.3.1), Vector Space Models approaches (section 2.3.2), Wordnet based approaches (section 2.3.3) and other techniques (section 2.3.4) relying on word embeddings or graph theory. In addition, the chapter presents the methods used to classify texts on the basis of their polarity (section 2.4), focusing on the classical methods (mostly addressed to classify product and movie reviews) and to the methods aimed at classify micro-blogging texts such as tweets. Finally, section 2.5 is dedicated to the classification of emotions, starting from the different theories and taxonomies to the emotion classification of documents (section 2.8).

Chapter 3 describes the tools used for the annotation of the corpora in the present work. Linguistic annotation up to dependency parsing has been carried out using $T2K^2$ which is described in section 3.1. Term extraction (section 3.2) has been carried out using the tool EXTra.

Chapter 4 presents distributional approaches to build emotive lexical resources that can be used for sentiment analysis and emotion detection. Section 4.1 describes the creation of the Italian emotive lexicon ItEM, by means a three stage process including the collection of highly emotive words to be used as seeds, their distributional expansion and the validation of the complete system. Section 4.2 presents a new Italian corpus composed of Facebook posts. The corpus has been created by crawling the Facebook pages of the most important Italian newspapers. Such pages typically include a low number of posts written by the journalists and a very high number of comments provided by web readers. Given the nature of the corpus, the emotive content of the posts is often neutral, but they can inspire long discussions among readers, which can become useful material for sentiment analysis and emotion detection.

Chapter 5 describes experiments aimed at the sentiment polarity classification. The Coling Lab system was originally developed for the participation in the constrained run of the EVALITA 2014 SENTIment POLarity Classification Task (SENTIPOLC [14]). The system is based on three main steps: the data preprocessing and annotation, feature extraction and the classification model. Subsequently, the Coling Lab system [149] has been enriched with emotive features in its second step (i. e. feature extraction). In order to assure a high coverage, such features have been derived from emotive resources built on the social media corpus FB-NEWS15 [148]. The chapter makes a comparison between the original Coling Lab system and the enriched one, showing that the emotive connotation of the words of a tweet helps in determine its general polarity.

CHAPTER 2

Literature review

In recent years, computational linguistics has seen a rising interest in subjectivity, opinions, feelings and emotions. Such a new trend is leading to the development of novel methods to automatically classify the emotions expressed in an opinionated piece of text [108, 144], as well as to the building of annotated lexical resources like SentiWordNet [7, 38, 56, 123, 194].

Emotion detection can be useful in several applications. As an example, in Customer Relationship Management (CRM) it can be used to track sentiments towards companies and their services, products or others target entities. Other possible applications are in Government Intelligence (GI), to collect people's emotions and points of views about government decisions as well as to monitor the brand reputation in the field of Business Intelligence (BI).

Today, most of the approaches are based on a binary categorization of emotions, articulated along the key opposition between POSITIVE and NEGATIVE emotions.

Typically, then, these systems would associate words like *pioggia* "rain" and *tradire* "betray" to the same emotion class in that they both evoke negative emotions, without further distinguishing between the SADNESS-evoking nature of the former and the ANGER-evoking nature of the latter.

In Sentiment Analysis the sentiment is articulated along the opposition between POSITIVE and NEGATIVE. It is very difficult to define what is an emotion, and there is no universally accepted definition. The word emotion itself can be interpreted in a number of ways, and it is used colloquially for different things. For this reason, we can find different taxonomies and thus different ways and algorithms aimed at a multi-label emotion classification of texts.

Sentiment and emotions, unlike factual information are highly subjective. Given a target entity such as a product or a service, different people often had different experiences and thus different opinions about that entity. For example, we can consider the case in which person A bought a smartphone of a particular brand and had a very good experience with it. She naturally has a positive opinion. Another person B bought the same product and she got a defective unit. B has a very negative opinion about the smartphone. Second, different interests and different ideologies impact on the sentiment. For example, when the price of a stock is falling, a person A (who bought the stock when the price was high) is very sad, but another person B (who bought the stock when the price was down) is very happy because she got the opportunity to make good profits by selling the stock [109].

In addition, different domains can have different degrees of difficulty. Opinions about concrete entities like products and services are usually the easiest to deal with. On the contrary, opinions about social and political issues are much harder because of the abstractness of the topic and of the presence of sentiment expressions such as sarcasms and irony, which are very difficult to deal with.

These differences explain why many commercial systems are able to perform a reasonably good sentiment analysis on texts containing opinions about products and services, but which work badly on opinionated social and political texts.

2.1 Sentiment analysis

The research in Sentiment analysis has been mainly carried out at three different granularity levels [109]: document level, sentence level, and aspect level.

In this work we assume to know the target of a certain text, therefore the analysis will be focused on the sentiment analysis at sentence level and document level.

1. Document level: In document-level classification, the goal is to classify whether a whole opinionated document expresses POSITIVE or NEGATIVE sentiment [145, 205]. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This level of analysis assumes that each document expresses opinions on a single target entity such as a product or a service.

2. Sentence level: Sentence-level classification determines whether each sentence expresses a POSITIVE, NEGATIVE or NEUTRAL opinion. This level of analysis is closely related to subjectivity classification [218], which distinguishes sentences that express factual information (objective sentences) from sentences that express subjective views and opinions (subjective sentences).

3. Aspect level. Aspect level classification, also called feature level [76,107] is aimed at discovering sentiments on entities and/or their aspects. Instead of looking at standard language units such as documents, paragraphs, sentences and phrases, aspect-level analysis takes into consideration the opinions about a certain opinion target. For example, in a sentence like *L'azienda X ha conseguito un utile superiore alle aspettative nel periodo di crisi del mercato italiano* "The result of Company X exceeded expectations in the Italian market during the crisis period" it has no sense to classify the entire sentence as positive or negative. In this case, in fact, we would like to be able to classify the sentence as POSITIVE for *Company*

X while at the same time as NEGATIVE for the *Italian market*. To deal with aspect level classification, researchers typically exploit feature based Support Vector Machine [98, 212] and neural network models [46, 99, 197, 198]. In recent years neural models are of growing interest for their capacity to learn text representation from data without careful engineering of features, and to capture semantic relations between aspect and context words in a more scalable way than feature based SVM [198].

Polanyi and Zaenen [164] reviewed the most important interacting factors that make difficult the determination of the point of view of an author. The major distinction is between Sentence Based Contextual Valence Shifters and Discourse Based Contextual Valence Shifters. Sentence Based Contextual Valence Shifters include:

1. Negatives and intensifiers (e. g. *Mario Rossi è intelligente* "Mario Rossi is clever" versus *Mario Rossi non è intelligente* "Mario Rossi is not clever";

2. Distinguish between realis and irrealis events (e. g. *Probabilmente ti ama* "He probably loves you" versus *Ti ama* "He loves you");

3. Presuppositional items (e. g. in the expression *Mancato successo* "Failure to success", the term *mancato* "failure" transforms the positive valence of *successo* "success" into a negative property;

4. Irony (e. g. in the sentence *Il brillante organizzatore non ha risolto il problema* "The very brilliant organizer failed to solve the problem" the extremely positive connotation of *brillante organizzatore* "very brilliant" is reversed by the meaning of *non ha risolto il problema* "failed to solve the problem".

Discourse Based Contextual Valence Shifters include:

1. Connectors (e. g. in the sentence *Sebbene Mario Rossi sia un ricercatore brillante, è un pessimo insegnante* "Although Mario Rossi is a brilliant researcher, he is a horrible teacher", in computing the author's attitude towards Mario Rossi, *Sebbene* "Although" neutralizes the effect of *è un ricercatore brillante* "is a brilliant researcher" resulting in a negative assessment score for the sentence;

2. Discourse structure such elaborations (e. g. a constituent gives more detail about a constituent which is in a structurally accessible position in a discourse stack. For instance in the sentence *Mario Rossi corre molto. La settimana scorsa ha corso per 70 km.* "Mario Rossi runs a lot. Last week he ran 43 miles.", the second sentence illustrates the concept expressed in the dominating sentence;

3. Multi-entity evaluation (e. g. a wide variety of entities are discussed – some of which are evaluated positively and others negatively);

4. Genre and Attitude Assessment (e. g. the assessment is related to the genre of the communication);

5. Reported speech (e. g. the valence of the reported material is not ascribed to the author);

6. Subtopics (e. g. it is possible to split the document into subtopics as well as the point of view of the author can then be made relative to each subtopic);
7. Genre Constraints (e. g. movie reviews, which contain both information about the events and situations in the story and information about the film which has been created to tell the story);
8. Cultural Constraints (e. g. a word such as *Rivoluzione* “Revolution” might be positive or negative depending on the context and background of the writer).

Independently from the level of analysis, one of the most useful indicators to extract the sentiment orientation of a piece of text are the words which compose the text itself. For example the presence of positive words like *buono* “good”, *gioioso* “joyful” and *divertente* “amazing” helps in determining that a particular text expresses a positive sentiment, while the presence of words like *cattivo* “bad”, *povero* “poor” and *terribile* “terrible” indicates that such text is probably expressing a negative sentiment.

2.2 Available Sentiment Lexicons

The simplest sentiment lexicons encode the words along one dimension which is known in literature as sentiment, semantic orientation, valence or polarity. The following freely available lexicons have been developed in several contexts and are not comparable in size. Some of them such as ANEW [25, 124] derive from psycholinguistic experiments, others are manually collected such as SO-CAL [196] and others derive from the aggregation of several resources such as the Subjectivity Lexicon [219].

Harvard General Inquirer: the lexicon [190] provides a binary classification (positive | negative) of 4k sentiment words. It provides manually-classified and part-of-speech tagged terms with several markers such as the semantic orientation, cognitive orientation and mood of the speaker;

ANEW: the resource [25] provides manually assigned valence scores for 1,034 English words in line with the psycholinguistic dimensional theory of emotions [140]. Although the resource has a very low coverage, its precision is maximized, therefore it has been adapted to many languages, including Italian [124];

Subjectivity Lexicon: it is a list [219] of clues belonging to system *OpinionFinder*. The clues were extracted by combining from several sources including the General Inquirer, the list produced by Hatzivassiloglou and McKeown [73] and automatically compiled lists [169][1]. Examples of clues are the Part of Speech (PoS), the prior polarity of the word (word polarity when considered out of context), the strength of the subjectivity.

SentiWordNet (SWN): Sentiwordnet [7,56] has been developed using WordNet knowledge base [60]. In SWN each WordNet synset is associated with the numerical scores Pos(s) and Neg(s), ranging from 0 to 1. The scores represent the positive and negative polarity of each entry, that takes the form $lemma\#pos\#sense-number$;

[1] The creation of the automatically created lexicons will be described in section 2.3

AFINN: the list includes English words rated for valence in a Likert scale ranging from -5 (negative) and $+5$ (positive). Overall, it contains 2,477 manually labeled words and phrases [134];

SO-CAL: the resource [196] has been manually tagged by a small number of annotators with a multi-class label ranging from *very negative* to *very positive*. The ratings were validated through crowdsourcing, ending up with a list of roughly 4k words;

NRC Word-Emotion Association Lexicon (EmoLex): is a lexicon [123] of 14,000 English words associated with the eight basic emotions proposed by Plutchick [159] and two sentiments (negative and positive), compiled by exploiting crowdsourcing and manual annotation.

Sentix: the Sentiment Italian Lexicon [15] is a lexicon for Sentiment Analysis in which 59,742 lemmas have been automatically annotated for their polarity and intensity, among other information. Polarity scores range from -1 (totally negative) to 1 (totally positive), while Intensity scores range from 0 (totally neutral) to 1 (totally polarized). The lexicon has been created by annotating a lexicon of senses for Italian starting from existing resources such as Wordnet [60], MultiWordNet [155] and SentiWordNet [7, 56].

2.3 Creating Sentiment Lexicons using Semi-Supervised Learning

The greatest problem in approaching Sentiment Analysis Tasks using existing resources is that very often their lexical coverage is too low. One of the most powerful ways to create new sentiment lexicons or to expand some of them is semi-supervised learning. The following algorithm (Figure 2.1) synthesizes the schema used for Sentiment lexicon generation. In practice, the systems identify a list of seeds which is a list of words (terms, phrases etc.) with a known polarity[2]. Once having compiled such list, it is divided into a positive set of seeds ($posSeeds$) and a negative set of seeds ($negSeeds$) and each group is expanded with its neighbors (words with similar polarity) until the expanded sets ($posLex$,$negLex$) have been computed. Different approaches diverge on how words of similar polarity are identified (FINDSIMILARWORDS), in the stopping criterion (Until done) and in the post-processing phase (POSTPROCESS) [87].

In next sections we introduce the three major methods to approach the task of generating a sentiment lexicon. In the first one (section 2.3.1) we show the methods based on pattern coordination, that combine linguistic cues to identify sentiment words and determine their orientations. A second approach (section 2.3.2) relies on the Harris's distributional hypothesis (DH), which states that at least part of the meaning of a linguistic item can be inferred from its distributional properties. This paradigm, in fact, assumes that words with similar polarity tend to occur together, so that to calculate the degree of positiveness and negativeness of a target word, it takes into consideration the neighborhood co-occurrence, wich is estimated using the Pointwise Mutual Information. A third approach (section 2.3.3) exploits external resources such as Wordnet [60] to infer the polarity of the words. The intuition is that the synonyms of a word probably

[2]In the simplest case, the system employs two single words as seed, a word for positive polarity, and a word for negative polarity.

function BUILDSENTIMENTLEXICON($posseeds$,$negseeds$)
returns $poslex$,$neglex$
  $posLex \leftarrow posSeeds$
  $negLex \leftarrow negSeeds$
  Until done
    poslex←poslex + FINDSIMILARWORDS($posLex$)
    neglex←neglex + FINDSIMILARWORDS($neglex$)
  $poslex$,$neglex$←POSTPROCESS($posLex$,$neglex$)

Figure 2.1: *Schematic algorithm for sentiment lexicon generation using semi-supervised learning methods [87]. The funcion* FINDSIMILARWORDS*, the stopping criterion and the* POSTPROCESS *step depend on the paradigm used for lexicon generation.*

share its polarity as well as its antonyms have an opposite one. Given a set of words with known polarity, these systems exploit synonymy and antonymy in Wordnet [60] to find candidate words with similar (or opposite) polarity.

Section 2.3.4 is left to discuss additional methods based on word embeddings, graph theory and on the combination of different of the previous schemes.

2.3.1 Pattern Coordination based approaches

The methods based on pattern coordination combine linguistic rules or conventions to identify sentiment words and determine their orientations. The first idea, named Sentiment Consistency, was by Hatzivassiloglou and McKeown [73]. The authors used a set of 1,336 hand-labeled seed adjectives with known polarity and a corpus to find other adjectives with the same (or opposite) orientation. Intuitively, if two adjectives are placed around the conjunction AND, probably they share the same orientation and, on the contrary, two adjectives placed around BUT have opposite sign. For example in the sentence *Questo appartamento è elegante **e** luminoso* "This apartment is elegant **and** bright" this seems to work very well. Along this line, other rules can be also designed for other connectives, such as OR, BUT, EITHER − OR, and NEITHER − NOR. However, although this idea is called sentiment consistency, in practice, it is not always consistent. In fact, the problem of this approach could easily be found in a sentence like *Questo appartamento è elegante e scomodo* "This apartment is elegant and uncomfortable". In order to solve this issue, the authors then applied a supervised classifier trained on a subset of the hand-labeled seed words and returning the probability that two words are of the same or opposite polarity. The system was trained on three features: (i) occurrence with AND; (ii) occurrence with BUT; (iii) morphological negations.

This approach has been extended by Kanayama and Nasukawa [91] in which the authors introduced the concepts of intra-sentential (within a sentence) and inter-sentential (between neighboring sentences) sentiment consistency, namely the coherency. The intra-sentential consistency corresponds basically to the sentiment consistency and inter-sentential consistency extends the approach to the neighboring sentences by exploiting expressions such as BUT and HOWEVER to capture the sentiment changes.

A similar criterion is also used by Kaji and Kitsuregawa [88, 89] to determine whether to add a word to the positive or negative lexicon. In particular, the authors exploited web pages and extracted many heuristics such as the surround text of the tables containing "Pros" and "Cons" of a particular entity and they used them to attribute a sentiment orientation to the target sentences.

2.3.2 VSM based approaches

This second approach is based on the Harris's Distributional Hypotesis (DH), which states that at least part of the meaning of a linguistic item can be inferred from its distributional properties. As an implication, the assumption is that words with similar polarity tend to occur together.

From a computational point of view, the DH is modeled using Vector Space Models (VSM), that will be described in the next subsection. When applied to Sentiment Analysis, VSMs allow us to calculate the polarity of a target word, starting from its neighborhood co-occurrence. The first research of this kind was by Turney [205], later expanded with Littman in [206]. In these works, the polarity (that they called Sentiment Orientation) is estimated using the Pointwise Mutual Information (PMI [31]) and Latent Semantic Analysis (LSA [100]). The method is described below.

Vector Space Models

Distributional semantics includes several approaches characterised by a usage-based perspective on meaning and by the idea that «difference of meaning correlates with difference of distribution» [71].

Vector space models (VSMs) of semantics are a very popular framework in computational lexical semantics. Their idea was initially proposed by Salton and colleagues in 1975 [180], who designed an Information Retrieval (IR) system, namely SMART, that anticipated many of the concepts used still today in the contemporary research on search engines[3].

VSMs have been widely extended in the following years to several Natural Language Processing tasks. According with the Harris's DH, these models «provide access to semantic content on the basis of an elementary principle which states that semantic proximity can be inferred from proximity of distribution» [58]. The construction of VSMs is performed in a four-step process [58, 207]:

1. for each target word, contexts are collected and counted;
2. raw frequencies are usually transformed into significance scores that are more suitable to reflect the importance of the contexts;
3. the resulting matrix tends to be very large and sparse, therefore it is reduced;
4. a similarity score is computed between the vector rows, using various similarity measures.

[3]The task of document retrieval was designed as a sort of overlapping between the words expressed in a query and the words belonging to a document, and the idea behind the VSM was to represent each document as a collection of vectors (a vector space), where the basis vectors of the space are words. The query is also represented as a vector, in the same space of the other documents, so it can be seen as a pseudo-document. The closer two vectors are, the higher is their semantic similarity. In this way, the system returns to the user the documents which are "similar" to their queries according to the similarity measure between the document and the query.

Different VSMs differ on several parameters that can be set up at each step of the process, such as the corpus, the definition of the context, the weighting schemes and the techniques used to smooth the matrix.

First of all, since they are corpus-based models, VSMs reflect the type of the corpus in which they are constructed (i. e. the corpus in which the words are counted).

Second, it's very important to define what is the *context*. For example, in document-based models, as in Latent Semantic Analysis (LSA) [100] words are similar if they appear in the same documents or in the same paragraphs. In word-based models the context is modeled in a *bag-of-word* window of collocates around the target words [178], so that the words are similar if they appear with the same context words (cf. below).

From a linguistic perspective [178] we can distinguish between syntagmatic and paradigmatic (similarity) relations: **Syntagmatic relations** concern positioning, and relate entities that co-occur in the text (relation *in praesentia*). This relation is linear, and applies to linguistic entities that occur in sequential combinations. **Paradigmatic relations** concern substitution, and relate entities that do not co-occur in the text (relation *in absentia*). Paradigmatic relations hold between linguistic entities that occur in the same context but not at the same time (see figure 2.2 taken by Sahlgren [178]).

	Paradigmatic relations Selections: "x or y or..."			
Syntagmatic relations	she	adores	green	paint
Combinations:	he	likes	blue	dye
"x and y and..."	they	love	red	colour

Figure 2.2: *The syntagmatic and paradigmatic relations. Taken by [178].*

From an algorithmically point of view, Sahlgren [178] makes a distinction between two types of algorithms for distributional meaning acquisition, the first one consisting in building distributional profiles for words based on which other words surround them, and the second one in building such profiles on the basis of the text regions in which the words occur. More specifically, the first approach, which is typical in information retrieval systems, exploits syntagmatic relations: the assumption is that words with a similar meaning tend to occur in the same contextual unit (the document) because they define the topic of that document[4]. On the contrary, the second approach uses paradigmatic relations, because in a small context window we don't expect that similar words such as synonyms co-occur, but we rather expect that their surrounding words will be the same.

The third parameter consists in the variation of weighting scores and similarity measures. In the field of IR, the most popular way to weight matrices is the Term Frequency $\times$ Inverse Document Frequency ($TF - IDF$) [85] in which an element gets a high weight when the corresponding term is frequent in the corresponding document (i. e. high TF), but the term is rare in other documents (i.e., low DF, and thus high IDF).

[4]If similar words are used in a query and in a document, this will probably mean that the document is relevant for the topic of the query.

Equation 2.1 shows the formula to compute $TF - IDF$, where TF is in Equation 2.2 as well as IDF is in Equation 2.3[5].

$$(TF - IDF)_{i,j} = TF_{i,j} \times IDF_i \tag{2.1}$$

$$TF_{i,j} = \frac{n_{i,j}}{|d_j|} \tag{2.2}$$

$$IDF_i = \frac{|D|}{d : t_i \in d} \tag{2.3}$$

In computational linguistics, the most common practice is to use Positive Pointwise Mutual Information (PPMI: [31, 204]) as weighting scheme (i. e. to measure syntagmatic co-occurrences) and cosine as similarity measure (i. e. to measure paradigmatic relations), which are typically credited for granting the best performances across a wide range of tasks that work with word–context matrices [27, 58, 207].

PMI is calculated as follows:

$$PMI(x, y) = log_2 \frac{O(x, y)}{E(x, y)} \tag{2.4}$$

where $O(x, y)$ is the observed co-occurrence frequency and $E(x, y)$ is the expected frequency under the null hypothesis of independence. In PPMI, all negative PMI values are replaced with zero [136].

In other cases [12], VSMs are weighted using the Local Mutual Information (LMI), whose formula is presented in Equation 2.5. The main difference between PMI and LMI is that the first favors more idiosyncratic, low-frequency expressions, while the latter has a greater bias towards frequent expressions [57].

$$LMI(x, y) = O * log_2 \frac{O(x, y)}{E(x, y)} \tag{2.5}$$

As in Equation 2.4, in 2.5, $O(x, y)$ is again the observed co-occurrence frequency and $E(x, y)$ is the expected frequency under the null hypothesis of independence.

Since explicit co-occurrence vectors are huge and sparse, to reduce their dimension and limit computational complexity, the most common approach consists in mapping the original sparse matrix into a low-dimensional dense matrix by exploiting reduction methods such as Singular Value Decomposition (SVD) [100]. Matrix reduction techniques smooth unseen data, remove noise and exploit redundancies and correlations between the linguistic contexts, thereby improving the quality of the resulting semantic space [58, 207].

In order to measure the paradigmatic similarity between two words, the standard approach is to calculate the angles between their vector representations, being the cosine similarity (cf. Equation 2.6). A score close to 1 indicates similarity, while a score close to 0 means they are completely unrelated.

[5] In Equation 2.2, $n_{i,j}$ is the number of times in which the term t_i appears in the document dj (whose length is at the denominator). In equation 2.3, $|D|$ is the cardinality of the collection and the denominator represents the documents which include the term t_i.

$$cos(\overrightarrow{\boldsymbol{w}_1}, \overrightarrow{\boldsymbol{w}_2}) = \frac{\overrightarrow{\boldsymbol{w}_1} \cdot \overrightarrow{\boldsymbol{w}_2}}{||\overrightarrow{\boldsymbol{w}_1}|| \cdot ||\overrightarrow{\boldsymbol{w}_2}||} \tag{2.6}$$

Turney and Turney and Littman's approach

A pioneering work in exploiting VSMs to create Sentiment Lexicons was made by Turney and Littman [206] in which the authors extended the PMI-based method described in [205] to compute the semantic orientation of a word[6].

Expanding the Turney's previous study [205], Turney and Littman computed the SO of a word by measuring its association with a set of positive words minus the strength of its association with a set of negative words. For their experiments, the authors used the following seed sets:

POSITIVE = {good, nice, excellent, positive, fortunate, correct, superior}

NEGATIVE = {bad, nasty, poor, negative, unfortunate, wrong,inferior}

According with the assumption of several approaches to Sentiment Analysis [73, 76, 107, 217], the terms selected from Turney and Littman were adjectives, that are considered the best indicators of the subjective content of a sentence. In their first experiment, in order to measure the association strength, the authors used the PMI [31]. The target words were classified as positive if $O(w) > 0$, and as negative otherwise.

$$SO(w) = \sum PMI(w, t(i)_{POSITIVE}) - \sum PMI(w, t(j)_{NEGATIVE}) \tag{2.7}$$

Since the PMI measures the degree of association between two words by the frequency with which they co-occur, by modeling the VSM using the PMI, Turney and Littman's assumption was that words tend to share the same SO of their neighbors. However the authors reflected on the possibility that such word-word co-occurrence approaches were able to capture *relatedness* of words, but not specifically to address *similarity of meaning*.

On the other hand, word-context co-occurrence approaches may correlate better with human judgments of semantic similarity [101], so that in the last part of their research the authors tried to model the semantic orientation as a function of the paradigmatic relations between the target words and the seeds, by implementing LSA.

In they experimental results, the authors applied the SO-PMI model and the SO-LSA measure to the smallest of their document sets, and they noticed improvement in the performance of the SO-LSA system.

The authors used an online demonstration of LSA [49][7] and they chose the corpus TASA-ALL, which was the largest among the available corpora[8]. The authors generated a word-context matrix smoothed using SVD to 300 dimensions.

[6]This approach was aimed at Sentiment Polarity Classification of costumer reviews using unsupervised learning. The method used by Turney in [205] is illustrated in section 2.4 (page 21).

[7]Available at `http://lsa.colorado.edu/`

[8]TASA-ALL is composed of a set of short English documents gathered from several sources and it contains approximately 10 million of words.

In order to compare the SO-PMI and SO-LSA models, Turney and Littmann used the corpus TASA with the lexicon of Hatzivassiloglou and McKeown [73] as a test set and they found that LSA outperforms PMI indipendently of the dimension of the test set. Table 2.1 shows the comparison on different portions of the dataset.

PERCENT OF TOT.	SIZE OF THE TEST SET	SO-LSA	SO-PMI
100%	1336	67.66%	61.83%
75%	1002	73.65%	64.17%
50%	668	79.34%	46.56%
25%	334	88.92%	70.96%

Table 2.1: *Comparison between the SO-LSA and the SO-PMI method in [206].*

In particular, the authors noticed that the performances have approximately the same accuracy when evaluated on the full test set (threshold 100%) but the LSA model is remarkably better when the percentage of the test set is decreased and that PMI seems to be less stable than LSA, especially when the percentage drops below 75%.

2.3.3 Wordnet based approaches

The third approach exploits external resources such as Wordnet [60] to infer the polarity of the words. The basic intuition is that the synonyms of positive words tend to have positive polarity and the synonyms of negative words, on the contrary, tend to have negative polarity. Given a set of words with known polarity, the systems based on Wordnet exploit synonymy and antonymy relations in Wordnet [60] to bootstrap the small lexicon with candidate words with similar (or opposite) polarity.

In this approach, the lexicons are in general built in an iterative process: First of all, the method requires a manual collection of a relatively small sets of labeled seeds with opposite polarity. Consequently, these sets are expanded by searching each of them in a dictionary, and by adding its synonyms to the set of the seeds having the same polarity, and its antonyms to the opposite one. The process of seed expansion ends when no more new words can be found [76, 208]. Once completed this expansion process, an additional manual step is required to fix the list by removing errors.

An extension of this work proposed by Kim and Hovy [95, 96] assigns a sentiment strength to each word using a probabilistic method: The authors built a sentiment lexicon by exploiting the structure of WordNet and three sets of seeds (positive, negative and neutral). They expanded the seeds by collecting synonyms from WordNet [60], but since the system produced many errors[9], the authors introduced a Bayesian formula to compute the closeness of each word to each category. In other words, they implemented an additional step to determine the most probable class (Positive, Negative, Neutral) for a given word.

Kamps and colleagues [90] used WordNet [60] to build a synonymy network able to connect pairs of synonymous words. The system works on adjectives: the semantic orientation of an adjective is decided on the basis of the shortest path connecting the adjective and two (opposite) seeds w^+ and w^- as follows:

[9]The errors are due to the multiple senses of the words in Wornet and to the fact that some common words (i. e. “great”, “strong”, “take”, and “get” occurred many times in both positive and negative categories.

$$SO(t) = \frac{d(t, w^-) - d(t, w^+)}{d(w^+, w^-)} \quad (2.8)$$

In this way, the word t is positive iff $SO(t)$ >0, and it is negative otherwise. Intuitively, the higher is the absolute value of $SO(t)$, the stronger is the semantic orientation of the adjective.

Esuli and Sebastiani [54] exploited the semantic structure of Wordnet [60] and supervised learning to infer the polarity of a given word. In order to build the training set, they expanded a given set P (positive seeds) and a given set N (negative seeds) with their synonyms and antonyms in WordNet. Once the sets of the seeds have been expanded, they built a binary classifier to decide if a word is positive or negative on the basis of the word vector representation.

Along this line, in a follow-up study the authors [55] included the category objective (neutral). The output of this algorithm is SentiWordNet [56], which is a lexical resource in which each synset of WordNet is associated with three numerical scores $Obj(synset)$, $Pos(synset)$ and $Neg(synset)$, describing the degree of membership to each set.

Guerini and colleagues [67] measured the word prior polarity scores for each sense of the word in SentiWordNet [7, 56]. They compared a number of techniques for the estimation of the polarity across tasks and datasets. In particular, the authors focused the research in determining the «prior polarities from the posteriors»: Sentiwordnet provides polarities scores according to each word sense. For example, the word *freddo* "cold" has a different posterior polarity for the meaning of *bassa temperatura* "having a low temperature" (i. e. *birra fredda* "cold beer") and for the meaning of *senza emozioni* "being emotionless" (i. e. *persona fredda* "cold person"). Since several methods to compute prior polarities starting from posterior ones have been used in the literature, the authors compared them across tasks and datasets and proposed a framework combining a number of these approaches. To address this issue, the authors trained a classifier using various aggregations of the scores as features. Their experimental results showed that outperforming results can be obtained by combining different approaches.

2.3.4 Other approaches

This section describes additional approaches to build Sentiment Lexicons, based on graph theory, word embeddings and on the combination of other schemes.

A first group of techniques relies to Graph based semi-supervised learning. For example, Blair-Goldensohn and colleagues [17], bootstrapped three seed sets (positive, negative, neutral) using an adapted version of the label propagation algorithm [226]. In particular, they used the neutral set to stop the propagation as follows: At the beginning, the edge weights were assigned on the basis of a scaling parameter for different types of edges which represent the relation of synonymy and antonymy. Each positive seed received the score of $+1$, each negative seed was given the score of -1, and all other words were given the score of 0. The input values were then modified during the propagation process and the final scores were derived by logarithmically scaling their polarity degrees.

Rao and Ravichandran [166] used three semi-supervised techniques [20, 21, 226] to separate the positive words from the negative ones in a graph generated by means of bootstrapping. In graph theory Mincut algorithms are used to divide a graph into partitions minimizing the number of nodes which have a strong probability to be placed in different partitions. Rao and Ravichandran [166] employed only the synonym relationship as a similarity metric between two nodes in the graph. The results show that mincut and randomized mincut produced better F scores, but label propagation gave significantly higher precision with low recall.

Hassan and colleagues [72] exploited random walks elements to generate a lexicon. The distance between two words in the graph was defined by considering the hitting time (h), which is the average number of hops to go from a node i to a node belonging to the set of the positive seeds (S^+) or negative seeds (S^-). A word $\langle w \rangle$ is classified as positive if it respects Equation 2.9, and as negative otherwise.

$$h(w|S^+) > h(w|S^-) \tag{2.9}$$

Other works combine both WordNet and heuristics to bootstrap a small lexicon. For example, given a set of seed words, Dragut and colleagues [47] instead of following the dictionary, proposed a deductive process based on inference rules to determine the sentiment orientations of the target words. The input of the algorithm is a set of words with known sentiment orientations and its output is a set of polarity-oriented synsets.

Volkova and colleagues [211] bootstrapped Twitter-specific subjectivity lexicons to generate a sentiment lexicon from tweets. Their algorithm is based on a set of polarity labeled tweets. On each iteration, some tweets in the unlabeled data are labeled using the current lexicon. If a tweet contains one or more terms belonging to the lexicon, it is marked as *subjective*, otherwise it is marked as *neutral*. Then, for every new term that has a frequency higher of a given threshold, they calculate the probability of that term being subjective and they added it to the *Extended Lexicon* according to a threshold-based filter. In order to determine the polarity of the new term, they calculated the probability of the term appearing in positive or negative tweets.

A most recent trend consists in creating Sentiment Lexicons by exploiting continuous word vector representations (word embeddings) by means of Deep Learning techniques [16, 77, 117, 203] instead of using more classical ones such as Latent Semantic Analysis and Latent Semantic Indexing [49]. Without entering in the details of Word Embeddings (WE), such methods, in the simplest case (as in the PMI approaches), use the vector representations to model the semantic orientation of the words starting from their proximity with a set of polarity-labeled seeds.

For example García Pablos and colleagues [141] recently published a simple technique to calculate a word polarity value for domain words using continuous WE [117]. Their goal was to see if the WE calculated on an in-domain corpus can be directly used to obtain a polarity measure of the domain vocabulary with no additional supervision. The work presents the experiments to compare different approaches proving that a lexicon built using WE can be easily adapted to new domains.

Rothe et al. [172] proposed a method to create Sentiment lexicons by exploiting WE[10]. In particular, they introduced DENSIFIER, which is «a method that learns an orthogonal transformation of the embedding space» [172]. Given a lexicon resource, their objective is to minimize the distances in the sentiment dimension between words

[10]The authors state that DENSIFIER does not need a text corpus, but can transform publicly available word embeddings. In practice this algorithm is independent from the embedding learning algorithm and therefore extensible to other word embedding models.

of the same polarity group and to maximize the distances between words of different groups. The input of the system is a set of embeddings and a small sentiment-annotated lexicon and the output is a lexicon consisting of all words covered by the embedding set, each associated with its one-dimensional (i. e. the polarity dimension) ultradense subspace representation.

Tang et al. [199] proposed to integrate the sentiment information of tweets into neural network to learn sentiment specific word embedding. They constructed a sentiment-specific word embedding based on the C&W [34] model but adding tweet sentiment information. They developed three neural network models to train sentiment-specific word embedding from massive distant-supervised tweets collected with positive and negative emoticons. These methods have shown promising results when combined with linear or neural classifiers [199, 209].

Hamilton and colleagues [70] proposed SENTPROP, a framework to learn accurate sentiment lexicons from small sets of seed words and domain-specific corpora. Algorithmically, the approach is inspired by Velikovich and colleagues [210], so that they derive the sentiment lexicon from the web by exploiting a graph propagation algorithm: The system assumes as input the sets of positive and negative seed phrases and calculates the sentiment of each of them by considering the cosine similarity between a target sentence and the set of the positive and negative seed phrases. Starting from this idea, Hamilton et al.'s approach, incorporates high-quality word vector embeddings to build the graph and an alternative label propagation algorithm as well as a bootstrapping method to obtain confidence scores.

2.4 Sentiment Polarity Classification

Sentiment Polarity Classification (SPC) is aimed at classifying a document such as a customer review or simply a social media post as expressing a positive or a negative sentiment (polarity). The task considers each document as a whole, therefore it belongs to the document-level analysis tasks.

SPC is considered the simplest sentiment analysis task because it treats sentiment classification as a traditional text classification problem by considering the polarities (positive and negative) as the classes. As reported by Liu [107, 109], the assumption of this task is that a document (for example a product review) reports opinions on a single entity from a single opinion holder, then its application is restricted to the documents respecting the constraints about (i) unique holder; (ii) unique opinion target.

From an algorithmically point of view, the task can be dealt using Unsupervised Learning (UL), consisting of classification methods based on syntactic patterns, web search and sentiment lexicons or using Supervised Learning (SL), which includes both standard Machine Learning algorithms and Custom Score Functions.

Unsupervised Learning

Unsupervised methods for SPC are based on two major approaches, the first one is based on the Turney's algorithm [205] described below, and the other one is based on Sentiment Lexicons.

In 2002, Turney [205] introduced a method to classify the semantic orientation (SO) of a given sentence by exploiting syntactic patterns. The algorithm is based on three steps:

1. Patterns of PoS tags such as [ADJECTIVE + NOUN] or [ADVERB + VERB] are extracted;

2. the SO of the extracted phrases is estimated using the PMI [31]);

3. the average SO of all phrases in the review is computed and the sign (positive or negative) of this value is used to classify the review.

In particular, the SO is calculated as follows:

$$SO(phrase) = PMI(phrase, excellent) - PMI(phrase, poor) \tag{2.10}$$

where the reference words "excellent" and "poor" were chosen because, in the five star review rating system, it is common to define one star as "poor" and five stars as "excellent". The SO is positive when the phrase is more strongly associated with *excellent* and negative when phrase is more strongly associated with *poor*. Variations of this scheme are in Feng and colleagues [61], in which the authors compared PMI with three additional association measures (Jaccard [80], Dice [187], and Normalized Google Distance [33]) using different corpora (Google indexed pages, Google Web IT 5-grams, Wikipedia, and Twitter). Their experimental results confirmed that the PMI is the best metric to measure the SO.

Other unsupervised approaches include methods based on Sentiment lexicons. These schemes assume to have available dictionaries of sentiment words and phrases, with their associated SO and strengths [76, 95]. In order to refine the computation of the SO of a given sentence or document, these system are integrated with intensification and negation [92, 195, 196]. Many variations of this approach differ in assigning a sentiment score to the sentiment expression, in handling negations and in considering additional information.

For example, Taboada and colleagues [196] proposed a Semantic Orientation CALculator (SO-CAL) based two assumptions: (i) individual words have a prior polarity, that is, a SO that is independent of the context; (ii) this SO can be expressed as a numerical value. The calculation of the SO value for each sentiment expression ranges from -5 (extremely negative) to $+5$ (extremely positive). Each intensifier or diminisher is associated with a positive or negative percentage weight respectively. In this work the authors refined the classical approach to negation, because instead of reversing the sign of the SO, they shifted it toward the opposite polarity by a fixed amount (i. e. they treated the negator as a diminisher). In addition, the authors proposed a new approach to deal with a number of markers indicating that the words appearing in a sentence might not be reliable for Sentiment Analysis, namely irrealis blocking. These markers usually indicate nonfactual contexts and are referred to irrealis moods such as modals, conditional markers, negative polarity items like *niente* "anything", *prevedere* "expect", *dubitare* "doubt", questions, and words enclosed in quotation marks. The semantic orientation of any word in the scope of an irrealis marker is ignored.

Other methods include approaches for specific applications. For example Tong [202], proposed an application able to generate sentiment timelines. In particular, the system tracked online discussions about movies and displayed a plot of the number of positive and negative messages (Y-AXIS) over time (X-AXIS). The messages were classified by matching manually compiled phrases that indicate the sentiment of the author toward a movie, for example, «great acting», «wonderful visuals», «uneven editing», and «highly recommend». Unfortunately, this lexicon is thus specific to the domain and needs to be compiled anew for each new domain. Of course, a possible extension of this approach is to exploit supervised learning to automatically collect the domain expressions and then to plot them in sentiment timelines.

Supervised Learning

Another natural approach to deal with SPC consists in using classification method such as naïve Bayes classification or support vector machines (SVM) [36, 84]. In this case, training and testing data are traditionally product reviews [143, 205] because most of online reviews are provided with Likert-X scale directly attributed by the reviewer. The most important features used to train SPC models are the following:

- Terms and frequency: Individual words (unigram) and n-grams with associated frequency counts. Sometimes, the frequency counts are weighted with $TF-IDF$ [85] or other weighting schemes;
- Part of speech (PoS): Different parts of speech have different impact in discriminating the sentiment orientation of a text. For example, several studies demonstrated that adjectives are the most important sentiment indicators [73, 76, 107, 217], but in general all PoS tags and their n-grams can be used as features to train a SPC model;
- Sentiment words and phrases: Sentiment words are natural features as they are words in a language for expressing positive or negative sentiments. For example, *buono* "good", *splendido* "wonderful", *sorprendente* "amazing" are positive sentiment words, and *cattivo* "bad", *povero* "poor" and *terribile* "terrible" are negative sentiment words. Most sentiment words are adjectives but also nouns (e.g. *rifiuti* "rubbish" and *cacca* "crap") and verbs (e.g. *odiare* "hate" and *amare* "love") can express sentiments. Besides individual words, there are also sentiment phrases and idioms such as *essere al settimo cielo* "to be on cloud nine" and *perdere le staffe* "hit the ceiling";
- Sentiment shifters: Expressions that are used to change sentiment orientation of a portion of text, for example, from positive to negative or vice versa (i. e. negation words as in *non buono* "not good");
- Syntactic dependency: Words' dependency-based features generated from parsing or dependency trees;
- Shallow features: numeric features such as counts or binary features that capture the presence of emoticons, abbreviations, and intensifiers (e.g., all caps or character repetitions). This class of features is often used to classify micro-blogging texts.

Although this set of features can be seen as quite standard, the way in which they are implemented and combined can make the difference in terms of overall performance.

In the following pages we show several works that are based on laborious feature engineering.

One of the well known approaches to classify reviwes was designed by Pang and Lee [143]. Their method was based on dividing the SPC problem into two sub-tasks, the first one aimed to recognize the degree of subjectivity expressed by a text, and the second one to attribute a polarity to the subjective texts. To implement this idea, instead of using the complete text of the review, Pang and Lee [143] applied ML methods to the portion of the texts previously classified as subjective. Such portions were in fact more likely to contain sentiments. In order to identify these portions in a complete review, they divided the review into sentences and applied a standard classification algorithm to classify each of them as subjective or objective.

In the same period, Gamon [63] classified customers' feedback data using NLPWin[11] to extract the following groups of features: (i) part-of-speech trigrams; (ii) constituent specific length measures (length of sentence, clauses, ratio between adverbial/adjectival phrases, and noun phrases); (iii) constituent structure in the form of context free phrase for each constituent in a parse tree; (iv) PoS information with semantic patterns (e.g. [VERB - SUBJECT - NOUN] indicating a nominal subject to a verbal predicate; (v) other features such as transitivity of a predicate and tense information.

Mullen and Collier [126] combined n-grams with other groups of features extracted by exploiting the research made by Turney [205] and Osgood and colleagues [140]. Besides n-gram features, they included: (i) SO calculated using the PMI following the approach in [205]; (ii) Osgood [140] Semantic Differential with WordNet [90] namely evaluation, potency, and activity[12]; (iii) Topic proximity and syntactic-relation features, in which they combined sentiment values of words or phrases in [205] and [140].

Beside these features, several systems exploit syntactic relations to enrich the classification scheme. Ng and colleagues [133] tested dependency relations ([ADJECTIVE-NOUN], [SUBJECT-VERB] and [VERB-OBJECT]), but no performance gain was obtained, so that they only included features based on unigrams, bigrams, trigrams, sentiment and objective words.

On the contrary, Joshi and Penstein-Rosé [86], obtained a gain in term of performance by adding dependency relations features and word unigrams. In particular, they organized their relations as follows: Given a sentence, its dependency parse is a set of triples expressing a grammatical relation $\{rel_i, w_j, w_k\}$, where w_j is the head word in the dependency triple, and the word w_k is the modifier word. In order to train the classifier, they generated *lexicalized dependency relation features* of the form [RELATION-HEAD-MODIFIER] and they used them in a standard bag-of-words paradigm (binary or frequency based).

Mejova and Srinivasan [116] performed a comparison between several feature selection. In particular, they tested stemming, term frequency versus binary weighting, negation-enriched features, and n-grams or phrases. Finally, they based a feature selection step on frequency-based vocabulary trimming, PoS, and lexicon selection. The experiments were performed on different datasets (product and movie review) and they found that for large datasets, a classifier trained on a small number of features that are ranked high by mutual information (MI) outperforms the models trained on all the

[11] NLPWin is the NLP system created by Microsoft.

[12] Evaluation considers the adjective pair «good-bad». The «strong-weak» adjective pair defines the potency factor. Adjective pair «active-passive» defines the activity factor.

features.

Most of the methods aimed at SPC use weighted unigrams and n-grams as features. The weighting schemes are often borrowed from Information Retrieval (IR). Kim and colleagues [94] studied pros and cons of different combinations of term weighting schemes: (i) PRESENCE (binary indicator for presence); (ii) TF (term frequency); (iii) VS.TF (normalized TF as in vector space model (VSM)); (iv) BM25.TF (normalized TF as in BM25[13] [170]); (v) IDF (inverse document frequency); (vi) VS.IDF (normalized IDF as in VS); (vii) BM25.IDF (normalized IDF as in BM25). According with previous studies [145], PRESENCE seems to represent the best scheme for SPC. In addition, the authors found that the best weighting combination is BM25.TF·VS.IDF, but its improvement over PRESENCE is actually minor (about 1.5%).

Martineau and Finin [112] proposed a new weighting scheme namely the Delta TFIDF, aimed at boosting the importance of words that are unevenly distributed between the positive and negative classes and discounts evenly distributed words. Their Delta TFIDF is obtained by contrasting the TFIDF of a given term t (e. g. a single word or a n-gram) against positive and negative training corpora. In particular, the feature value (Vt, d) for a term/word t in a document d is the difference of the TFIDF scores in positive and negative portions of the training set.

Paltoglou and Thelwall [142] compared different weighting schemes including TF and IDF variants in the SMART system [181], the variants in [170], their SMART Delta TFIDF versions, and their BM25 Delta TFIDF which are based on [112] but included smoothing. The comparison of these schemes on different corpora proved that Delta versions with smoothing performed significantly better than other variants of TFIDF.

Semi-Supervised Learning

Other systems exploit sentiment lexicons and self-training methods to infer the polatity of a given text. Qiu and colleagues [165] proposed an algorithm that works as follows: (i) the system exploits a lexicon-based iterative method, in which some reviews are classified on the basis of the sentiment lexicon, and then more reviews are classified through an iterative process by manipulating the ratio between positive and negative scores; (ii) a supervised classifier is trained on the reviews classified in the first phase and is then applied to other reviews. In practice, the authors exploited supervised learning techniques despite not having annotated data.

Similarly, Dasgupta and Ng [39] proposed a two-step algorithm based on the intuition that some texts (i. e. the unambiguous ones) could be easily classified as positive or negative on the basis of a sentiment lexicon. For this reason, they projected an algorithm able to exploit the unambiguous texts to classify the others. Their algorithm works as follows: (i) The system divides ambiguous from unambiguous texts and labels them with their polarity; (ii) The labeled data are used for the annotation of ambiguous texts. This approach allowed the authors to exploit supervised algorithms to manage unlabeled data.

Along these lines Zhou and colleagues [225] proposed a semi-supervised learning algorithm namely Active Deep Networks (ADN) to address SPC. In particular, they choose the proper training data to be labeled manually, and exploited the embedding

[13]BM25 is a ranking bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document.

information from the large amount of unlabeled data. Their results show that the proposed method need fewer manual labeled reviews to reach a relatively high accuracy. Other works [74] based on semi-supervised learning exploit word vectors in order to improve the classification by capturing their latent aspects.

Scoring techniques

A final method to approach SPC consists in using scoring techniques. For example, Dave and colleagues [40] scored the vocabulary as follows: (i) each term (or n-gram) is attributed with a polarity depending on the probability of this word occurring in Positive or Negative reviews; (ii) the document is attributed with a global polarity score by summing up the scores of all terms and using the sign of the total to determine the class. In order to calculate the score, the authors computed the formula in 2.11.

$$Score(f_i) = \frac{P_r(f_i|C) - P_r(f_i|C')}{P_r(f_i|C) + P_r(f_i|C')} \tag{2.11}$$

where f_i is a feature (term), C is a class, C' is its complement $\neg C$, and $P_r(f_i|C)$ is the conditional probability of term f_i in class C, which is computed by considering the number of times that a term f_i occurs in class C divided by the total number of terms in the reviews of class C.

In addition, the authors tested several substitution strategies aimed at improving the generalization of the system. For example, they replaced product names with a token ($\_productname$), rare words with a token ($\_unique$), numeric tokens with the token $\_number$ and so on. This last characteristic of their work is very interesting because one of the most important limits of using sentiment lexicons for SPC is to guarantee that such lexicons have a high coverage on the domain to which the text belongs to. By introducing this method, Dave and colleagues incorporated low-frequency terms in a generalized way instead of cutting out them.

Brief Comparison of the Approaches

Table 2.2 reports accuracies obtained using several approaches. Most of the works have been tested on movie reviews (MR) and customer reviews (CR), using supervised learning (SL) and feature engineering (FE) or Unsupervised Learning (UL). Only the best reported accuracy for each approach is presented. When evaluation was done for several datasets, we selected the results on the well known movie review dataset by Pang et al. (2002) [145]. The list is not exhaustive and because of differences in training/testing data, the results are not directly comparable. It is produced here only for reference.

2.4.1 SPC in Micro-blogging

In recent years, the huge amount of information from online micro-blogging platforms such as Twitter, attracted many researchers both for English [130] and Italian [14]. For this reason it is becoming very common to organize shared tasks in which several research units can participate with their own method to label the same annotated data. If we consider the SemEval-2015 shared task on English [171], the results for message-

AUTHOR	METHOD	YEAR	DT	DD	ACC.
Turney [205]	UL	2002	CR	410 reviews	0,74
Dave et al [40]	SL – FE	2003	CR	448 reviews	0,85
Gamon [63]	SL – FE	2004	CR	40,884 reviews	0,67
Mullen and Collier [126]	SL – FE	2004	MR	1,380 reviews	0,85
Pang and Lee [143]	SL – FE	2004	MR	2,000 reviews (balanced)	0,86
Ng et al. [133]	SL – FE	2006	MR	2,000 reviews (balanced)	0,88
Joshi and Penstein Rose [86]	SL – FE	2009	CR	2,200 review sentences	0,68
Kim et al. [94]	SL – FE	2009	MR	2,000 reviews (balanced)	0,85
Martineau and Finin [112]	SL – FE	2009	MR	2,000 reviews (balanced)	0,88
Qiu et al. [165]	SL – FE	2009	CR	248,535 sentences	0,7
Dasgupta and Ng [39]	SL – FE	2009	MR	2,000 reviews (balanced)	0,86
Paltoglou and Thelwall [142]	SL – FE	2010	MR	2,000 reviews (balanced)	0,88
Zhou et al [225]	SL	2010	MR	2,000 reviews (balanced)	0,77
Mejova-Srinivasan [116]	SL – FE	2011	MCR	2,000 reviews (balanced)	0,79
Taboada et al. [196]	UL	2011	CR	400 reviews	0,8

Table 2.2: *Accuracy obtained using several approaches. From left to right, the table presents author, method, year, dataset type, dataset dimension and accuracy.*

level polarity show the following models at the top, with a macro-averaged F1-score ranging from 64.3 to 64.8.

1. The best model [68], with a F1-score of 64.84 is the ensemble system that re-implemented and combined three state of the art models presented in a previous edition of SemEval;

2. The second model [184] with a F1-score of 64.59, was based on deep learning. The approach includes the following steps: (i) word embeddings are initialized using a neural language model [117] which is trained on a large unsupervised collection of tweets; (ii) a convolutional neural network is used to refine the embeddings on a large distant supervised corpus [64]; (iii) the word embeddings and other parameters of the network obtained at the previous stage are used to initialize the network that is then trained on the supervised corpus from Semeval-2015.

3. The third model [69] with a F1-score of 64.27 exploited a Logistic Regression classifier trained on several groups of features including lexical, syntactic, lexicon based, Z score and semantic features. In addition, the authors applied several weighting schemes for positive and negative labels in order to take into account the unbalanced distribution of tweets between the positive and negative classes.

If we consider the same task on 2016 [129], we notice that out of the 10 top-ranked teams, 5 teams used deep Neural Networks of some sort, and 7 teams used either general purpose or task-specific word embeddings, generated via word2vec [117] or GloVe [154], with a global performance between 0.59 and 0.63 [129].

Switching to Italian language, the first shared task on Sentiment Analysis has been organized in 2014 with SENTIPOLC 2014 [14]. The task was based on SemEval [130] and designed for Italian, which was an under-resourced language in the field of Sentiment Analysis [15, 24]. SENTIPOLC 2014 [14] shows the following top-3 models:

1. Basile and Novielli [13] proposed a supervised approach based on keywords, lexicons and micro-blogging features as well as the representation of the tweets in a word space, and achieved F1-score of 0.68;
2. The second model made by Hernandez-Farias et al. [59] proposed a two-step supervised learning approach. In particular, they trained a first classifier to decide if a tweet was subjective or objective. The result of the subjectivity step was then passed as a feature to the second classifier with the purpose of assigning the polarity label. The authors faced the lack of resources in Italian by translating (mostly automatically) existing English lexical resources.
3. The third model [149], based on supervised learning with lexical and unlexical features is described in detail in section 5.

The limited availability of sentiment resources for languages other than English forced researchers to transfer them from English [59], in spite of the connotation of lexical items is a cultural phenomenon that may vary greatly between different languages and time spans [38]. The organization of the shared task allowed for the development of a standard sentiment corpus for Italian that promoted the research in this field [8,14].

2.5 Classifying Emotions

One of the most important affective classes is **emotion**. Emotion detection is aimed at inferring the emotions influencing the author of a text. As in other sentiment analysis tasks, the intuition is that if a person is happy, he tends to use positive words. Likewise, if a person is angry or sad, the kind of words he uses express negative emotions.

This section investigates state-of-the-art methods for emotion detection, focusing on the existing emotion taxonomies, the algorithms aimed at the classification of texts and the available emotive lexical resources.

2.5.1 Taxonomies from Different Theorists

It is very difficult to define what is an emotion, and there is no universally accepted definition. The word *emotion* is used to refer to both the inner feelings of a person and to their outer displays. In affective science, different words are used to distinguish different, related phenomena [66]:

– AFFECT is a wide umbrella word, often used to encompass the different phenomena concerning valenced internal states (i.e. positive and negative);

– ATTITUDES are the most stable beliefs held by an individual about the valence of things.

– MOODS are passing and more long-term states, often not directly or simply related to a specific cause.

– EMOTIONS are the most short-lived reactions and responses to events and situations, reflecting the current goals, attitudes and mood of the individual, and they work to appraise the situation. Emotions can further be explained as the conscious feeling, the behaviour the emotion causes, and its physiological manifestation.

Several psychological studies argue that there are some universalities in emotions, even if language and culture can shape emotions and the way in how they manifest themselves, for example in facial expression [175].

It is possible to identify two broad classes of influencing theories for developing models of emotions being the *dimensional models* (any emotion can be described in terms of dimensions) and *basic emotions* (there are basic emotions that serve evolutionary functions, have a neurological basis and that cannot be reduced to more primitive components).

One of the most important theorists in the class of the *dimensional models*, is Russell, who developed a circumflex model of affect, in which affective states emerge from cognitive interpretations of core neural sensations that are the product of two independent neuro-physiological systems. In his works [174, 176], he plotted constructs of emotions against two dimensions being activation (the degree of intensity or arousal of an emotion) and pleasantness (the valence - how positive or negative the emotion is).

Figure 2.3 shows the Russell's circumplex model, consisting of six emotional dimensions organized into a circle, corresponding to the 2-dimensional emotion scale of arousal and valence.

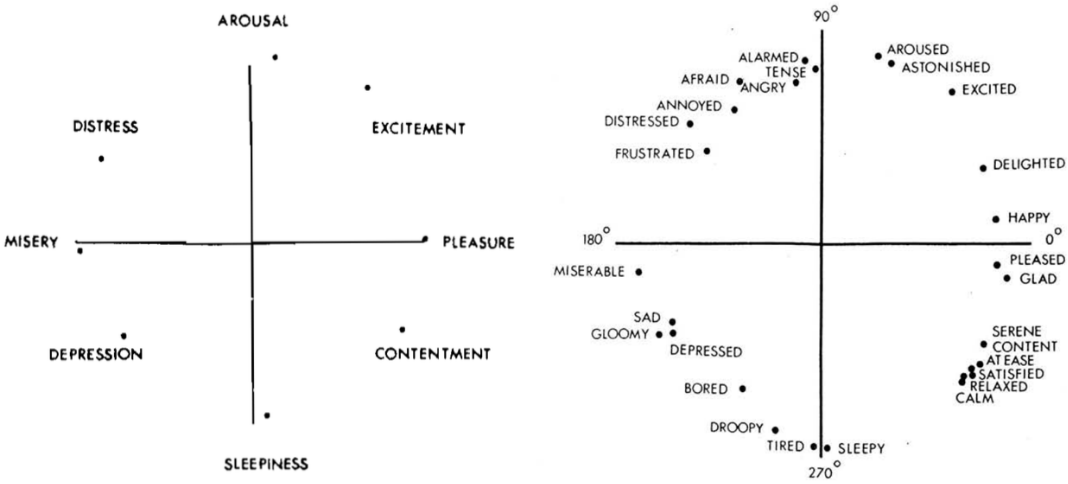

Figure 2.3: *Russell's circumplex model. On the left: Russell's model. On the right: emotion words mapped onto the circle. Image from J. A. Russell [174].*

Consecutive refinements of the model include three dimensions which are affiliation (communion, warmth, evaluation, etc.), power (agency, dominance, competence, potency, etc.), and activation (arousal, action readiness, affect intensity).

Watson and Tellegan [214] proposed four dimensions: pleasantness, engagement, positive and negative affect.

On the other hand, a pioneering work in the group of the theories concerning the *basic emotions* is the Charles Darwin's book "The Expressions of the Emotions in Man and Animals". Studies by evolutionary biologists [37] and psychologists show that emotions have evolved to improve the reproductive fitness for a species, as they are triggers for behavior with high survival value. The more complex brains of primates and humans are capable of experiencing not just the basic emotions such as FEAR and

JOY, but also more nuanced emotions such as REMORSE and SERENITY.

William James [83] was among the first scientist to approach a psychophysiological theory of emotions. Emotions produce in fact physical sensations which are considerably different for the various emotions. James proposed a somatic theory of emotions, stating that «the bodily changes follow directly the PERCEPTION of the exciting fact, and that our feeling of the same changes as they occur IS the emotion».

While in general it is not possible to observe clear boundaries between the emotions, and emotions don't usually occur in isolation, during the years psychologists have proposed a number of theories to classify human emotions into taxonomies in which some of them are considered basic, whereas others are considered complex.

In psychology, the beginning of the eighties of the twentieth century marked the beginning of a fertile period for a theoretical discussion about the emotions: Zajonc [223] have classified emotions into those that we can perceive (instinctual), and those that we reach after thinking and reasoning (cognitive). However, others do not agree with such a distinction and argue that emotions do not precede cognition [102, 103].

Plutchik [160] states that this debate may not be resolvable because it does not lend itself to empirical proof and the problem is a matter of definition.

Whatever, it seems to be a high correlation between basic and instinctual emotions, as well as between complex and cognitive emotions. Many of the basic emotions are also instinctual. A number of theories have been proposed on which emotions are basic [52, 83, 147, 160].

The longstanding dispute concerning whether emotions are better conceptualized in a discrete taxonomy, produced among years a long list of different theories including from two to twenty basic human emotions.

Ortony and Turner [139] make a detailed review of many of these models. Table 2.3 adapts the schema in [139] concerning the different taxonomies for basic emotions proposed by several theorists.

One of the most influencing taxonomies in the fields of Computational Linguistics and Facial Recognition has been proposed by Ekman and colleagues [52]. The taxonomy contains six basic emotions which are HAPPINESS, SADNESS, FEAR, DISGUST, SURPRISE and ANGER. Figure 2.4 shows the facial expressions connected to each emotion.

According to Darwin's observations [37], the Ekman's theory has evolved in humans from animals. These emotions are in fact referred as *universal* as they were found to be universal across human ethnicities and cultures.

Plutchik [159–162] proposed a theory with eight basic emotions which include the Ekman's six emotions as well as TRUST and ANTICIPATION. Plutchik organized the emotions in a wheel, in which the radius indicates intensity: the closer to the center, the higher the intensity (Figure 2.5). In addition, Plutchik's hypotesis was that the eight basic emotions form four opposing pairs: (i) JOY versus SADNESS; (ii) TRUST versus DISGUST; (iii) FEAR versus ANGER; (iv) ANTICIPATION versus SURPRISE.

This kind of emotion opposition is displayed in Figure 2.5 by the spatial opposition. In the Plutchik's model, certain emotions (primary dyads, in the white spaces between the basic emotions), which can be thought as combinations of the adjoining emotions, so that, combining the basic emotions each other, we can obtain several *second order emotions* such as LOVE and AGGRESSIVENESS.

Source	Emotion classes
Arnold (1960) [3]	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman et al. (1982) [52]	Anger, disgust, fear, joy, sadness, surprise
Gray (1982) [65]	Anxiety, joy, rage, terror
Izard (1971) [79]	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James (1884) [83]	Fear, grief, love, rage
McDougall (1926) [115]	Anger, disgust, elation, fear, subjection, tender emotion, wonder
Mowrer (1960) [125]	Pain, pleasure
Oatley and Jobnson-Laird (1987) [137]	Anger, disgust, anxiety, happiness, sadness
Panksepp (1982) [146]	Expectancy, fear, rage, panic
Plutchik (1980) [159]	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins (1984) [53]	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson (1930) [215]	Fear, love, rage
Weiner and Graham (1984) [216]	Happiness, sadness
Parrott (2001) [147]	Anger, fear, joy, love, sadness, surprise

Table 2.3: *Emotion classes from different theorists.*

Apart from the basic emotions, Parrott [147] proposes secondary and tertiary emotions, useful if the set of basic emotions is not considered fine-grained enough. Although the words describe different emotions or simply states of mind, they can also be included in emotion lexicons to spot different kinds of emotions [109]. This taxonomy could be a starting point for the construction of an emotive lexicon, for example after a significant expansion to include synonymous words, phrases and the top nearest neighbors of each emotion. Table 2.4 shows primary, secondary and tertiary emotions proposed by Parrott [147].

Similarly, the Human-Machine Interaction Network on Emotion [78] identified 48 emotions and classified them by taking into account different kinds of positive (and negative) orientations as well as the emotion intensity. Unfortunately, not even this taxonomy is balanced by respect to positive and negative emotions and some emotions do not have positive or negative orientations (i. e. SURPRISE and INTEREST) and this is why some psychologists observed that these are not emotions [139] simply because they do not have positive or negative orientations or valences.

2.6 Available Emotion Lexicons

Compared to sentiment lexicons, far less emotive lexicons have been produced, and most of them have lower coverage.

WordNetAffect: the resource [194] is one of the most used both in Italian and English. It contains information about the emotions that the words convey. Emotion labels are taken from Ekman's taxonomy [52] and they refer to WordNet

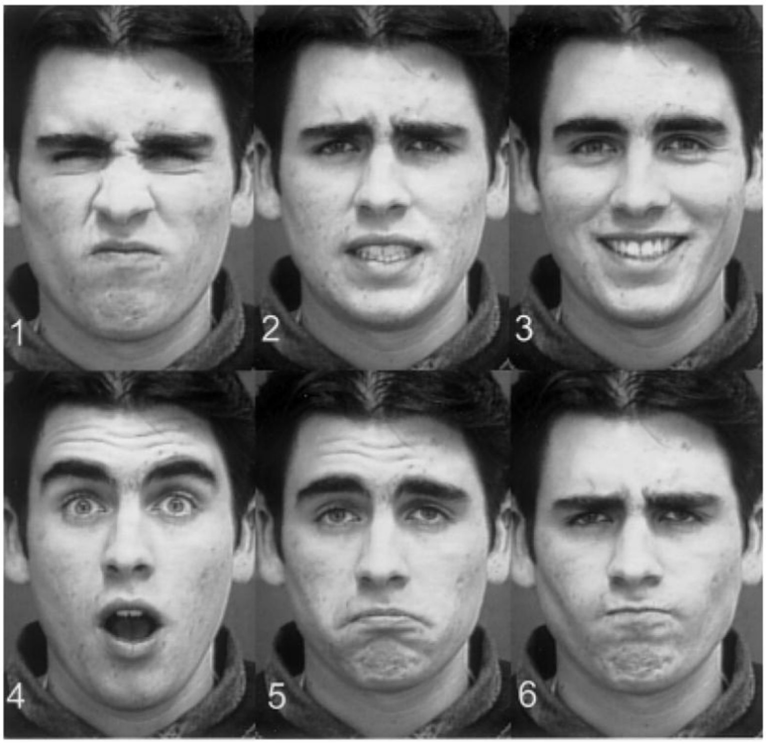

Figure 2.4: *Facial expressions representing the big six basic emotions: disgust, fear, joy, surprise, sadness and anger. Image from Schmidt and Cohn [183].*

synsets. The resource provides 900 annotated synsets and 1.6k words in the form $lemma\#PoS\#sense$, corresponding to roughly 1 thousand $lemma\#PoS$.

Affect database: is an extension of SentiFul [131] and contains 2.5K words in the form $lemma\#PoS$. The resource provides words annotated also with emotion scores rather than only with label.

AffectNet: is part of the SenticNet project [29]. It contains 10k words taken from ConceptNet [110] and it is aligned with WordNetAffect. This resource extends WordNetAffect labels to the concepts expressed with multi-words expressions (MWE).

NRC Word-Emotion Association Lexicon (EmoLex): is a lexicon [123] of 14,000 English words associated with the eight basic emotions proposed by Plutchick [159] and two sentiments (negative and positive), compiled by exploiting crowdsourcing and manual annotation.

Hashtag Emotion Lexicon: is a lexicon created by Mohammad and Kiritchenko [121] by exploiting twitter data and hashtags. They experimented fine-grained emotion sets (up to almost 600 emotion labels), to create the lexicon which consists of 11,418 word types annotated for the Ekman's six basic emotions [52].

2.7 Datasets annotated for emotions

In order to develop and evaluate emotion detection systems, the three commonly used datasets are:

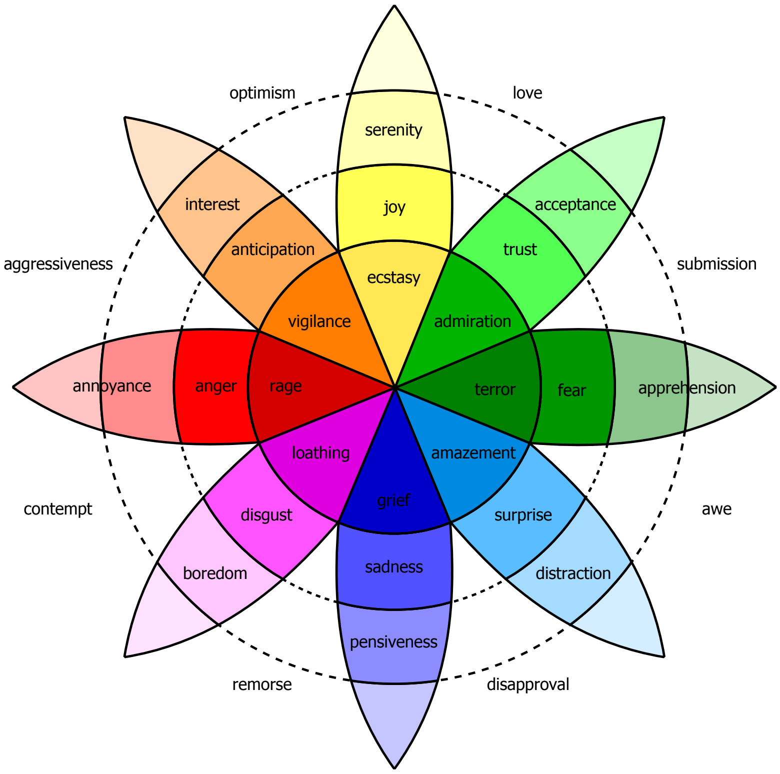

Figure 2.5: *Plutchik's wheel of emotions.*

(i) Affective Text dataset; (ii) Fairy Tales dataset; (iii) ISEAR.

Affective Text dataset: Task 14 at SemEval 2007 [192] was concerned with the classification of emotions and valence in news headlines. The headlines where collected from Google news, The New York Times, BBC News and CNN. The selected taxonomy was in line with the Ekman's standard model [52]. Valence was to be determined as positive or negative. Classification of emotion and valence were treated as separate tasks. Emotion labels were not considered as mututally exclusive, and each emotion was assigned a score from 0 to 100. Training/developing data amounted to 250 annotated headlines, while systems were evaluated on another 1000. Evaluation was done using two different methods: a fine-grained evaluation using Pearson's r to measure the correlation between the system scores and the gold standard. With a coarse-grained method, each emotion score was converted to a binary label, and precision, recall, and f-score were used for evaluation.

Fairy Tales dataset: The dataset was collected by Alm [2] and consists of about 1,000 sentences from fairy tales by B. Potter, H.C. Andersen and Grimm. The sentences were annotated with the same six emotions of the Affective Text dataset, though with different names: Angry, Disgusted, Fearful, Happy, Sad, and Surprised.

ISEAR: The International Survey on Emotion Antecedents and Reactions (ISEAR) is a dataset created for in the context of a psychology project of the 1990s, by collecting questionnaires answered by people with different cultural backgrounds.

Primary	Secondary	Tertiary emotions
Love	Affection	Adoration, affection, love, fondness, liking, attraction, caring, tenderness, compassion, sentimentality
	Lust	Arousal, desire, lust, passion, infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, bliss, cheerfulness, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
	Zest	Enthusiasm, zeal, zest, excitement, thrill, exhilaration
	Contentment	Contentment, pleasure
	Pride	Pride, triumph
	Optimism	Eagerness, hope, optimism
	Enthrallment	Enthrallment, rapture
	Relief	Relief
Surprise	Surprise	Amazement, surprise, astonishment
Anger	Irritation	Aggravation, irritation, agitation, annoyance, grouchiness, grumpiness
	Exasperation	Exasperation, frustration
	Rage	Anger, rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, loathing, scorn, spite, vengefulness, dislike, resentment
	Disgust	Disgust, revulsion, contempt
	Envy	Envy, jealousy
	Torment	Torment
Sadness	Suffering	Agony, suffering, hurt, anguish
	Sadness	Depression, despair, hopelessness, gloom, glumness, sadness, unhappiness, grief, sorrow, woe, misery, melancholy
	Disappointment	Dismay, disappointment, displeasure
	Shame	Guilt, shame, regret, remorse
	Neglect	Alienation, isolation, neglect, loneliness, rejection, homesickness, defeat, dejection, insecurity, embarrassment, humiliation, insult
	Sympathy	Pity, sympathy
Fear	Horror	Alarm, shock, fear, fright, horror, terror, panic, hysteria, mortification
	Nervousness	Anxiety, nervousness, tenseness, uneasiness, apprehension, worry, distress, dread

Table 2.4: *Inventory of emotions in Parrott [147].*

Students were asked to report situations in which they had experienced all of 7 major emotions (joy, fear, anger, sadness, disgust, shame and guilt). The dataset contains reports on seven emotions, by approximately 3000 respondents from all over the world. In total there are 7665 sentences labelled with an emotion.

2.8 Emotion Classification of documents

Emotion detection investigates how to classify an opinionated piece of text into a set of emotion labels such as ANGER and JOY. This kind of application is useful for example in the fields of social sciences and economics to recognize (and measure) the emotional

Negative & forceful	Positive & lively
Anger	Amusement
Annoyance	Delight
Contempt	Elation
Disgust	Excitement
Irritation	Happiness
	Joy
Negative & not in control	Pleasure
Anxiety	
Embarrassment	**Caring**
Fear	Affection
Helplessness	Empathy
Powerlessness	Friendliness
Worry	Love
Negative thoughts	**Positive thoughts**
Doubt	Courage
Envy	Hope
Frustration	Pride
Guilt	Satisfaction
Shame	Trust
Negative & passive	**Quiet positive**
Boredom	Calm
Despair	Content
Disappointment	Relaxed
Hurt	Relieved
Sadness	Serene
Agitation	**Reactive**
Shock	Interest
Stress	Politeness
Tension	Surprise

Table 2.5: *Inventory of emotions in HUMAINE [78].*

changes of population in large scale [22, 45].

Most of the studies on emotions are focused on emotion lexicon constructions, such as WordNet-Affect [194], which creates a mapping between emotive words to the Wordnet concepts, and the lexicon built by Mohammad and Turney [122] by exploiting crowdsourcing techniques.

A number of techniques for text emotion classification aim to focus on the writer emotional state, in other words they study the text from the perspective of the author. If we consider the studies on English, a number of emotion classification researches focused for example on blog postings [93, 118, 221]. The LiveJournal blog corpus is very useful for the evaluation, because it offers a text classified according to the mood reported by its author during composition and it is tagged with 132 emotion categories.

Mishne [118] implemented binary SVM classifiers for mood classification of the top forty frequent moods. Mishne and de Rijke [119] lift this work to the aggregate level, and use ML to estimate aggregate mood levels from the text of blog entries.

Wu and colleagues [221] represented the sentence emotional state as a sequence of semantic labels and attributes. In a first step they used an Apriori algorithm to derive the emotion association rules (EARs) for each emotion, and in a second one they developed a separable mixture model to calculate the similarity between the sentence and the EARs associated to each emotion.

A second approach to text emotion classification is to focus on the perspective of the reader (e.g. what are the emotions expressed through the text which the reader is able to capture?). Lin and collegues [106] classified news articles provided by Yahoo! Chinese news. Readers voted on the articles based on the readers' perceived emotions. The algorithm used supervised learning with an SVM trained on features such as: (i) all Chinese character bigrams; (ii) all words produced by a Chinese word segmentation tool; (iii) articles meta-data (e.g. news reporter, news category, location, publication time, and name of the news agency); (iv) emotion categories of words, obtained from an emotion lexicon previously constructed [222].

In order to deal with the data sparseness problem in sentence-level emotion classification, Tokushisa and colleagues [201] proposed a data driven method for inferring the emotion of an utterance sentence, consisting in collecting a huge corpus of emotion-provoking event instances from the web.

Strapparava and Mihalcea [193] constructed a data set of news titles annotated for emotions, and proposed a methodology for fine-grained and coarse-grained evaluations. They presented several algorithms, ranging from simple heuristics (e.g., directly checking specific affective lexicons) to more refined algorithms exploiting latent semantic representations of emotions.

2.9 Summary

The chapter presents a review of different state-of-the-art methods for sentiment analysis. In particular, it focuses on the existing lexicons (section 2.2) and on the techniques aimed at the creation of new sentiment lexicons (section 2.3) using several methods including rule based approaches, PMI-based approaches, Wordnet based approaches and other techniques relying to word embeddings or graph theory.

In addition, the chapter presents the methods used to classify texts on the basis of their polarity, focusing on the classical methods (mostly addressed to classify product and movie reviews) and to the methods aimed to classify micro-blogging texts such as tweets.

Finally, section 2.5 is dedicated to the classification of emotions, starting from the different theories and taxonomies to the emotion classification of documents 2.8.

CHAPTER 3

NLP tools

A crucial prerequisite to extract any kind of information from text is an accurate linguistic annotation, and this is true also for inferring its emotive connotation.

For the linguistic annotation of the resources described in this work we used two main resources namely T2K^2 and EXTra. The linguistic pipeline up to the dependency parsing has been carried out using the set of tools belonging to T2K^2, and in particular for sentence splitting, tokenization, lemmatization, part of speech tagging and dependency parsing.

In order to extract more complex linguistic items such as multi-words expressions and collocations, we used the tool EXTra, which is a software based on statistical analysis of words frequency and association, aimed at recognizing complex terms starting from patterns of parts of speech.

3.1 NLP Analyzer: T2K^2

T2K^2 is a set of linguistic tools described by Dell'Orletta and colleagues [42], which is an extension of the ontology learning system T2K (Text–to–Knowledge, [43, 104]). Linguistic pre–processing of texts is performed by a collection of annotation tools developed by the ItaliaNLP Lab and the Department of Computer Science of the University of Pisa. Each uploaded text is linguistically annotated at increasingly complex levels of analysis, represented by: (i) Sentence splitting; (ii) Tokenization; (iii) Lemmatization; (iv) Part–Of–Speech tagging; (v) Dependency parsing;

Sentence splitting is the task of dividing a string of text into sentences. In Italian the use of the punctuation (in particular of the full stop character) supplies a reasonable approximation of the problem, but the non trivial nature of the problem is due to the use of the full stop character for abbreviations, which may or may not also terminate a sentence. Naturally, for the languages that don't contain punctuation this task is exponentially more complicated.

Lemmatization is the process of grouping together different inflected forms of a word. Algorithmically, the process may involve complex tasks such as understanding context and determining the part of speech of the words in a sentence.

Part-of-speech tagging is the process of assigning unambiguous part-of-speech and possibly a set of morphosyntactic features to each token in a text. Tokens are normally words, but also include punctuation and numbers.

In T2K^2, morpho–syntactic tagging is carried out by the POS tagger developed by Dell'Orletta [41] and dependency parsed with the DeSR parser in [4]. In particular, we used the configuration that exploits Support Vector Machine and Multilayer Perceptron as learning algorithms. As reported in Dell'Orletta [41] and Attardi [5], both the instruments represent state–of–the–art tools for Italian and English.

The output of T2K^2 can be downloaded by the user in CoNLL format [135], where the sentences are separated by a blank line and each token starts on a new line. Each token is annotated with the following information: (i) lemma; (ii) coarse part of speech; (iii) fine grained part of speech; (iv) morphological features; (v) syntactic dependency. In Figure 3.2 all columns but the last one reflect the information expressed in the previous list. The last column contains information about the presence of Multi-Word Expressions (section 3.2).

3.2 Multi-Word Expressions

A general definition of term is «a surface representation of a specific domain concept» [82, 153]. In general, a term can be either a single word or a multiword expression (MWE). In this section the focus is on the latter kind of terms, under the assumption that MWE (i.e. complex terms) range from completely opaque idioms to semantically compositional word combinations [57].

The Principle of Compositionality states that the meaning of an expression is determined by the meaning of its constituents and by its grammatical structure [200]. By extension, sentiment composition is the determining of sentiment of a phrase, based on its constituents [97]. Multiword terms, in fact, are less ambiguous and less polysemous than single word terms, yielding a better representation of the document content. Moreover, the lion share of domain concepts are normally expressed through multiword terms, which represent a crucial component of natural language lexicons [81].

The recognition of complex terms from texts is performed on the basis of different criteria. Major differences exist between algorithms that take into account only the distributional properties of terms, such as frequency and $TF - IDF$ [182], and those using contextual information such as syntactic, terminological and semantic features [23, 42, 62, 113]. The common trait of most of the strategies above is the identification of a set of ranked candidates from texts, and then the application of a filtering function to separate real terms from non-terms. In this latter phase, the candidates are usually sorted according to their association strength as an estimate of their degree of termhood.

The term extractor EXTra [150] takes into account the linguistic structure of multiword terms by implementing a candidate selection step that uses manually-defined structured PoS-patterns. Moreover, in order to tackle the complexity of term phrases, EXTra adopts a new association measure that promotes terms composed by one or more sub-terms. The intuition is that the degree of termhood of a candidate pattern is a function of the statistical distribution of its parts, and of the presence of highly weighted sub-terms. The last step of EXTra applies a filtering function to separate real terms from wrong candidates.

EXTra includes several parameters that allow the user to optimize the extracted terms with respect to the target corpus and domain. In particular, the user can specify the set of structured patterns that guide the extraction process, a list of stopwords, and the thresholds for the association measure and the n-gram frequency. In the configuration file, the user also selects the association measure used by the weighting algorithm, which can be selected between the Pointwise Mutual Information [31], the Local Mutual Information [57], the Log Likelihood Ratio [50], and an identity function weighting the n-grams with their raw frequency. In order to assure the flexibility of the system, a further parameter affects the importance given to long terms by the weighting algorithm.

The input of EXTRA is a PoS-tagged and lemmatized text in a tab-delimited CONLL [135] format and its output consists of two files: the CONLL enriched with the extracted multiword terms, and a list of terms ranked according to their termhood. EXTra works in three main steps: (i) Candidate selection; (ii) Weighting; (iii) Filtering.

Candidate selection. Candidate terms are identified using manually-defined *structured PoS patterns* that represent the recursive phrase structure of terms. A structured PoS pattern is a bracketed list of constituents, where each constituent can be either a sequence of two content PoS or another bracketed constituent. This structure defines long term patterns as a composition of smaller patterns[1]. The following is an example of structured PoS pattern:

Example 1.

$$[[noun(-s), prep(-e),\ \ noun(-s)], prep(-ea),\ \ [noun(-s), adjective(-a)]]$$

The pattern is composed by two constituents:

(i) $[noun(-s), preposition(-e), noun(-s)]$;

(ii) $[noun(-s), adjective(-a)]$.

This structured pattern identifies the candidate *Politica di sviluppo delle Risorse Umane* “human resource development policy”.

Following the pattern structure and ignoring prepositions, we can isolate two embedded sub-terms $[politica{-}s\ \ di{-}e\ \ sviluppo{-}s]$ and $[risorse{-}s\ \ umane{-}a]$. From a computational point of view, during the candidate selection phase, EXTra first stores the statistical information of each sub-pattern (e.g., the frequencies of the embedded pairs and then registers the frequency of the aggregate pair $(Politica\_di\_Sviluppo$, $Risorse\_Umane)$

[1]The content PoS are specified in the configuration file. In the configuration, the user is also able to exclude from the termhood computation particular classes of PoS such as articles and prepositions.

Weighting The structure of the PoS patterns is also used to guide the process of statistical term weighting by following the same order of incremental composition. Following a recursive structure, the weighting algorithm assigns a termhood score to each of the embedded phrases, and then computes the global score for the complex term by combining the partial weights of its constituents.
EXTra's term weighting algorithm is applied recursively to the internal structure of the patterns: At the base step it measures the association strength σ of each candidate two-word term (w_1, w_2) by computing standard association measures, such as the PMI. The candidates whose score σ is above an empirically fixed threshold are added to the set of the terms $T = \{t1, ..., tn\}$.

In the recursive step, EXTra measures the association strength σ of any n-word candidate term (c_1, c_2) by combining the association strengths of its sub-elements. The termhood of a candidate is calculated using the formula in 3.1, where $S(c_i)$ follows Equation 3.2 if $c_i \in T$ and it follows Equation 3.3 otherwise.

$$\sigma(c_1, c_2) = S(c_1) \cdot S(c_2) \tag{3.1}$$

$$S(c_i) = \frac{log_2\sigma(c_i)}{k} \tag{3.2}$$

$$S(c_i) = 1 \tag{3.3}$$

The weighting scheme formalizes the assumption that the termhood of longer terms depends on the degree of termhood of their parts. The parameter k controls the contribution of sub-terms to the weight of longer terms: The smaller the k, the higher the weight assigned to longer terms containing them.

Coming back to Example 1, and supposing that the selected association measure is the PPMI, at the base step EXTra measures the scores for the pairs $(risorse - s, umane - a)$ and $(politica - s, sviluppo - s)$ and it stores their termhood value.

In the recursive step, the system calculates the score σ between the sub-candidates $politica\_di\_sviluppo$ and $risorse\_umane$ by applying Equation 3.1. Since both the sub-terms belong to the set of accepted terms T, their termhood score σ is calculated according to Equation 3.2.

Filtering Candidate multiword terms are filtered by using three main filters. First of all, an optional stoplist is used to exclude the terms containing one or more words in the blacklist during indexing. Then, patterns with a frequency below a frequency threshold are discarded before computing their strength of association.

Finally, the association measure filter defines the minimum strength of association that an n-gram must have to be considered as a multiword term: The candidates whose score σ is above that empirically fixed threshold are added to the set of terms T and discarded otherwise.

Figure 3.1 shows the top ranked terms with the relative association scores in the corpus FB-NEWS15, which will be described in section 4.2. Figure 3.2 shows an example of annotated text taken from the same corpus. From left to right, the columns are:

```
avere-v#ragione-s (avere ragione) 629212.44
rendere-v#conto-s (rendere conto) 335553.85
pagare-v#il-rd#tassa-s (pagare le tasse) 274965.32
avere-v#bisogno-s (avere bisogno) 227073.72
persona-s#di-ea#stesso-a#sesso-s (persone dello stesso sesso)184884.8
fare-v#schifo-s (fare schifo) 179046.52
scoperta-s#di-ea#acqua-s#caldo-a1 (scoperta dell'acqua calda) 78923.24
avere-v#paura-s (avere paura) 163754.21
chiedere-v#scusa-s (chiedere scusa) 147487.11
augurio-s#di-e#pronto-a#guarigione-s (auguri di pronta guarigione) 139083.79
leggere-v#articolo-s (leggere l'articolo) 136408.36
pezzo-s#di-e#merda-s (pezzo di merda) 134584.41
risolvere-v#problema-s (risolvere il problema) 129331.5
sapere-v#cosa-s (sapere le cose) 126072.5
avere-v#diritto-s (avere il diritto) 118700.71
forza-s#di-e#ordine-s (forze dell'ordine) 118495.87
biglietto-s#di-e#solo-a#andata-s (biglietto di sola andata) 105630.47
rompere-v#il-rd#coglione-s (rompere i coglioni) 103201.98
rompere-v#palla-s (rompere le palle) 96675.65
avere-v#il-rd#coraggio-s (avere il coraggio) 95611.72
```

Figure 3.1: *Top ranked terms using EXtra (Corpus: FB-NEWS; Association measure: LMI).*

(i) id of the token in the sentence (sentences are separated with a blank line); (ii) token; (iii) lemma; (iv) coarse part of speech; (v) fine grained part of speech; (vi) morphological features; (vii) syntactic dependency; (viii) MWE. The columns (i)-(vii) come from $T2K^2$ and MWEs from EXTra. MWEs are in the form $lemma_1 - pos\#lemma_2 - pos\#...\#lemma_n - pos$ and they are repeated for each token belonging to the MWE.

3.3 Summary

The chapter describes the tools used for the annotation of the corpora in the present work. Linguistic annotation up to the dependency parsing has been carried out using $T2K^2$ which is described in section 3.1. Multiword terms have been identified using the tool EXTra, which is a domain indipendent term extractor.

```
<doc user="<newspaper(string)>"  id="<id_post(string)>" type="comment"
parent_post="<id_post(string)>" parent_comment=""
date="AAAA-MM-DD HH:MM:SS" likes="1" comments="0">

1 Grazie grazie I I _ 5 mod _
2 a a E E _ 1 comp_loc _
3 Dio Dio S SP _ 2 prep _
4 gli il R RD num=p|gen=m 5 det _
5 scienziati scienziato S S num=p|gen=m 6 subj _
6 sono essere V V num=p|per=3|mod=i|ten=p 0 ROOT _
7 troppo troppo B B _ 8 mod _
8 occupati occupare V V num=p|mod=p|gen=m 6 pred _
9 a a E E _ 8 arg _
10 fare fare V V mod=f 9 prep fare-v#esperimento-s
11 esperimenti esperimento S S num=p|gen=m 10 obj fare-v#esperimento-s
12 nella in E EA num=s|gen=f 11 comp _
13 speranza speranza S S num=s|gen=f 12 prep _
14 di di E E _ 13 arg _
15 migliorare migliorare V V mod=f 14 prep migliorare-v#il-rd#vita-s
16 la il R RD num=s|gen=f 17 det migliorare-v#il-rd#vita-s
17 vita vita S S num=s|gen=f 15 obj migliorare-v#il-rd#vita-s
18 di di E E _ 17 comp _
19 tutti tutto D DI num=p|gen=m 20 mod _
20 voi voi P PE num=p|per=2|gen=n 18 prep _
21 per per E E _ 15 mod _
22 leggere leggere V V mod=f 21 prep _
23 questi questo D DD num=p|gen=m 24 mod _
24 commenti commento S S num=p|gen=m 22 obj _
25 . . F FS _ 6 punc _
_
1 Complimenti complimento S S num=p|gen=m 0 ROOT _
2 ai al E EA num=p|gen=m 1 comp _
3 vincitori vincitore S S num=p|gen=m 2 prep _
4 di di E E _ 3 comp _
5 quest'questo D DD num=s|gen=n 6 mod _
6 anno anno S S num=s|gen=m 4 prep _
7 che che P PR num=n|gen=n 14 subj _
8 ,,F FF _ 9 punc _
9 con con E E _ 14 comp _
10 le il R RD num=p|gen=f 12 det _
11 loro loro A AP num=n|gen=n 12 mod _
12 scoperte scoperta S S num=p|gen=f 9 prep _
13 ,,F FF _ 9 punc _
14 permettono permettere V V num=p|per=3|mod=i|ten=p 6 mod_rel _
15 di di E E _ 14 arg _
16 salvare salvare V V mod=f 15 prep salvare-v#il-rd#vita-s
17 la il R RD num=s|gen=f 18 det salvare-v#il-rd#vita-s
18 vita vita S S num=s|gen=f 16 obj salvare-v#il-rd#vita-s
19 vita vita S S num=s|gen=f 18 mod _
20 milioni milione S S num=p|gen=m 19 mod milione-s#di-e#persona-s
21 di di E E _ 18 comp milione-s#di-e#persona-s
22 persone persona S S num=p|gen=f 21 prep milione-s#di-e#persona-s
23 . . F FS _ 1 punc _

</doc>
```

Figure 3.2: *Example of annotated text taken from the corpus FB-NEWS15.*

CHAPTER 4

Creating emotive (lexical) resources

4.1 ItEM: A distributional emotive resource

With the proliferating use of social media, textual emotion analysis is becoming increasingly important. Emotion detection can be useful in several applications.

Emotion lexicons, in which lemmas are associated to the emotions they evoke, are knowledge sources that can help the development of detection algorithms and prediction systems. Almost all languages but English lack a high-coverage high-quality emotion inventory of this sort. The English language, in fact, has available many resources [29, 123, 194] compared to the resources available in the other languages, but the connotation is a cultural phenomenon that may vary greatly between languages and between different time spans [38], then the simple transfer from English to another language represents a temporary solution. In this context, WordNet Affect [194] represents an exception because it is aligned with Wordnet, so that it can be used as an emotive resource in several languages.

Building emotive resources is very costly and requires a lot of manual effort by human annotators. Crowdsourcing is usually able to speed the process and dramatically lower the cost of human annotation [128, 188].

Mohammad and Turney [122,123] show how the «wisdom of the crowds» can be effectively exploited to build a lexicon of emotion associations for more than 24,200 word senses. For the creation of their lexicon, EmoLex, they use the following resources: Macquarie Thesaurus [51], General Inquirer [191], WordNet Affect Lexicon [194] and Google n-gram corpus [26].

The terms selected from these resources have been manually annotated by means of a crowdsourcing experiment, thus obtaining, for every target term, an indication of its polarity and of its association with one of the eight Plutchik's basic emotions. The methodology proposed by Mohammad and Turney [122, 123], however, cannot be easily exported to languages where even small emotive lexicons are missing. Moreover, a potential problem of a lexicon built solely on crowdsourcing techniques is that its update requires a re-annotation process.

To bypass the problem, we propose an approach that jointly exploits corpus-based methods and human annotation. The output will be ItEM, a high-coverage emotion lexicon for Italian, in which each target term is provided with an association score with eight basic emotions (cf. below).

Given the way it is built, ItEM is not only a static lexicon, since it also provides a dynamic method to continuously update the emotion value of words, as well as to increment its coverage. The resource is comparable in size to EmoLex [123], with the following advantages:

– it requires a minimal use of external resources to collect the seed terms;

– little annotation work is required to build the lexicon;

– its update is mostly automatized.

Following the approach in Mohammad and Turney [122, 123], we borrow our emotions inventory from Plutchik [161], who distinguishes eight "basic" human emotions: JOY, SADNESS, FEAR, DISGUST, SURPRISE, ANGER, TRUST and ANTICIPATION.

Positive characteristics of this classification include the relative low number of distinctions encoded, as well as its being balanced with respect to positive and negative feelings. In addition, this taxonomy is a superset of the Ekman [52]'s one which is used by several systems oriented to Emotion Recognition.

An emotive lexicon implementing the Plutchik's taxonomy will encode words like *ridere* "laugh" or *festa* "celebration" as highly associated to JOY while words like *rain* "pioggia" or *povertà* "poverty" will be associated to SADNESS, and words like *rissa* "fight" or *tradimento* "betray" will be encoded as ANGER-evoking entries.

4.1.1 System overview

Distributional semantics is grounded on Harris's distributional hypothesis [71], which states that semantically similar words tend to appear in similar contexts. From a computational point of view, each word is represented by a weighted feature vector, where features correspond to other words that co-occur with the target word in the surrounding context [12, 207]. In order to build ItEM, we exploited the distributional hypothesis, which we have generalized to emotions:

> *A word $\langle w \rangle$ is strongly associated with an emotion $\langle e \rangle$ if it co-occurs in similar contexts of other words strongly associated with $\langle e \rangle$.*

In order to implement this hypothesis, we represented each emotion as a centroid vector built starting from a set of seed words strongly associated to the target emotion and we measured the paradigmatic similarity between the word and the emotion. Besides co-occurring in similar contexts, words with the same (or similar) emotive connotation also tend to occur together. For this reason, we have introduced a *syntagmatic boost* to promote the most informative contexts of each emotion.

Insert a ⟨**PoS**⟩ that you associate to the emotion ⟨**e**⟩*

1. ⟨PoS⟩ Lemma
2. ⟨PoS⟩ Lemma
3. ⟨PoS⟩ Lemma
4. ⟨PoS⟩ Lemma
5. ⟨PoS⟩ Lemma

Definition of the emotion ⟨**e**⟩.

Figure 4.1: *Form used to collect the seeds lemmas. Each form corresponds to the pair langlePoS, emotionrangle where a PoS could be a Noun, an Adjective or a Verb, and an emotion belongs to the scheme in Plutchik [162].*

ItEM has been built with a three stage process. The first one has been implemented using an online feature elicitation paradigm, and it is aimed at collecting and annotating a small set of emotional seed lemmas. A second phase, exploits distributional semantic methods to expand the seeds and populate the resource. Finally, the automatically extracted emotive annotations have been evaluated with crowdsourcing.

4.1.2 Seed selection and annotation

The goal of the first phase is to collect a small lexicon of *emotive lemmas*, highly associated to one or more Plutchik's basic emotions. To address this issue, we used an online feature elicitation paradigm, in which 60 Italian native speakers were asked to list, for each of our eight basic emotions, 5 lemmas for each of our PoS of interest (Nouns, Adjectives and Verbs). Figure 4.1 shows the form used to collect the seed lemmas[1].

In this way, we collected a lexicon of 347 lemmas strongly associated with one or more Plutchik's emotions. For each lemma, we calculated its emotion distinctiveness as the production frequency of the lemma (i.e. the numbers of subjects that produced it) divided by the number of the emotions for which the lemma was generated. In order to select the best set of seed to use in the bootstrapping step, we only selected from ItEM the terms evoked by a single emotion, having a distinctiveness score equal to 1.

In addition, we expanded this set of seeds with the names of the emotions such as the nouns *gioia* "joy", *rabbia* "anger" and *fiducia* "trust" and their synonyms attested in Multiwordnet [155] and Treccani Online Dictionary[2].

Globally, we selected 555 emotive seeds, whose distribution towards emotion and PoS is described in Table 4.1. Table 4.2 shows the most frequent (i. e. the number of elicitations) adjectives for each emotion in Plutchik [161].

[1] The original form was in the Italian language.
[2] http://www.treccani.it/vocabolario/.

EMOTION	N. OF SEEDS	ADJ	NOUNS	VERBS
JOY	61	19	26	19
ANGER	77	32	30	16
SURPRISE	60	25	17	22
DISGUST	80	40	21	25
FEAR	78	37	20	27
SADNESS	77	39	22	26
TRUST	62	25	21	17
ANTICIPATION	60	15	22	23

Table 4.1: *Distribution of the seeds lemmas.*

EMOTION	ADJ	VERBS	NOUNS
JOY	*allegro* "joyous"	*cantare* "sing"	*soddisfazione* "pleasure"
	spensierato "cheerful"	*abbracciare* "hug"	*mare* "sea"
	appagato "satisfied"	*condividere* "share"	*sole* "sun"
ANGER	*furioso* "furious"	*picchiare* "beat up"	*violenza* "aggressiveness"
	violento "aggressive"	*gridare* "shout"	*odio* "hate"
	nervoso "irritable"	*aggredire* "attack"	*collera* "anger"
SURPRISE	*inaspettato* "unespected"	*stupire* "amaze"	*regalo* "present"
	sorprendente "surprising"	*improvvisare* "extemporize"	*festa* "celebration"
	piacevole "pleasant"	*regalare* "donate"	*gioia* "joy"
DISGUST	*schifoso* "ripugnant"	*vomitare* "throw up"	*vomito* "puke"
	marcio "rotten"	*rifiutare* "refuse"	*schifo* "revulsion"
	nauseante "nauseating"	*allontanare* "distance"	*nausea* "aversion"
FEAR	*oscuro* "dark"	*tremare* "shiver"	*terrore* "terror"
	spaventoso "frightful"	*scappare* "run away"	*brivido* "creeps"
	terribile "terrible"	*affrontare* "deal with"	*panico* "panic"
SADNESS	*oscuro* "dark"	*disperarsi* "be depressed"	*pianto* "crying"
	noioso "boring"	*isolarsi* "isolate"	*amarezza* "bitterness"
	deprimente "depressing"	*cadere* "fall apart"	*infelicità* "unhappiness"
TRUST	*affidabile* "reliable"	*credere* "believe"	*speranza* "chance"
	sicuro "sure"	*affidare* "entrust"	*sicurezza* "safety"
	amichevole "friendly"	*dare* "give"	*ottimismo* "optimism"
ANTICIPATION	*fiducioso* "confident"	*immaginare* "image"	*fiducia* "trust"
	trepidante "anxious"	*illudersi* "deceive"	*desiderio* "whish"
	incerto "uncertain"	*pazientare* "wait patiently"	*prospettiva* "perspective"

Table 4.2: *Top 3 nouns, verbs and adjectives produced by the speakers for each of the emotion categories.*

4.1.3 Bootstrapping

In this phase, we exploited distributional semantic methods to expand the seeds collected in the first phase and populate ItEM.

The seed lemmas collected in the first phase have been used to bootstrap ItEM using a corpus-based model inspired to Turney and Littmann [206] to automatically infer the semantic orientation of a word from its distributional similarity with a set of positive and negative words.

Even if we employ a larger number of emotion classes, our model is based on the

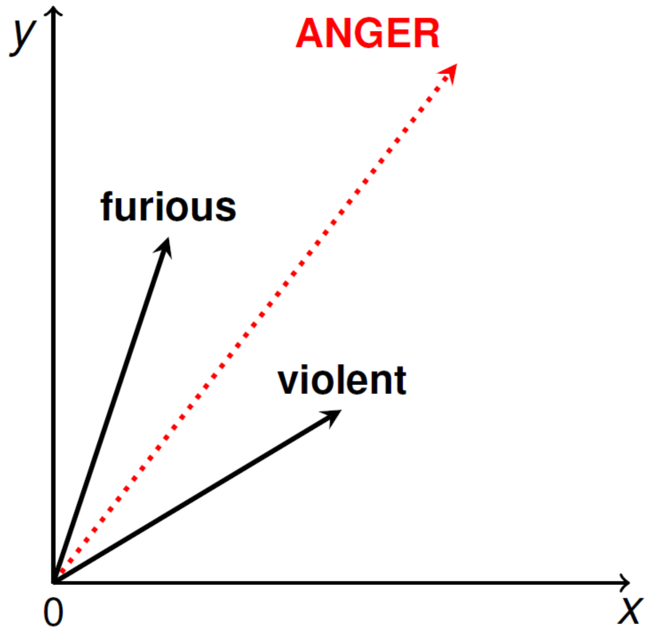

Figure 4.2: *Cosine similarity between two target words and the emotive centroid "*ANGER*-adjectives".*

same assumption that, in a vector space model [179, 207], words tend to share the same connotation of their neighbors.

In a first version of the system, we extracted from La Repubblica corpus [10] and itWaC [11], the list of the 30,000 most frequent nouns, verbs and adjectives, which were used as targets and contexts in a co-occurrence matrix collected using a five-word window centered on the target lemma. Differently from Turney and Littmann [206], however, we did not calculate our scores by computing the similarity of each new vector against the whole sets of seed terms. On the contrary, for each ⟨emotion, PoS⟩ pair we built a centroid vector from the vectors of the seeds belonging to that emotion and PoS, obtaining in total 24 centroids.

We re-weighted the co-occurrence matrix using the Pointwise Mutual Information [31] as in Equation 2.4 on page 15, and in particular the Positive PMI (PPMI), in which negative scores are changed to zero [136].

To optimize the vector space, we followed the approach in Polajnar and Clark [163] and we selected the top 240 contexts for each target word. As a last step, we applied singular value decomposition (SVD), to reduce the matrix to 300 dimensions.

The VSM allowed us to calculate our emotive scores by measuring the cosine similarity between the target lemmas and the centroid vectors: depending on the PoS of the target lemma, we measured the cosine similarity between the lemma and the eight emotive centroids corresponding to the target PoS (e.g. the centroid of "ANGER-adjectives" in Figure 4.2). The output of this stage is a list of words ranked according to their emotive score.

Tables 4.3-4.5 show respectively the top most associated adjectives verbs and nouns for each emotion, with their association scores, calculated as the cosine similarity between the word and the corresponding centroid vector.

EMOTION	ADJECTIVES	COS.SIM.
JOY	*gioioso* "joyful"	0.85
	scanzonato "easygoing"	0.68
	spiritoso "funny"	0.66
ANGER	*insofferente* "intolerant"	0.72
	impaziente "anxious"	0.67
	permaloso "prickly"	0.66
SURPRISE	*perplesso* "perplexed"	0.81
	sgomento "dismayed"	0.73
	allibito "shocked"	0.73
DISGUST	*immondo* "dirty"	0.6
	malsano "unhealthy"	0.58
	insopportabile "intolerable"	0.58
FEAR	*impotente* "helpless"	0.6
	inquieto "restless"	0.57
	infelice "unhappy"	0.55
SADNESS	*triste* "sad"	0.8
	tetro "gloomy"	0.65
	sconsolato "surrowful"	0.62
TRUST	*disinteressato* "disinterested"	0.65
	rispettoso "respectful"	0.65
	laborioso "hard-working"	0.64
ANTICIPATION	*inquieto* "agitated"	0.7
	ansioso "anxious"	0.58
	desideroso "desirous"	0.56

Table 4.3: *Top 3 Adjectives for emotion with their association scores.*

As expected, a lot of target words have a high association score with more than one emotive class, and therefore some centroids are less discriminating because they have a similar distributional profile. Figure 4.3 shows the cosine similarity between the emotive centroids. In addition, we can observe for example, a high similarity between SADNESS and FEAR, as well as between SURPRISE and JOY. On one hand, this is consistent with the close relatedness between these emotions, on the other one, the implication is that *similar* emotions tend to share the same distributional profile, and then that it is difficult to discriminate between them.

4.1.4 Evaluation

In order to evaluate the system, we used a two-step crowdsourcing approach: first, for each ⟨emotion, PoS⟩ pair we ranked the target words with respect to their cosine similarity with the corresponding emotive centroid. We then selected the top 50 words for each centroid and we collected human ratings about the association of the target with each emotion: Given a target word ⟨w⟩, for each Plutchik's emotion ⟨e⟩, three annotators were asked to answer the question illustrated in Figure 4.4. The results of this experiment are provided in section 4.1.5, in which we refer to this experiment as **ParModel**.

The annotators rated words on a Likert scale ranging from 1 (not associated) to 5

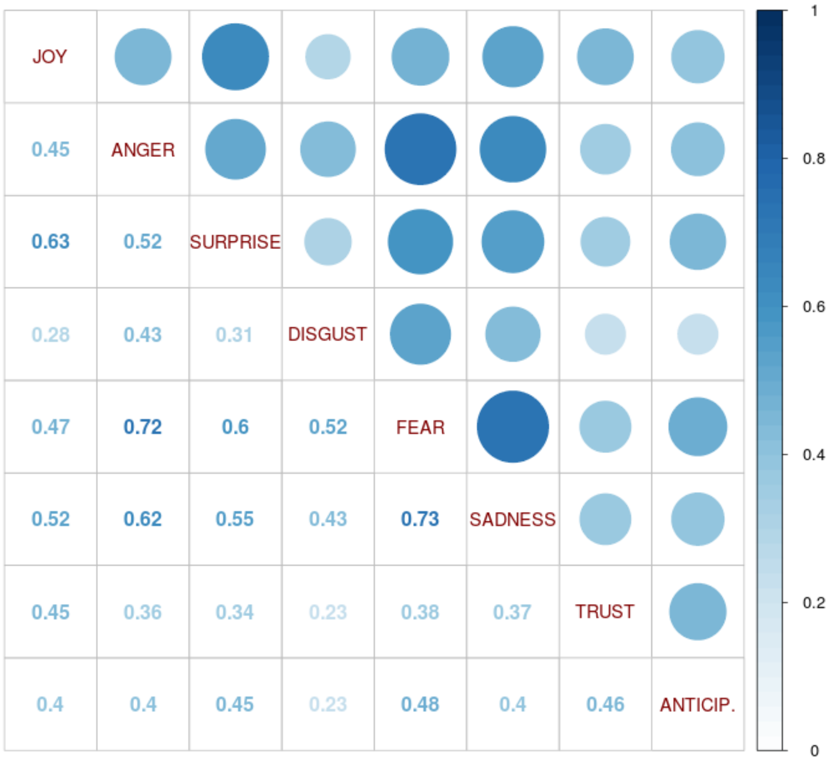

Figure 4.3: *Cosine similarity between the emotive centroids.*

How much is $\langle \mathbf{w} \rangle$ associated with the emotion $\langle \mathbf{e} \rangle^*$?

○ ○ ○ ○ ○

* Definition of the emotion $\langle \mathbf{e} \rangle$.

Figure 4.4: *Form used to validate the candidates extracted from the VSM. Each form corresponds to the pair ⟨word, emotion⟩ where an emotion belongs to the scheme in Plutchik [162].*

EMOTION	VERBS	COS.SIM
JOY	*rallegrare* “make happy”	0.6
	sentenceItconsolare “comfort”	0.54
	apprezzare “appraise”	0.53
ANGER	*inveire* “inveigh”	0.59
	maltrattare “abuse”	0.58
	offendere “offend”	0.56
SURPRISE	*stupefare* “amaze”	0.82
	sconcertare “disconcert”	0.81
	rimanere “remain”	0.79
DISGUST	*scandalizzare* “shock”	0.63
	indignare “make indignant”	0.53
	disapprovare “disapprove”	0.5
FEAR	*stupefare* “amaze”	0.7
	scioccare “shock”	0.68
	sbalordire “astonish”	0.68
SADNESS	*deludere* “betray”	0.78
	amareggiare “embitter”	0.75
	angosciare “anguish”	0.72
TRUST	*domandare* “ask”	0.64
	dubitare “doubt”	0.59
	meravigliare “amaze”	0.58
ANTICIPATION	*sforzare* “force”	0.56
	confortare “comfort”	0.56
	degnare “deign”	0.55

Table 4.4: *Top 3 verbs for emotion with their association scores.*

(highly associated). Since words are often associated with more than one emotion, we calculated a distinctiveness score in order to estimate the average degree of the association between a word the various emotions.

After ranking the words according to this emotive score, we selected the top 10 distinctive nouns, adjectives and verbs for each ⟨emotion, PoS⟩ pair, in order to further expand the set of the seeds used to build the distributional space. In [151] we showed that the process of stepwise seed expansion used to calculate the emotive centroids may be repeated several times, in order to optimize the system and improve its performance.

4.1.5 Context Selection Strategies

In order to optimize the emotive VSM and to discriminate better between the various emotions, we applied a context selection step starting from the idea that the emotive connotation of a target word is a function of both the paradigmatic similarity between the word and the emotive centroids, and of the syntagmatic associations between the target word and the top neighbors of the emotion seeds.

Different context selection strategies have been tested in order to select the best approach. The best performing models are based on a syntagmatic emotive score, combined with the cosine similarity, which is the original measure used to calculate the emotive association between a word and an emotion centroid vector.

In these models, for each word-emotion pair ⟨w,EMOTION⟩, we calculated a syn-

EMOTION	NOUNS	COS.SIM
JOY	“joy”	0.83
	ilarità “cheerfulness”	0.73
	tenerezza “tenderness”	0.72
ANGER	*impazienza* “impatience”	0.8
	dispetto “prank”	0.76
	rancore “resentment”	0.75
SURPRISE	*sgomento* “dismay”	0.74
	trepidazione “trepidation”	0.74
	turbamento “turmoil”	0.74
DISGUST	*fetore* “stink”	0.84
	escremento “excrement”	0.83
	putrefazione “rot”	0.82
FEAR	*disorientamento* “disorientation”	0.82
	angoscia “anguish”	0.81
	turbamento “turmoil”	0.79
SADNESS	*tristezza* “sadness”	0.91
	sconforto “discouragement”	0.88
	disperazione “desperation”	0.88
TRUST	*serietà* “seriousness”	0.91
	prudenza “caution”	0.9
	mitezza “mildness”	0.89
ANTICIPATION	*oracolo* “oracle”	0.77
	premonizione “premonition”	0.74
	preveggenza “presage”	0.73

Table 4.5: *Top 3 nouns for emotion with their association scores.*

tagmatic emotive score (SintScore) based on the association measure (AM) such as the PPMI between ⟨w⟩ and the seeds used to compute the vector of ⟨EMOTION⟩:

$$SintScore = \sum_{seed \in \text{EMOTION}} \text{AM}(w, seed_{\text{EMOTION}}) \tag{4.1}$$

The syntagmatic emotive score has been calculated for each word-emotion pair, and it has been used as a filter to compute the filtered VSMs **SintParModel** and **Top1000EmoPos**.

SintParModel is the model in which we restricted the contexts to the words with a sufficiently high cosine similarity with the emotive centroid vectors and a sufficiently high syntagmatic emotive score. In particular, we selected the contexts having, at least for one emotion:

$$CSim * SintScore > 1 \tag{4.2}$$

where

$$CSim(\overrightarrow{w}, \overrightarrow{\text{EMOTION}}) = \frac{\overrightarrow{w} \cdot \overrightarrow{\text{EMOTION}}}{||\overrightarrow{w}|| \cdot ||\overrightarrow{\text{EMOTION}}||} \tag{4.3}$$

and the association measure is the PPMI. Globally, this VSM includes 10,114 contexts before applying Singular Value Decomposition to 300 dimensions.

Top1000EmoPos is the model in which we adapted the algorithm proposed in Zhitomirsky-Geffet and Dagan [224] to bootstrap the emotive contexts starting from a standard approximation of the similarity space. In particular, we adapted their *Bootstrapped Feature Weight* (BFW) to capture both syntagmatic and paradigmatic properties of the words.

$$BFW(w, f) = \sum_{v \in \mathrm{WS}(f) \cap \mathrm{N}(w)} \textsc{Sim}(w, v) \tag{4.4}$$

The authors demonstrated that the definition of a bootstrapping scheme assures improved feature weights, and hence higher quality feature vectors. We applied their scheme in two steps: in the first one we promoted the most important contexts for each word, and in the second one we generalized the intuition to emotions, by defining a sort of *emotion neighborhood*, as the top 1,000 associated words for each of them.

In the experiments below, we defined $\mathrm{WS}(f)$ as the set of words having a positive PMI with f (i.e., the words for which f is an active feature) and N(w) as the set of the words *v* having a cosine similarity (SIM) with *w* greater than 0.2, which is an empirically fixed threshold (i.e., the semantic neighbourhood of *w*). Once the bootstrapped weights have been computed, we calculated the syntagmatic emotive score according to the formula (4.1), using the BFW as association measure.

Starting from these new weights, we ranked the contexts according to these values and we restricted the contexts of the matrix to the top 1,000 nouns, adjectives and verbs for each emotion. Globally, we selected 15,116 distinct contexts, and we applied SVD reducing the matrix to 300 dimensions.

To study the differences among the spaces, we performed two different experiments carried out by setting up crowdsourcing tasks on the Crowdflower (CF) platform[3]. In the first one we repeated the evaluation on the top 50 nouns, adjectives and verbs extracted from the three VSMs as in [151]. For each of the VSMs we measured the Precision at a particular rank ($P@K$) with K in [10, 20, 30, 40, 50].

In the second experiment, we compared the performance of the models on a random set of words, including also possibly neutral words, associated with human ratings about their association or lack of association with emotions. In this case, we measured Precision, Recall and F1-score. The metric used to compare the models is the F1-score at different values of K (F1@K).

In both the experiments, we employed *competition ranking* so that items that compare with equal CF score receive the same ranking number, and a gap is left in the ranking numbers. For example, if A ranks ahead of B and C (which compare equal) which are both ranked ahead of D, then A gets ranking number 1, B and C get ranking number 2 and D gets ranking number 4.

Experiment 1: Precision@K

Precision has been calculated by comparing the vector space model's candidates against the annotation obtained with crowdsourcing.

[3] `http://www.crowdflower.com.`

Select the emotions that the word $\langle$**w**$\rangle$ ($\langle$**PoS**$\rangle$) expresses.

☐ NO EMOTION

☐ JOY
☐ ANGER
☐ SURPRISE
☐ DISGUST
☐ FEAR
☐ SADNESS
☐ TRUST
☐ ANTICIPATION

Figure 4.5: *Form used to validate 1,170 words selected randomly from the list of the 30,000 targets. The list includes 390 nouns, 390 adjectives and 390 verbs balanced with respect to their frequency in la Repubblica [10] + itWac [11].*

For each $\langle$emotion, PoS$\rangle$ pair we ranked the target words with respect to their cosine similarity with the corresponding emotive centroid. We then selected the top k words for each centroid vector and we asked the annotators to provide an emotive score for the selected words.

In this experiment, True positives (TP) are the words found among the top k neighbours for a particular emotion and PoS, for which the annotators provided an average association score greater than 3 and False positives (FP) are the words found in the top k nouns, adjectives and verbs, but for which the aggregate evaluation of the annotators is equal or lower than 3.

In addition, as a measure of quality of seed lemmas, we report the results of this experiment restricted to the seeds used to build the centroid vectors (Tables 4.10 and 4.9).

Experiment 2: Performance on a random sample (F1-score)

In the second experiment, we compared the models on a sample of words selected randomly from the 30,000 most frequent targets extracted from itWac [11] and la Repubblica [10], according to the following strategy: We divided the list of the 30,000 target into 30 frequency ranges, and from each of them we selected randomly 39 words (13 nouns, 13 adjectives and 13 verbs), for a total of 1170 items.

To obtain highly reliable emotive ratings, we increased the number of annotators from 3 to 20. Given a target word $\langle$w$\rangle$, for each Plutchik's emotion $\langle$e$\rangle$ (plus the neutral one), the annotators were asked to select the emotions expressed by $\langle$w$\rangle$ using a multi-selection button. Besides the eight emotions, a null option was available in case $\langle$w$\rangle$ was considered emotionally neutral (cf. Figure 4.5).

For each annotated word, CF provides a confidence score describing the level of agreement between multiple raters. The aggregate answer returned by CF is the majority vote, weighted by each contributors' trust scores. The aggregate answer is chosen by considering the response with the greatest confidence.

In Table 4.6, we report the levels of agreement (mean and standard deviation per emotion) and the number of words for which the aggregate answer corresponds to the target emotion. The last row lists the values for the neutral words.

EMOTION	MEAN	ST.DEV	ITEMS
ANTICIPATION	0.52	0.07	6
DISGUST	0.67	0.22	23
TRUST	0.55	0.15	21
JOY	0.66	0.19	34
FEAR	0.6	0.19	37
ANGER	0.58	0.14	40
SURPRISE	0.69	0.14	5
SADNESS	0.7	0.19	20
NO EMOTION	0.82	0.16	984

Table 4.6: *Agreement (mean and standard deviation per emotion).*

In Table 4.7 we report on the distribution of the CF scores in the random sample. High (CF score) is assigned to the words receiving more than 10 ratings for the target emotion (plus the neutral one). Medium is assigned to the words receiving between 5 and 10 ratings, Low is assigned to the words receiving between 1 and 4 ratings and Zero to the others.

EMOTION	HIGH	MEDIUM	LOW	ZERO
ANTICIPATION	3	28	303	836
DISGUST	17	31	153	969
TRUST	11	67	305	787
JOY	23	49	255	843
FEAR	23	76	215	856
ANGER	21	56	214	879
SURPRISE	4	12	179	975
SADNESS	15	23	210	922
NO EMOTION	923	152	72	23

Table 4.7: *Distribution of CF scores in the random sample.*

To compute Precision, Recall and F1, given a $\langle$emotion, word-PoS$\rangle$ pair, we used a fixed threshold on the number of CF raters and a variable threshold on the cosine similarity (Table 4.8). In particular, we sorted the candidates according to their cosine similarity for each $\langle$emotion, PoS$\rangle$ pair, and the threshold was fixed at the 3^{rd} quartile for each group. Gold standard entries have been generated according to the values in Table 4.8.

4.1.6 Results and discussion

A preliminary analysis is aimed at evaluating the quality of the seed lemmas used to build the centroid vectors. If we look at the human ratings on the 228 seed lemmas belonging to the top 50 emotion-pos neighbors for at least one of the models, we can notice that the precision varies depending on the emotions and parts of speech. Table

CF SCORE	COSINE SIMILARITY	DEFINITION
>5	> 3^{rd} Quartile	True Positive
<=5	> 3^{rd} Quartile	False Positive
>5	<= 3^{rd} Quartile	False Negative
<=5	<= 3^{rd} Quartile	True Negative

Table 4.8: *Gold Standard entries.*

4.9 shows the precision aggregated by emotion and Table 4.10 shows the precision aggregated by PoS. Overall, a half of the seed lemmas are in the top 50 neighbors for the corresponding emotion, and the human ratings collected for them are very high, except for the case of the emotion TRUST. A more detailed analysis showed that the global precision of the emotion TRUST depends on the verbs, for which the attested precision is 0.33. Although this is an incomplete assessment, because we have not evaluated the entire set of the seed lemmas, the evaluation demonstrates that a tuning of the seed selection phase could optimize the global performance of the system. As for the remaining seed lemmas, for which we don't know the human judgements, we cannot conclude neither that they follow the same trend, nor that they are rated worse. The fact that they are not in the top 50 neighbors for the corresponding emotion does not mean that the system assigned them a low emotive score, but it could be due to their frequency in the corpus.

EMOTION	TRUE POSITIVE	FALSE POSITIVE	PRECISION
ANTICIPATION	27	2	0.931
DISGUST	49	3	0.942
TRUST	25	8	0.758
JOY	35	0	1.000
FEAR	49	3	0.942
ANGER	33	4	0.892
SURPRISE	36	1	0.973
SADNESS	44	0	1.000

Table 4.9: *Seed lemmas. Precision aggregated by Emotion.*

POS	TRUE POSITIVE	FALSE POSITIVE	PRECISION
ADJECTIVES	95	3	0.969
NOUNS	63	4	0.94
VERBS	81	5	0.942

Table 4.10: *Seed lemmas. Precision aggregated by PoS.*

If we consider the differences between the filtered models we can notice that overall, the best model seems to be the one in which we filtered the contexts having a sufficiently high cosine similarity and syntagmatic emotive score (SintParModel).

Figures 4.6 and 4.7 show the global performance for the two experiments.

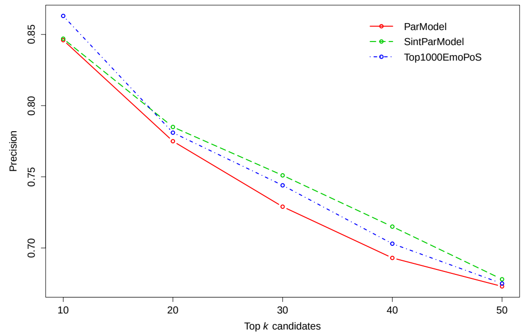

Figure 4.6: *Precision@k by VSM.*

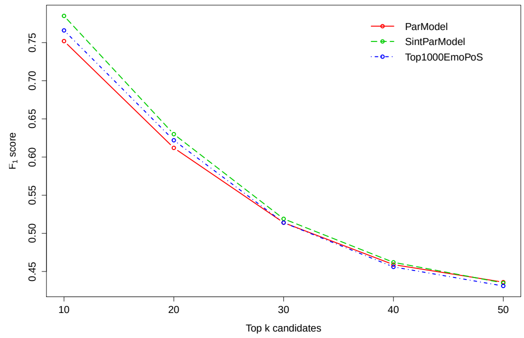

Figure 4.7: *F1-score@k by VSM.*

Globally, the filtered models show a more *graceful degradation* of precision when K increases and reach their highest performance with the top 30 neighbours. For example, the precision on the top 30 neighbours raises by 0.4 percent in the case of nouns, and it raises by 0.6 percent in the case of verbs. The same trend (though less pronounced) is observed if we consider the F1-score on the random sample.

With regard to the Precision@K on the top candidates and F1-score@K on the random sample (cf. Table 4.11 and Table 4.13), most of the emotions seem to take advantage from the context selection, but, at the same time, the different results prompt us to further investigate the quality of the seeds representing emotions (cf. Figures 4.8 and 4.10).

For example, for the emotions JOY, DISGUST SADNESS and SURPRISE context selection seems to improve their performance on the top K candidates, but such an improvement is not found in the random sample. These effects could be due to the fact that in our feature elicitation experiment ANGER terms tend also to be associated with DISGUST and many JOY terms are also associated with TRUST and SURPRISE.

In other words, the distinctiveness of the seeds impacts on the final results, but the effect is more evident if we consider a random sample of words which includes emotive words with a moderate intensity. In addition, it is important to stress that by considering top 50 neighbours, the filtered models reach a higher precision level for all the emotion (except the ANTICIPATION).

Moving to the results by PoS (cf. Tables 4.12 and 4.14, Figures 4.9 and 4.11), though for adjectives the best model is the complete one (ParModel), verbs and nouns are better represented by the filtered models (SintParModel and Top1000EmoPoS).

This evaluation demonstrates the possibility to improve ItEM by exploiting context selection methods based on syntagmatic and paradigmatic distributional associations. In addition, these results may suggest that different PoS require different approaches, e.g. using filtered spaces for verbs and nouns only. At the same time, we noticed that the improvement is not balanced with respect to the emotions. In fact the emotions JOY, DISGUST SADNESS and SURPRISE reach a higher F1-score in the full model. Thus, a future optimization might try to combine different approaches for the various emotions.

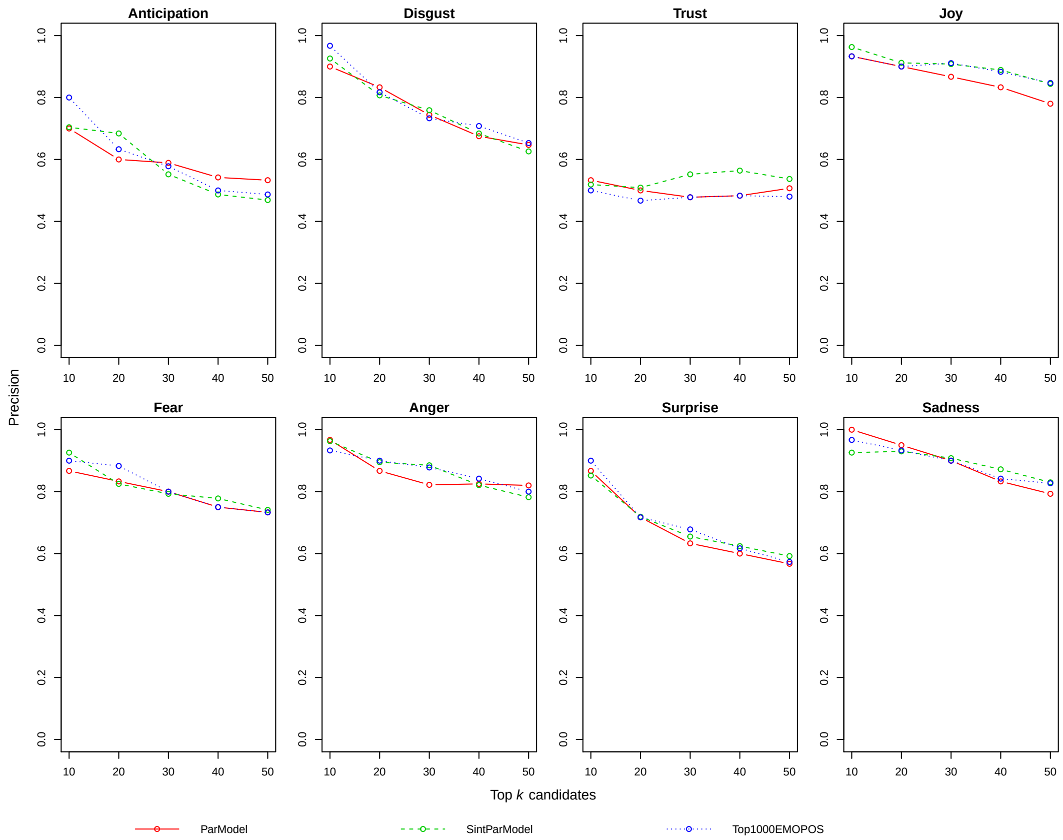

Figure 4.8: *Precision@k by Emotion (Experiment 1).*

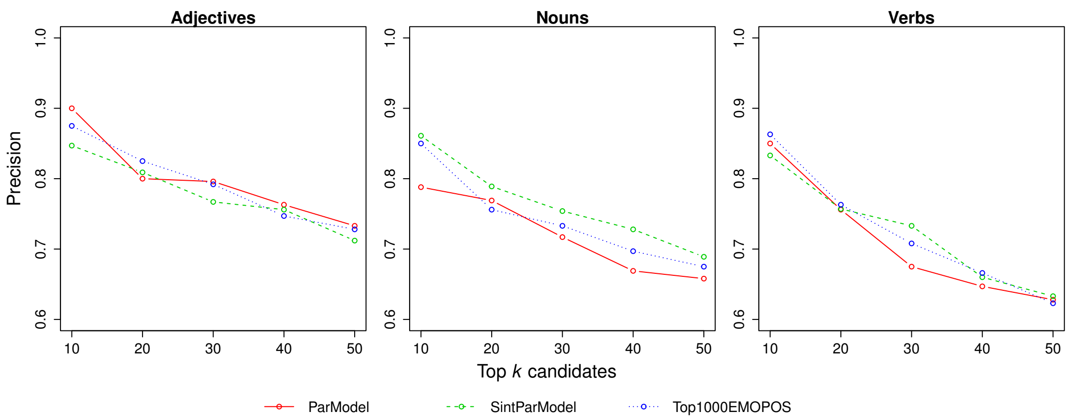

Figure 4.9: *Precision@k by PoS (Experiment 1).*

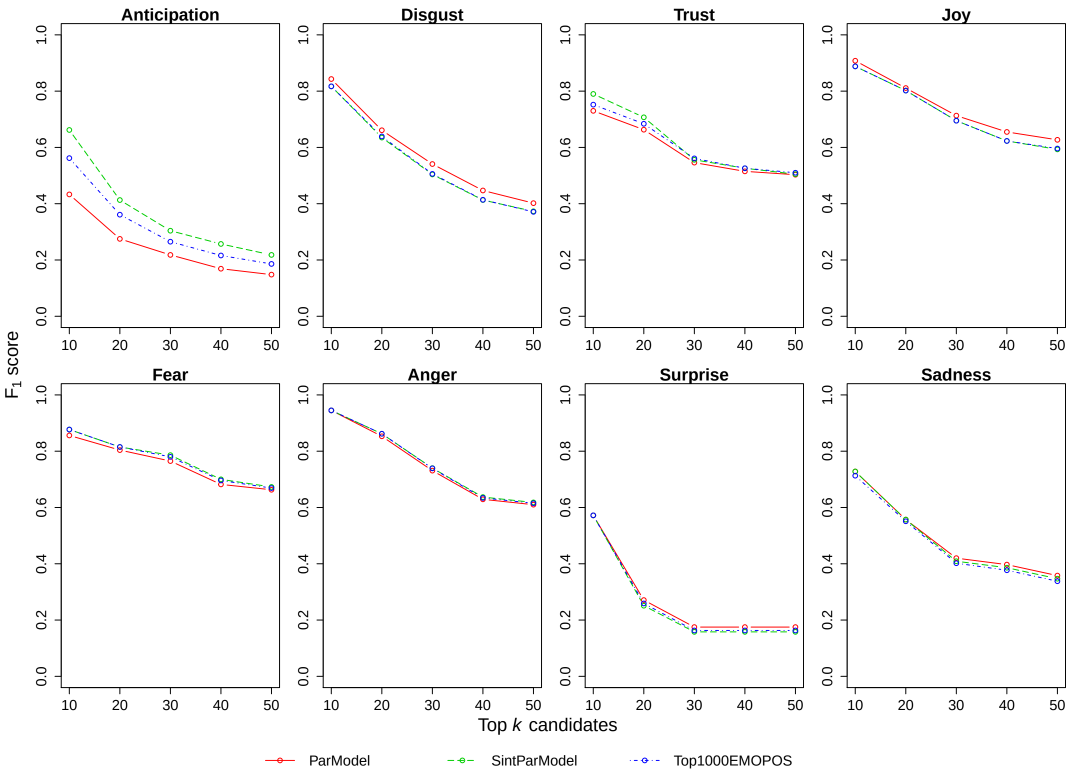

Figure 4.10: *F1-score@k by Emotion (Experiment 2).*

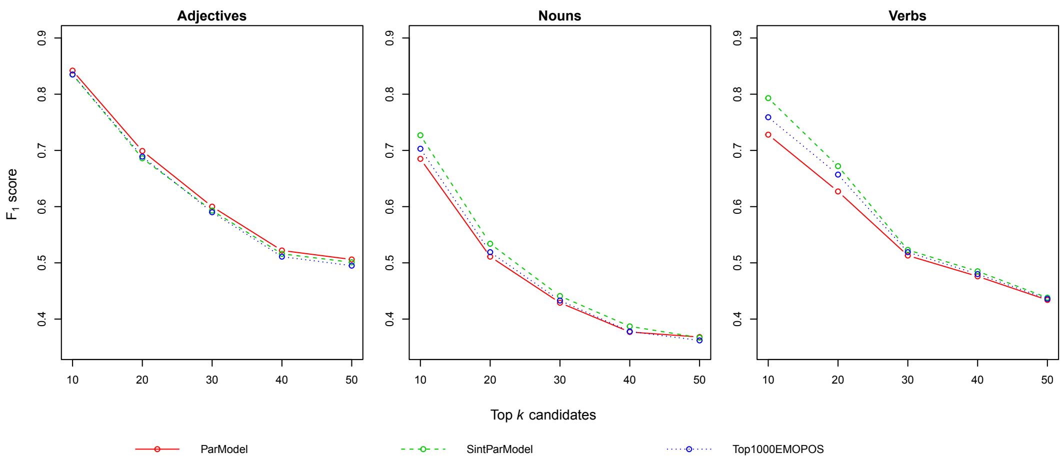

Figure 4.11: *F1-score@k by PoS (Experiment 2).*

EMOTION	MODEL	P@10	P@20	P@30	P@40	P@50
	ParModel	0.7	0.6	**0.589**	**0.542**	**0.533**
ANTICIPATION	SintParModel	0.704	0.684	0.552	0.487	0.469
	Top1000EmoPoS	**0.8**	**0.633**	0.578	0.5	0.487
	ParModel	0.9	**0.833**	0.744	0.675	0.647
DISGUST	SintParModel	0.926	0.807	**0.759**	0.684	0.626
	Top1000EmoPoS	**0.967**	0.817	0.733	**0.708**	**0.653**
	ParModel	**0.533**	0.5	0.478	0.483	0.507
TRUST	SintParModel	0.519	**0.509**	**0.552**	**0.564**	**0.537**
	Top1000EmoPoS	0.5	0.467	0.478	0.483	0.48
	ParModel	0.933	0.9	0.867	0.833	0.78
JOY	SintParModel	**0.963**	**0.912**	0.908	**0.889**	0.844
	Top1000EmoPoS	0.933	0.9	**0.911**	0.883	**0.847**
	ParModel	0.867	0.833	**0.8**	0.75	0.733
FEAR	SintParModel	**0.926**	0.825	0.793	**0.778**	**0.741**
	Top1000EmoPoS	0.9	**0.883**	**0.8**	0.75	0.733
	ParModel	**0.967**	0.867	0.822	0.825	**0.82**
ANGER	SintParModel	0.963	0.895	**0.885**	0.821	0.782
	Top1000EmoPoS	0.933	**0.9**	0.878	**0.842**	0.8
	ParModel	0.867	0.717	0.633	0.6	0.567
SURPRISE	SintParModel	0.852	**0.719**	0.655	**0.624**	**0.592**
	Top1000EmoPoS	**0.9**	0.717	**0.678**	0.617	0.573
	ParModel	**1**	**0.95**	0.9	0.833	0.793
SADNESS	SintParModel	0.926	0.93	**0.908**	**0.872**	**0.83**
	Top1000EmoPoS	0.967	0.933	0.9	0.842	0.827

Table 4.11: *Precision@k aggregated by emotion and VSM (Experiment 1).*

MODEL	POS	P@10	P@20	P@30	P@40	P@50
	ParModel	**0.9**	0.8	**0.796**	**0.763**	**0.733**
ADJECTIVES	SintParModel	0.847	0.809	0.767	0.756	0.712
	Top1000EmoPoS	0.875	**0.825**	0.792	0.747	0.728
	ParModel	0.788	0.769	0.717	0.669	0.658
NOUNS	SintParModel	**0.861**	**0.789**	**0.754**	**0.728**	**0.689**
	Top1000EmoPoS	0.85	0.756	0.733	0.697	0.675
	ParModel	0.85	0.756	0.675	0.647	0.628
VERBS	SintParModel	0.833	0.757	**0.733**	0.66	**0.633**
	Top1000EmoPoS	**0.863**	**0.763**	0.708	**0.666**	0.623

Table 4.12: *Precision@k aggregated by PoS and VSM (Experiment 1).*

4.2 FB-NEWS15: A Corpus of Facebook News and Comments

The use of Social Networks (SN) platforms like Facebook and Twitter has developed overwhelmingly in recent years. SNs are exploited for different purposes ranging from the sharing of contents among friends and useful contacts to the newsgathering about different domains such as politics and sports [1, 35, 186]. Many journalists indeed use SN platforms for professional reasons [75, 138].

Several recent studies provide insights on how the popularity of blogs and other user generated content impacted the way in which news are consumed and reported.

EMOTION	MODEL	F1@10	F1@20	F1@30	F1@40	F1@50
ANTICIPATION	ParModel	0.433	0.275	0.218	0.169	0.148
	SintParModel	**0.662**	**0.413**	**0.304**	**0.257**	**0.218**
	top1000EMOPOS	0.562	0.361	0.265	0.216	0.186
DISGUST	ParModel	**0.843**	**0.661**	**0.541**	**0.447**	**0.402**
	SintParModel	0.817	0.635	0.504	0.413	0.373
	top1000EMOPOS	0.817	0.639	0.506	0.414	0.371
TRUST	ParModel	0.73	0.663	0.546	0.515	0.503
	SintParModel	**0.79**	**0.707**	0.556	**0.526**	0.505
	top1000EMOPOS	0.752	0.684	**0.561**	**0.526**	**0.51**
JOY	ParModel	**0.908**	**0.811**	**0.713**	**0.655**	**0.627**
	SintParModel	0.888	0.802	0.695	0.623	0.593
	top1000EMOPOS	0.888	0.802	0.695	0.623	0.596
FEAR	ParModel	0.856	0.804	0.765	0.682	0.663
	SintParModel	**0.877**	**0.815**	**0.786**	**0.7**	**0.673**
	top1000EMOPOS	**0.877**	**0.815**	0.78	0.696	0.669
ANGER	ParModel	**0.945**	0.853	0.731	0.629	0.61
	SintParModel	**0.945**	**0.862**	**0.74**	**0.637**	**0.618**
	top1000EMOPOS	**0.945**	**0.862**	0.739	0.634	0.615
SURPRISE	ParModel	**0.572**	**0.271**	**0.175**	**0.175**	**0.175**
	SintParModel	**0.572**	0.251	0.158	0.158	0.158
	top1000EMOPOS	**0.572**	0.259	0.163	0.163	0.163
SADNESS	ParModel	**0.728**	**0.557**	**0.42**	**0.397**	**0.358**
	SintParModel	**0.728**	**0.557**	0.409	0.386	0.347
	top1000EMOPOS	0.713	0.551	0.402	0.377	0.338

Table 4.13: *F1-score@k aggregated by Emotion and VSM (Experiment 2).*

MODEL	POS	F1@10	F1@20	F1@30	F1@40	F1@50
ADJECTIVES	ParModel	**0.842**	**0.699**	**0.6**	**0.522**	**0.506**
	SintParModel	0.835	0.686	0.593	0.516	0.501
	top1000EMOPOS	0.835	0.689	0.59	0.511	0.495
NOUNS	ParModel	0.685	0.511	0.429	0.377	**0.368**
	SintParModel	**0.727**	**0.534**	**0.441**	**0.387**	0.367
	top1000EMOPOS	0.703	0.519	0.433	0.378	0.362
VERBS	ParModel	0.728	0.627	0.513	0.476	0.434
	SintParModel	**0.793**	**0.672**	**0.523**	**0.485**	**0.438**
	top1000EMOPOS	0.759	0.657	0.519	0.48	0.436

Table 4.14: *F1-score@k aggregated by PoS and VSM (Experiment 2).*

Picard [156] states that SN platforms provide an easy and affordable way to take part in discussions with larger groups of people and, consequently, the bond between SN and information is becoming increasingly stronger.

Mass information is gradually moving towards general platforms, and official websites are losing their lead position in providing information. As noted by Newman and colleagues [132], even though the use of internet in the years 2009-2012 has grown, the same is not reflected in the consumption of online newspapers, probably because of the increasing use of SN for news diffusion and gathering. If on the one hand this apparent decline of the traditional news platforms may lead to a decline in quality and news coverage [32], on the other hand the rise of SN as platforms to spread news promotes a more fervid debate between users [185].

This issue is central for the present work. In fact, user's comments very often contain their own opinions about a certain problem or event. In addition, because of the colloquial style in comments, they contain large amounts of words and collocations with a high subjective content, mostly concerning the author's emotive stance.

Facebook is one of the most popular online SN in the world with 1 billion active users per month and it offers the possibility to collect data from people of different ages, educational levels and cultures. From a linguistic point of view, previous studies [105] demonstrated that the language in Facebook is more emotional and interpersonal compared for example to the language in Twitter. Probably, this is due to the fact that in Facebook there is a stronger psychological closeness between the author and audience because of the different structure of the SNs (bidirectional vs. unidirectional graphs).

The debate among users in commenting news and posts on Facebook offers a lot of subjective material to study the way in which people express their own opinions and emotions about a target event. The corpus FB-NEWS15 collects these linguistic items expressing the whole range of positive and negative emotions.

In analyzing a news corpus, however, it is not simple to aggregate the posts on the basis of a certain fact, since several posts relate to the same event. For this reason, we decided to organize the corpus into clusters of topically related news identified with Latent Dirichlet Allocation (LDA: [19]). This approach allows us to infer the most debated news in the corpus, and, in a second step, to discover the readers' sentiment about a particular topic.

4.2.1 Overview of the corpus

For the creation of the corpus we followed the most important Italian newspapers. Since we were interested in building a corpus as heterogeneous as possible, we decided to focus on major newspapers with different political orientations, and which have in general heterogeneous readers (followers).

Facebook allow users to post states, links, photos and videos on their own wall. In general, users can be divided into two macro-categories: People and Pages. People are often individuals, and the interaction with them is usually bidirectional (user A can read what user B publishes if A and B have a *friendship relation*). Conversely, pages are typically used to represent organizations, public figures (web stars), companies or, as in our case, newspapers. In this case, the relationship is unidirectional, in the sense that user A can access the timeline of the page P by putting a *Like* on P. Unlike a single-user, who usually publishes photos, videos and links about his *private* life, the timeline of a newspaper Facebook page, in general contains news titles with a link to the official website of the newspaper, where the user can read the full article.

The corpus keeps tracks of the threefold hierarchical structure of Facebook, which includes the news posts by the newspaper, the users' comments to the posts and the replies to the comments. In this context, it is clear that the emotive content of the post is often neutral, but this post can inspire long discussions among readers, which can become useful material for sentiment analysis and emotion detection. Figure 4.12 shows a post, with some of its comments and replies.

In order to create the FB-NEWS15, we decided to download the timeline of the

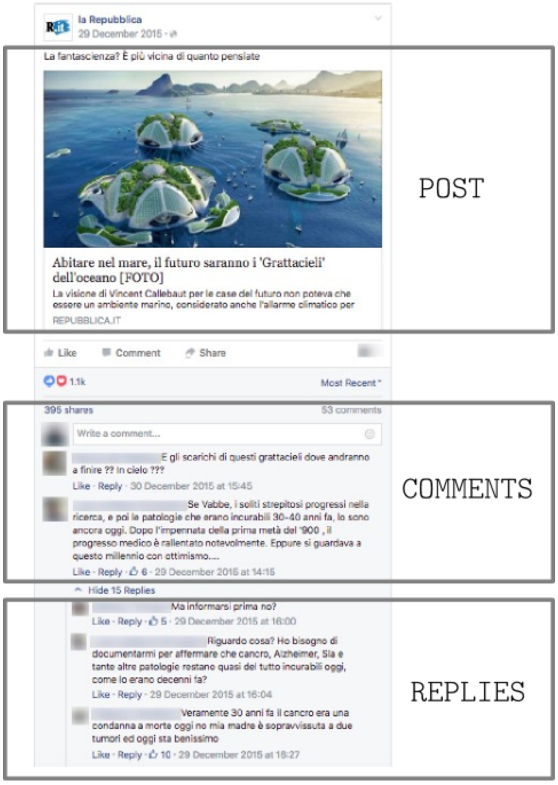

Figure 4.12: *Example of post in Facebook with the relative comments and replies.*

following newspapers, from 1 January 1 to 31 December 2015:

- La Repubblica
- Il Giornale
- L'Avvenire
- Libero
- Il Fatto Quotidiano
- Rainews24
- Corriere Della Sera
- Huffington Post Italia

4.2.2 Crawling

Facebook offers developers Application Programming Interfaces (APIs) for creating apps with Facebook's native functionalities. In order to develop the crawler, we exploited the Graph API, which provides a simple view of the Facebook social graph by showing the objects in the graph and the connections between them. The Graph API allows us to navigate through the graph of the social network, which is organized into

```
<doc user="<newspaper(string)>"  id="<id_post(string)>"  type="post"
 parent_post=""  parent_comment="" date="AAAA-MM-DD HH:MM:SS"
 location=""  likes="662" comments="54" shares="322">

Un business truffaldino [E ora finitela con l'eco-balla dei
controlli sulle emissioni]

</doc>
```

Figure 4.13: *Example of crawled text.*

nodes (Users, Pages, Photos and Comments) and Edges (Connections such as Friendship or Likes). The graph is navigated by exploiting HTTP requests, that may be implemented using any programming language. The native APIs offered by Facebook has some drawbacks: (i) the maintenance of the app, since the APIs change over time, making it necessary to update the code of the crawler; (ii) only public data can be accessed without requiring the user's consent; (iii) Facebook places limitations on the number of requests through a given period of time. For each post, comment and reply, we stored the following information offered by Facebook:

– `message`: the text of the post;

– `story`: additional information about the message such as the presence of a photo, a link or a tag referring to another user;

– `created time`: timestamp of the message;

– `type`: the metadata specifies whether the text is a post or a comment (comment or response);

– `likes.summary (true)`: number of likes obtained from the post;

– `shares`: total number of shares obtained by post (available for posts only);

– `comments.summary (true)`: number of comments (available for posts only);

For each comment (or response), we stored the message, the creation date, the number of likes and (recursively) its answers. Figure 4.13 shows an example of post.

Linguistic annotation

A very basic preprocessing phase has been applied to the corpus before linguistic annotation, in order to replace urls with the tag `"_URL_"`. The text has been subsequently feed to the pipeline of general-purpose NLP tools described in chapter 3. In particular, the corpus has been POS-tagged with the Part-Of-Speech tagger described in [41] and dependency-parsed with the DeSR parser [6]. In addition, complex terms like *forze dell'ordine* (security force) or *toccare il fondo* (hit rock bottom) have been identified using the EXTra term extraction tool [150].

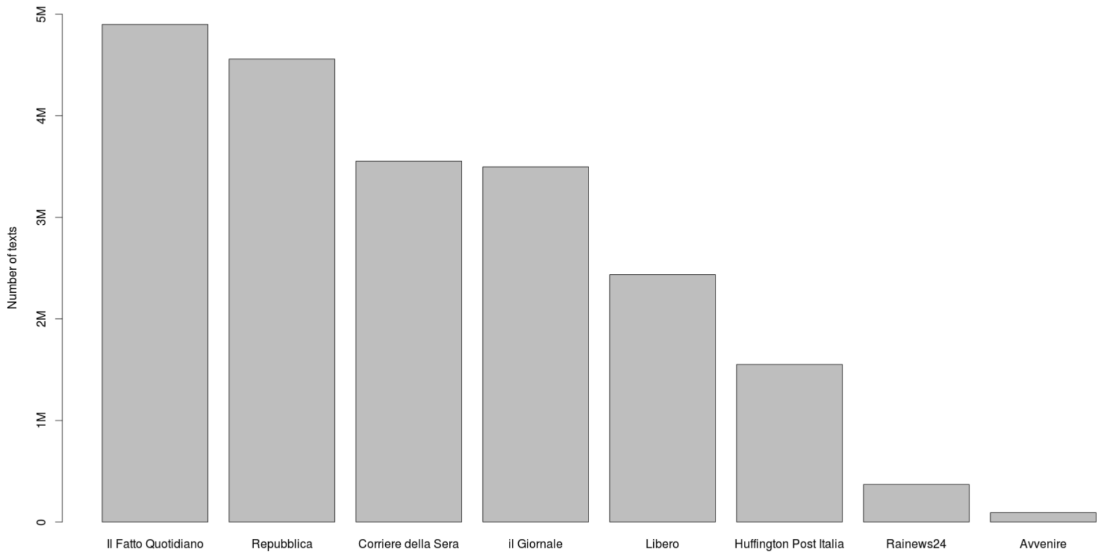

Figure 4.14: *Number of texts aggregated by Newspaper in FB-NEWS.*

4.2.3 Corpus Analysis

In general, the number of posts is very low compared to the number of comments and replies. The average number of posts for each newspaper is 27,341.25, while for comments and replies is respectively 2,016,243.38 and 576,498.5. Figure 4.14 shows the number of texts (including posts, comments and replies) for each newspaper, and Figure 4.15 represents the cumulative percentage of posts, comments and replies for each of them.

Table 4.15 shows the total number of tokens for each page and the average number of texts, including comments and replies, produced for each post in a particular Facebook page. We can notice that the most followed newspapers on Facebook are Il Fatto Quotidiano and La Repubblica.

NEWSPAPER	TOKENS	TEXTS/POSTS
La Repubblica	96,059,756	182.61
Avvenire	2,611,899	12.65
Il Giornale	64,345,260	77.93
Libero	41,166,457	81.87
Il Fatto Quotidiano	99,025,541	193.33
Rainews24	7,735,908	10.21
Huffington Post	32,587,065	84.06
Corriere della Sera	64,197,579	95.01
OVERALL	407,729,465	94.83

Table 4.15: *Tokens and texts/posts ratio for page.*

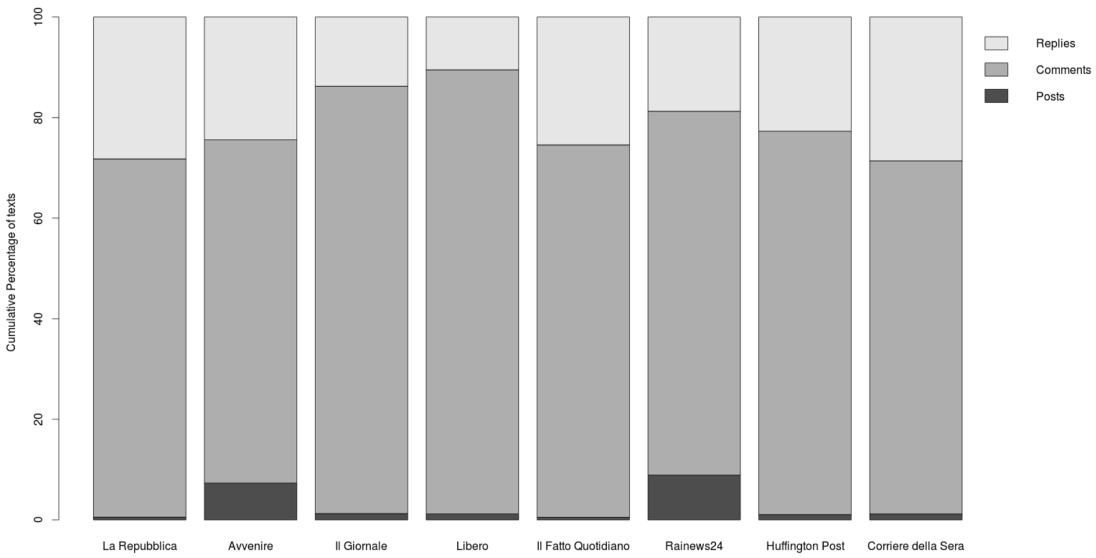

Figure 4.15: *Cumulative distribution of posts, comments and replies in FB-NEWS15 for each Newspaper.*

4.2.4 Topics in FB-NEWS15

FB-NEWS15 contains texts referring to a large variety of events. In order to organize the corpus into clusters of thematically related news, we used LDA [19]. LDA represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words. These random mixtures express a document semantic content, and document similarity can be estimated by looking at how similar the corresponding topic mixtures are. For the topic identification we used the software Mallet [114] which is a Java implementation of LDA.

LDA [19] is a probabilistic model that assumes that each document is a mixture of latent topics. For each latent topic T, the model learns a conditional distribution $p(w|T)$ for the probability that word w occurs in T. In this way, one can obtain a k-dimensional vector representation of words by first training a $k - topic$ model and then filling the matrix with the normalized probabilities $p(w|T)$.

Selecting the vocabulary

Since we were interested in extracting the topics from the news articles, we have built the model on the portion of FB-NEWS15 containing the posts (FB-NEWS15_posts[4]) published by the newspaper. In particular, we used entropy [48] as a global term weighting and we selected for training the terms (nouns, adjectives, verbs and complex terms) with a high informative value (threshold fixed to 0.3), while using the remaining words as stopwords in Mallet [114].

[4]FB-NEWS15_posts is the portion of the corpus FB-NEWS15 that contains only posts. We approximate the number of posts with the number of posts published by the newspapers.

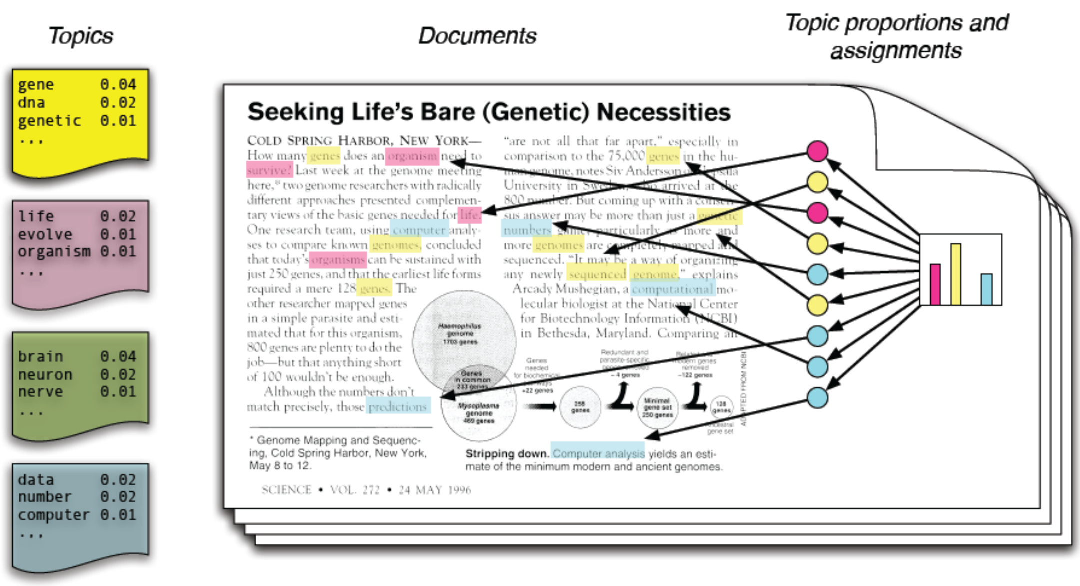

Figure 4.16: *The intuitions behind latent Dirichlet allocation. Illustration from Blei (2012). "Probabilistic Topic Models" [18].*

Extracting topics from posts

In order to determine the most debated topics in 2015, we extracted 50 topics from FB-NEWS15_posts and we navigated the graph to assign a topic to each comment and reply. Each post has been assigned to a topic T if its probability of belonging to T was higher than the 90^{th} percentile of its topic distribution. On average, each post has been assigned to 3.06 topics. Finally, comments and replies have inherited the probability of belonging to the topic T from their parent post.

Among the extracted topics ranked according to the sum of these probabilities we can find national and foreign politics, terrorism and church but also food, football, cinema and weather forecast. We report some topics below, with the relative ranking and number of comments and replies (texts).

NATIONAL POLITICS (2,516,640 TEXTS, RANK 1): *{Renzi, presidente, premier, Mattarella, riforma, Alfano, senato, camera, Boschi, aula}* (Renzi, president, Mattarella, reform, Alfano, senate, chamber, Boschi, hall);

SCHOOL (1,707,145 TEXTS, RANK 2):*{scuola, giovane, studente, protesta, corso, mancare, sospendere, inglese, spiegare, lezione}* (school, young, protest, class, lack, suspend, English, explain, lesson);

CRIME (1,543,735 TEXTS, RANK 7): *{uccidere, polizia, arrestare, fermare, sparare, uomo, poliziotto, colpo, ferire, agente}* (kill, police, detain, stop, open fire, man, policeman, bump, wound, police officer);

MIGRANTS (1,376,298 TEXTS, RANK 12): *{migrare, profugo, mare, migrante, immigrato, immigrazione, treno, emergenza, Sicilia bloccare}* (migrate, refugee, sea, migrant, immigrant, immigration, train, emergency, Sicily, block);

FOREIGN POLITICS (1,389,080 TEXTS, RANK 10): *{Europa, Grecia, piano, accordo, Tsipras, crisi, votare il sondaggio, Merkel, Germania, Atene }* (Europe, Greece, plan, pact, Tsipras, crisis, vote on the survey, Merkel, Germany, Athens);

ISIS (1,267,749 TEXTS, RANK 16): *{Isis, guerra, Siria, minaccia, U.S.A., Libia, colpire, islamico, usare, jihadisti }* (Isis, war, Syria, threat, U.S.A., Libya, damage, islamic, use, jihadist);

CHURCH (1,320,797 TEXTS, RANK 17): *{papa, san, chiesa, aprire, Vaticano, appello, incontrare, porta, pace, miliardo di euro}* (pope, saint, church, open, Vatican, plea, meet, door, peace, billions of euros);

FOOD (949,520 TEXTS, RANK 40): *{mangiare, ricetta, cibo, preparare, consiglio, evitare, perfetto, trucco, salute, semplice}* (eat, recipe, food, prepare, advice, avoid, perfect, trick, health, simple);

PARIS TERROR ATTACK (955,923 TEXTS, RANK 37): *Parigi, morto, strage, attacco, vittima, Francia, ferito, terrorismo, terrorista, bomba}* (Paris, dead, massacre, attack, victim, France, hurt, terrorism, terrorist, bomb);

FOOTBALL (606,560 TEXTS, RANK 50): *{seguire la diretta, guardare il video, campo, calcio, serie, Napoli, Milan, segnare, battere, partita}* (follow the live, look at the video, football field, football, league, Naples, Milan).

The structure of FB-NEWS15, in which each comment is explicitly associated with a particular post, allows us to study the differences in terms of readers' perceptions about a particular topic. Differently from other social media like Twitter, Facebook contains larger texts including lot of subjective expressions that are very useful for the construction of sentiment and emotive lexicons.

Starting from the construction of ItEM [151, 152] we used the FB-NEWS15 corpus to build a new lexical resources for sentiment analysis which includes both words and complex terms. The resource has been used to build new emotive features for sentiment polarity classification (see section 5).

4.3 Additional resources

Textual communication lacks many of the factors charachterizing face-to-face communication. To combat this, users often exploit new strategies to substitute physical cues in online settings [167]. Some of these are very creative such as acronyms, new idioms, hashtags, swear words, emoticons and emojis. Among this group, acronyms, idioms and hashtags have often a short life cycle and emojis are very difficult to deal with using only text processing tools. On the contrary, emoticons and common bad words are very popular and also very affirmed in social media communication, in the sense that users tend to use a low number of them in their posts [213].

Emoticons and bad words, however, often don't reflect the polarity of the post in which they are inserted. For example, bad words can be used in positive posts as an emphasis (i. e. *Sei fottutamente sveglio! :D* "you are smart as fuck! :D", or negative emoticons can be used in positive sentences (i. e. *:'-(Piango di felicità :'-(* ":'-(I am crying of joy :'-(".

Even though their polarity is not a complete indicator of the polarity of the texts in which they appear, these features can be very useful in classification tasks.

4.3.1 Emoticons

Emoticons, such as `:)`, `;)`, `:-)` and `:(` are frequently used online in social media platforms ranging from Skype to Twitter and Facebook. Because they are commonly used in online communication, and they seems to be direct signals of sentiment, emoticons have been widely used in NLP [64] as features to train machine learning algorithms for sentiment analysis and to build lexicons for rule-based approaches.

LexEmo (Lexicon of Emoticons) represents a relatively comprehensive inventory of common emoticons taken from previous studies and wikipedia. In LexEmo, emoticons are marked with their polarity score, which can be 1 (positive), -1 (negative), 0 (neutral) and possibly with their associated emotions. The lexicon includes in total 113 emoticons (60 positive; 36 negative; 17 neutral).

Table 4.16 shows the attested emoticons in the corpus FB-NEWS-15.

EMOTICON	FREQUENCY	POLARITY	EMOTIONS
:)	102204	1	joy
:-)	54828	1	joy
<3	43759	1	_
:(	36906	-1	sad
:-(	18808	-1	sad
XD	10564	1	joy
xD	6760	1	joy
^_^	4910	1	_
DX	855	-1	disgust/sad
O_o	524	0	surprise
:|	520	0	_
O_O	399	0	surprise
o_O	395	0	surprise
%)	189	-1	_
XP	110	1	joy
xp	106	1	joy
v.v	76	-1	disgust/sadness
:-|	74	0	_
o_o	35	0	surprise
\o/	24	1	_

Table 4.16: *Top 20 emoticons attested in the corpus FB-NEWS-15, with the relative Frequency, Polarity and Associated emotions.*

4.3.2 Bad words

Although swear words sometimes appear in positive posts, a small lexicon of bad words is often useful to capture the sentiment orientation of a given message. For this reason, we collected a list of 126 Italian swear words (60 lemmas) from the Internet.

As an indication to negative sentiment, we counted the number of occurrences of those bad words in tweets and we used it as a feature in the experiments presented in chapter 5.

4.4 Summary

The chapter presents a distributional approach to build emotive lexical resources that can be used for sentiment analysis and emotion detection. Section 4.1 describes the creation of ItEM, which is a high-coverage Italian emotive lexicon. It has been built with a three stage process including the collection of highly emotive words to be used as seeds, the distributional expansion of these seeds and the validation of the system.

Section 4.2 presents a new Italian corpus composed of Facebook posts. The corpus has been created by crawling the Facebook timeline of the most important Italian newspapers. Such timelines typically include a low number of posts written by the journalists and a very high number of comments provided by web readers. Given the nature of the corpus, it is clear that the emotive content of the posts is often neutral, but this post can inspire long discussions among readers, which can become useful material for sentiment analysis and emotion detection.

Finally, additional in-house built lexicons (emoticons and bad words) are described in the last section.

CHAPTER 5

Sentiment Polarity Classification of tweets

Nowadays social media and microblogging services are extensively used for rather different purposes, from news reading to news spreading, from entertainment to marketing. As a consequence, the study of how sentiments and emotions are expressed in such platforms, and the development of methods to automatically identify them, has emerged as a great area of interest in the Natural Language Processing Community.

In this context, the research on sentiment analysis and detection of speaker-intended emotions from Twitter messages (tweets) appears to be a task on its own, rather distant from the previous sentiment classification research that focused on classifying longer pieces of texts, such as movie reviews [145].

As a medium, Twitter presents many linguistic and communicative peculiarities. A tweet, in fact, is a really short informal text (140 characters), in which the frequency of creative punctuation, emoticons, slang, specific terminology, abbreviations, links and hashtags is higher than in other domains and platforms.

Twitter users post messages from many different media, including their smartphones, and they “tweet” about a great variety of topics, unlike what can be observed in other sites, which appear to be tailored to a specific group of topics [64].

5.1 Description of the task

This chapter describes a system developed for the participation in the constrained run of the EVALITA 2014 SENTIment POLarity Classification Task (SENTIPOLC14 [14]) which has been extended with emotive features extracted from ItEM and FB-NEWS15.

The chapter is organized as follows: Section 5.2 describes the CoLing Lab system, starting from data preprocessing and annotation, to the adopted classification model. Section 5.3 reports on an additional set of features derived from Emotive VSMs built using the corpus FB-NEWS described in section 4.2. Section 5.5 shows the results obtained in EVALITA 2014 and compares such results with the additional models.

5.2 The ColingLab System

The CoLing Lab system for polarity classification of tweets includes the following three basic steps, that will be described in this section:

1. a preprocessing phase, aimed at separating linguistic and nonlinguistic elements in the target tweets and at the linguistic annotation of a "cleaned" version of the original texts;

2. a feature extraction phase, in which the relevant characteristics of the tweets are identified;

3. a classification phase, based on a Support Vector Machine (SVM) classifier with a linear kernel.

5.2.1 Preprocessing

The aim of the preprocessing phase is the identification of the linguistic and nonlinguistic elements in the tweets and their annotation.

While the preprocessing of nonlinguistic elements such as links and emoticons is limited to their identification and classification (cf. section 5.2.2), the treatment of the linguistic material required the development of a dedicated rule-based procedure, whose output is a normalized text that is subsequently feed to a pipeline of general-purpose linguistic annotation tools.

In details, the following rules have been applied in the linguistic preprocessing phase:

- Emphasis: tokens presenting repeated characters like *bastaaaa* “stooooop” are replaced by their most probable standardized forms (i.e. *basta* “stop”);

- Links and emoticons: they are identified and removed;

- Punctuation: linguistically irrelevant punctuation marks are removed;

- Usernames: the users cited in a tweet are identified and normalized by removing the @ symbol and capitalizing the entity name;

- Hashtags: they are identified and normalized by simply removing the # symbol;

The output of this phase are “linguistically-standardized” tweets, that are subsequently POS tagged with the Part-Of-Speech tagger described in [41] and dependency-parsed with the DeSR parser [6].

Figure 5.1 shows an example of the nonlinguistic elements identified in the preprocessing step and Figure 5.2 shows the same linguistically annotated tweet. In this example, the following rules have been applied: (i) The emoticon (`<3`) has been removed and marked before annotation; (ii) exclamation marks (`!!!!!!!!`) have been reduced to a single mark `!`; (iii) the word *bellissimaaaaa* “wonderfuuuul” has been replaced by its most probable standardized form *bellissima* “wonderful”; (iv) the hashtag `#love` has been replaced with the form "Love".

```
id_tweet = "<ID_TWEET>"
text = "Mi sono innamorata della canzone \""Mirror\"""" #love è
bellissimaaaaa!!!!!!!! <3 """
normalized_text = "Mi sono innamorata della canzone "Mirror". Love  è
bellissima!"
emoticons = "<3_p1"
hashtags = "#love"
emphasis = "bellissimaaaaaa"
fancy_punctuation = "!!!!!!!!"
```

Figure 5.1: *Example of nonlinguistic elements identified in the preprocessing step. Tweet adapted from Evalita 2014. The field "emoticons" shows the emoticon (<3) and its polarity (p1).*

```
<doc id="<ID_TWEET>" polarity="10" text = ""Mi sono innamorata della
canzone \""Mirror\""""
#love è bellissimaaaaa!!!!!!!! <3 ">

ID Token Lemma C-POS F-POS Morphosyntactic feats HEAD DEP
1 Mi mi P PC num=s|per=1|gen=n 3 clit
2 sono essere V VA num=s|per=1|mod=i|ten=p 3 aux
3 innamorata innamorare V V num=s|mod=p|gen=f 0 ROOT
4 della di E EA num=s|gen=f 3 comp
5 canzone canzone S S num=s|gen=f 4 prep
6 " " F FB _ 7 punc
7 Mirror Mirror S SP _ 5 mod
8 " " F FB _ 7 punc
9 . . F FS _ 3 punc
2 1 Love Love S SP _ 2 subj
2 è essere V V num=s|per=3|mod=i|ten=p 0 ROOT
3 bellissima bello A A num=s|gen=f 2 pred
4 ! ! F FS _ 2 punc
```

Figure 5.2: *Example of annotated tweet taken from Evalita 2014.*

5.2.2 Feature extraction

By exploiting the linguistic and non-linguistic annotations obtained in the preprocessing phase, a total of 1,239 features have been extracted to be feed to the classifier. The inventory of features can be organized into the five classes described in this subsection.

Lexical Features

Lexical features represent the occurrence of bad words or of words that are either highly emotional or highly polarized. Relevant lemmas were identified from two in-house built lexicons (cf. below), and from Sentix [15], a lexicon of sentiment-annotated Italian words. Lexical features include:

ItEM seeds: Lexicon of 347 highly emotional Italian words built by exploiting an online feature elicitation paradigm. Native speakers were requested to list nouns, adjectives and verbs that are strongly associated with the eight basic positive and negative emotions defined in [162]. The features are, for each emotion, the total count of strongly emotional tokens in each tweet.

Bad words lexicon: By exploiting an in house built lexicon of common Italian bad words4.3.2, we included, for each tweet, the frequency of the bad words belonging to the lexicon (a feature for each bad word), as well as the total amount of these lemmas.

Sentix: Sentix (Sentiment Italian Lexicon: [15]) is a lexicon for Sentiment Analysis in which 59,742 lemmas are annotated for their polarity and intensity, among other information. Polarity scores range from -1 (totally negative) to 1 (totally positive), while Intensity scores range from 0 (totally neutral) to 1 (totally polarized). Both these scores appear informative for the classification, so that we derived, for each lemma, a Combined score C_{score}:

$$C_{score} = Intensity * Polarity \tag{5.1}$$

on the basis of which the selected lemmas have been organized into the following five groups:

- strongly positives: $1 \leq C_{score} < 0.25$
- weakly positives:$0.25 \leq C_{score} < 0.125$
- neutrals: $0.125 \leq C_{score} \leq -0.125$
- weakly negatives: $-0.125 < C_{score} \leq -0.25$
- highly negatives: $-0.25 < C_{score} \leq -1$

Since Sentix relies on WordNet sense distinctions, it is not uncommon for a lemma to be associated with more than one $\langle$Intensity,Polarity$\rangle$ pair, and consequently to more than one C_{score}.

In order to handle this phenomenon, the lemmas have been splitted into three different ambiguity classes: Lemmas with only one entry or whose entries are all associated with the same C_{score} value, are marked as "Unambiguous" and associated with their C_{score}.

Ambiguous cases were treated by inspecting, for each lemma, the distribution of the associated C_{scores}: Lemmas which had a Majority Vote (MV) were marked as "Inferable" and associated with the $C_{score}$of the MV. If there was no MV, but the highest number of senses in Sentix occurred simultaneously in both the positive or negative groups, lemmas were marked as “Inferable” and associated with the mean of the C_{scores}. All other cases were marked as "Ambiguous" and associated with the mean of the C_{scores}.

To isolate a reliable set of polarized words, we focused only on the “Unambiguous” or “Inferable” lemmas and selected only the 250 topmost frequent according to the PAISÀ corpus [111], a large collection of Italian web texts.

Other Sentix-based features in the ColingLab model are: the number of tokens for each C_{score} group, the C_{score} of the first token in the tweet, the C_{score} of the last token in the tweet and the count of lemmas that are represented in Sentix.

Negation

Negation features have been developed to encode the presence of a negation and the morphosyntactic characteristics of its scope.

The inventory of negative lemmas (e.g. "non") and patterns (e.g. "non…mai") have been extracted from [168]. The occurrences of these lemmas and structures have been counted an inserted as features to feed the classifier.

In order to characterize the scope of each negation, we used the dependency parsed tweets produced by DeSR [6]. The scope of a negative element is assumed to be its syntactic head or the predicative complement of its head, in the case the latter is a copula. Although it is clearly a simplifying assumption, the preliminary experiments show that this could be a rather cost-effective strategy in the analysis of linguistically simple texts like tweets.

This information has been included in the model by counting the number of negation patterns encountered in each tweet, where a negation pattern is composed by the PoS of the negated element plus the number of negative tokens depending from it and, in case it is covered by Sentix, either its Polarity, its Intensity and its C_{scores} value.

For instance, the negation pattern instantiated in the phrase *non tornerò mai* “I will never come back” has been encoded, as "neg-negVPOSPOL", "neg-negVHIGHINT" and "neg-negVPOSCOMB", meaning that a verb with high positive polarity, high intensity and a high C_{scores} is modified by two negative tokens.

Morphological features

The linguistic annotation produced in the preprocessing phase has been exploited also in the population of the following morphological statistics:

- number of sentences in the tweet;
- number of linguistic tokens;
- proportion of content words (nouns, adjectives, verbs and adverbs);
- number of tokens for Part-of-Speech.

Shallow features

This group of features has been developed to describe some distinctive characteristic of the web communication. The group includes:

Emoticons: We used the lexicon LexEmo (cf. section 4.3.1) to mark the most common emoticons, such as `:-(` and `:-)`, marked with their polarity score: 1 (positive), -1 (negative), 0 (neutral).

LexEmo is used both to identify emoticons and to annotate their polarity.
Emoticon-related features are the total amount of emoticons in the tweet, the polarity of each emoticon in sequential order and the polarity of each emoticon in reversed order. For instance, in the tweet *:-(quando ci vediamo? mi manchi anche tu! :*:** “:-(when are we going to meet up? I miss you, too :*:*” there are three emoticons, the first of which (`:-(`) is negative while the others are positive (`:*; :*`).
Accordingly, the classifier has been fed with the information that the polarity of the first emoticon is -1, that of the second emoticon is 1 and the same goes for the third emoticon. At the same way, another group of feature specifies that the polarity of the last emoticon is 1, as it goes for that of the last but one emoticon, while the last but two has a polarity score of -1.

Links: These features contain a shallow classification of links performed using simple regular expressions applied to URLs.
In particular, links are classified as following: `video`, `images`, `social` and `other`.
For example, URLs containing substrings such as `youtube.com` or `twitcam` are classified as `video`.
Similarly URLs containing substrings such as `imageshack`, or `jpeg` are classified as “images”, and URLs containing `plus.google` or `facebook.com` are classified as “social”. Unknown links are inserted in the residual class “other”. We also use as feature the absolute number of links for each tweet.

Emphasis: The features report on the number of emphasized tokens presenting repeated characters like *bastaaaa*, the average number of repeated characters in the tweet, and the cumulative number of repeated characters in the tweet.
For instance, in the message *Bastaaa! Sono stufaaaaaaaaa* “Stop! I have had enough”, there are 2 empathized tokens, the average number of repeated characters is 5, and the cumulative number of repetitions is 10.

Creative Punctuation: Sequences of contiguous punctuation characters, like “`!!!`”, “`!?!?!?!!?!????!`” or “`.......`”, are identified and classified as a sequence of dots, exclamations marks, question marks or mixed.
For each tweet, the features correspond to the number of sequences belonging to each group and their average length in characters.

Quotes: The number of quotations in the tweet.

Twitter features

This group of features describes some Twitter-specific characteristics of the target tweets.

Topic: This information marks if a tweet has been retrieved via a specific political hashtag or keywords. It is provided by organizers as an attribute of the tweet;

Usernames: The number of @username in the tweet;

Hashtags: This group of features is aimed at inferring the polarity of an hashtag by generalizing over the polarity of the tweets in the same thread. In other words, the terms associated with a target hashtag t are expanded by using t as a search key to download the most recent tweets in which it occurs.
The polarity of the retrieved tweets is calculated by simply counting the number of positive and negative words in them.
In this way, the polarity of the hashtag t is assumed to be the same of the words it typically co-occurs with.
This, of course, does not take into account any kind of contextual variability of words meaning and is an oversimplifying assumption; nevertheless, in most cases, the polarity of the hashtag will reflect the polarity of its typical word contexts.
Moreover, tweets were assumed to be positive if they contained a majority of positive words, negative if they contained a majority of negative words, neutral otherwise. In these features, the polarity of a word, is inferred from Sentix: Words with a positive score ≤ 0.7 got a score of 1, while words with a negative score ≤ -0.7 received the score of -1. All the other words got a score of 0 (neutrality).
The hashtags with a small number of tweets (< 20) were filtered, leaving us with 279 polarity-marked hashtags.
By relying on this hashtag-to-polarity mapping, the hashtag-related features in th model consisted in the total amount of hashtag for tweet, the polarity of each hashtag in sequential order and the polarity of each hashtag in reversed order.

Table 5.1 shows the groups of features used in three different configurations of the ColingLab system. The Full model is the model that has been submitted for the evaluation, Lexical is the model containing only lexical features such as ItEM, Sentix anf Badwords and the Shallow model is based on non-linguistic features. The table also provides the number of features extracted for each group.

5.3 Introducing emotive features

In order to add emotive features to the previous model, several emotive spaces have been created from FB-NEWS15 following the strategy illustrated in section 4.1.

In particular, the list T of the 30,000 most frequent nouns, verbs and adjectives has been extracted from FB-NEWS15 (crf. section 4.2). The lemmas in T were subsequently used as target and contexts in a square matrix of co-occurrences extracted within a five word window (± 2 words, centered on the target lemma, before removing the words not belonging to T). In addition, we extended the matrix to the nouns, adjectives and verbs in the corpus of tweets (i. e. lemmas not belonging to T).

For each $\langle$emotion, PoS$\rangle$ pair we built a centroid vector from the vectors of the seeds belonging to that emotion and PoS, obtaining in total 24 centroids. We constructed different word spaces according to PoS because the context that best captures the meaning of a word, differs depending on the word to be represented [173].

Group	Features	N. of features	Full	Lexical	Shallow
Lexical	Badwords	28	*	*	
	ItEM	9	*	*	
	Sentix	1023	*	*	
Negation	Negation	53	*	*	
Morph. features	Morph. features	18	*		*
Shallow	Emoticons	17	*		*
	Emphasis	3	*		*
	Links	5	*		*
	Punctuation	6	*		*
	Quotes	1	*		*
	Slang	10	*		*
Twitter	Hashtags	63	*	*	
	Topic	1	*		*
	Usernames	2	*		*
Total number of features		**1239**	**1239**	**1176**	**63**

Table 5.1: *Features used in EVALITA 2014 to train the models.*

Following the configuration in [151, 152], the co-occurrence matrix has been re-weighted using the Pointwise Mutual Information [31] as in 2.4, and in particular the Positive PMI (PPMI), in which negative scores are changed to zero [136].

Starting from these spaces, several groups of features have been extracted. The simplest ones include general statistics such as the number of emotive words and the emotive score of a tweet. More sophisticated features are aimed at inferring the degree of distinctivity of a word as well as its polarity from their own emotive profile.

Number of emotive words: Words belonging to the emotive Facebook spaces;

Emotive/words ratio: The ratio between the number of emotive words and the total number of words in the tweet;

Strongly emotive words: Number of words having a high (greater than 0.4) emotive score for at least one emotion;

Tweet emotive score: Score calculated as the ratio between the number of strongly polarized words and the number of the content words in the tweet (cf. Equation 5.2). The feature assumes values in the interval [0, 1]. In absence of strongly emotive words, the default value is 0.

$$Emotivity(Tweet) = \frac{Count(Strongly\ emotive\ words)}{Count(Content\ words)} \tag{5.2}$$

Maximum values: The maximum emotive value for each emotion (8 features);

Quartiles: The features take into account the distribution of the emotive words in the tweet. For each emotion, the list of the emotive words has been ordered according to the emotive scores and divided into quartiles (e. g. the fourth quartile contains the most emotive words and the first quartile the less emotive ones.). Each feature registers the count of the words belonging to the pair $\langle emotion, quartile \rangle$ (32 features in total);

ItEM seeds Boolean features registering the presence of words belonging to the words used as seeds to build the vector space models. In particular, the features include the top 4 frequent words for each emotion (32 boolean features in total);

Distintive words: 32 features corresponding to the top 4 distinctive words for each emotion. The degree of distinctivity of a word for a given emotion is calculated starting from the VSM normalized using Z-scores. In particular, the feature corresponds to the proportion of the emotion $\langle emotion_i \rangle$ against the sum of total emotion score $[e_1, ..., e_8]$;

Polarity (count): The number of positive and negative words. The polarity of a word is calculated by applying the formula 5.3, in which positive emotions are assumed to be JOY and TRUST, and negative emotions are assumed to be DISGUST, FEAR, ANGER and SADNESS.

$$Polarity(w) = \frac{\text{JOY+TRUST}}{2} - \frac{\text{DISGUST+FEAR+ANGER+SADNESS}}{4} \quad (5.3)$$

Polarity (values): The polarity (calculated using 5.3) of the emotive words in the tweet. The maximum number of emotive words is assumed to be 20.

5.4 Classification

Due to the better performance of SVM-based systems in analogue tasks (e.g. [130]), we chose to base the CoLing Lab system for polarity classification on the SVM classifier with a linear kernel implementation available in Weka [220], trained with the Sequential Minimal Optimization (SMO) algorithm introduced by Platt [158].

The classification task proposed by the organizers could be approached either by building two separate binary classifiers relying of two different models (one judging the positiveness of the tweet, the other judging its negativeness), or by developing a single multiclass classifier where the possible outcomes are Positive Polarity (Task POS:1, Task NEG:0), Negative Polarity (Task POS:0, Task NEG:1), Mixed Polarity (Task POS:1, Task NEG:1) and No Polarity (Task POS:0, Task NEG:0).

In Evalita 2014 [149] we tried both approaches in our development phase, and found no significant difference, so that we opted for the more economical setting, i.e. the multiclass one.

For sake of comparison, the same paradigm has been maintained to compare the system with models based on the Coling Lab system, but including distributional emotive features.

5.5 Results and comparison

In Evalita 2014, the submitted model obtained a macro-averaged F1-score of 0.6312 on the test set (cf. section 5.2).

Besides the submitted model, two additional configurations were presented in order to measure the impact of nonlinguistic and lexical elements for the classification.

A Shallow model includes non-linguistic features such as emoticons, creative punctuation etc., as well as a Lexical model which includes different features based on external resources such as Sentix or a list of bad words.

The best model to predict the polarity of a tweet was the one combining lexical and shallow information (Full model).

Task	Class	Precision	Recall	F-score
POS	0	**0.7976**	0.7806	0.789
POS	1	0.581	**0.4109**	**0.4814**
POS task		0.6893	**0.5957**	**0.6352**
NEG	0	0.6923	0.6701	**0.681**
NEG	1	**0.6384**	0.5201	0.5732
NEG task		**0.6654**	**0.5951**	**0.6271**
GLOBAL		0.6774	**0.5954**	**0.6312**

Table 5.2: *CoLing Lab system results.*

Even though the Full model achieved a better F1-score, the global precision of the Shallow model was higher than the precision of the Full one, despite the much smaller numbers of features. In particular, the Shallow model recognized positive tweet more accurately.

It is worth noticing that the class of positive tweets was the one in which our systems scored worst. Besides the fact that the tweet class distribution was unbalanced in the training corpus, positive lexical features were likely to be not as able to predict tweets positivity, as negative features were with respect to negative tweets.

Task	Class	Precision	Recall	F-score
POS	0	0.7599	0.7755	0.7676
POS	1	0.4913	0.2981	0.371
POS task		0.6256	0.5368	0.5693
NEG	0	0.66	**0.6861**	0.6728
NEG	1	0.6218	0.4522	0.5237
NEG task		0.6409	0.5692	0.5983
GLOBAL		0.6333	0.553	0.5838

Table 5.3: *CoLing Lab Lexical system results.*

Task	Class	Precision	Recall	F-score
POS	0	0.7578	**0.8679**	**0.8092**
POS	1	**0.7184**	0.2205	0.3374
POS task		**0.7381**	0.5442	0.5733
NEG	0	**0.7369**	0.5174	0.608
NEG	1	0.5778	**0.6582**	**0.6154**
NEG task		0.6574	0.5878	0.6117
GLOBAL		**0.6978**	0.566	0.5925

Table 5.4: *CoLing Lab Shallow system results.*

The comparison between the three experiments demonstrated on the one hand that significant improvements can be obtained by using lexical information. On the other hand the preliminary results highlighted that the lexical coverage of the available resources such as Sentix and the seeds of ItEM had to be increased in order to obtain a more accurate classification. In order to improve the system results, we focused in particular on the recognition of positive tweets, which was the worst performing class (i.e. Task POS:1, Task NEG:0). In the second version of the system, all the emotive features were added, including lexical and statistic ones.

Starting from the Shallow model as a baseline, introducing emotive features allow the system to improve its performance especially by considering the recall of the positive items. Table 5.5 shows the comparison between such baseline and the improved model.

Task	Class	Precision	Recall	F-score
POS	0	+0.0256	-0.0374	-0.0029
POS	1	-0.0632	+0.1146	+0.106
POS task				+0.0515
NEG	0	-0.0019	+0.1282	+0.0794
NEG	1	+0.071	-0.0587	+0.0078
NEG task				+0.0436
GLOBAL				+0.0476

Table 5.5: *CoLing Lab Shallow + Emotive features.*

Task	Class	Precision	Recall	F-score
POS	0	**0.7834**	0.8305	0.8063
POS	1	0.6552	**0.3351**	**0.4434**
POS task				**0.6248**
NEG	0	0.735	**0.6456**	**0.6874**
NEG	1	**0.6488**	0.5995	**0.6232**
NEG task				**0.6553**
GLOBAL				**0.6401**

Table 5.6: *CoLing Lab Shallow + Emotive features. System results.*

Finally, as a last experiment relating to sentiment polarity classification, a full model has been created starting from shallow, lexical and emotive features. Table 5.7 shows the results of this model, after feature ablation. The feature ablation experiments, in fact, demonstrated a modest improvement in classification accuracy when the distinctivity features are dropped. In this case, the recall for the positive tweet is increased to 0.44. In this case (Lexical + Emotive), the emotive features have been extracted from the z-score normalized VSM, and such features have been joined to the lexical and shallow ones used in the original CoLingLab (Full) system.

Although from a computational point of view this experiment is not preferable since it reaches the little overall gain compared to the model Shallow + Emotive against a much larger number of features, from a linguistic point of view it proves once again that improving the coverage of the lexical resources helps in obtaining better performance in sentiment polarity classification tasks.

Task	Class	Precision	Recall	F-score
POS	0	0.8059	0.7674	0.7862
POS	1	0.5744	0.4427	0.5
POS TASK				0.6431
NEG	0	0.717	0.6758	0.6958
NEG	1	0.6599	0.5581	0.6047
NEG TASK				0.6503
GLOBAL				0.6467

Table 5.7: *CoLing Lab Shallow + Lexical + Emotive features. System results.*

5.6 Summary

The chapter describes the Coling Lab system, which has been developed for the participation in the constrained run of the EVALITA 2014 SENTIment POLarity Classification Task (SENTIPOLC [14]). The system includes three main steps being the data preprocessing and annotation, feature extraction and the classification model.

The Coling Lab system has been enriched with emotive features in its second step (i. e. feature extraction). In order to assure a high coverage, such features have been derived from emotive resources built on the social media corpus FB-NEWS15.

The chapter makes a comparison between the original Coling Lab system and the enriched one, showing that the emotive connotation of the words of a tweet helps in determine its general polarity.

CHAPTER 6

Conclusion and future perspectives

In this chapter, we provide a summary of the techniques and methodologies presented in this dissertation and the conclusions we reached. In this experimental work some benefits of using distributional methods to deal with affective systems emerged, but there are still a number of issues that require consideration. Since in every study there is room for error, extension and improvement, we will finally discuss several ideas for future work.

6.1 Contributions

In this work we first presented a review of different state-of-the-art methods for sentiment analysis and emotion detection, focusing on the existing lexicons and on the techniques that can be used to create new ones. We showed that sentiment lexicons can be induced using several methods including (i) rule based approaches; (ii) corpus-based approaches such as the Vector Space Models; (iii) existing linguistic resources like Wordnet; (iv) word embeddings; (v) hybrid approaches. We moreover reviewed some of the strategies that can be used to classify opinionated texts on the basis of their polarity and emotions.

After a brief description of the NLP tools used for the linguistic annotation of the material presented in this thesis, we presented ItEM, which is a high-coverage Italian emotive lexicon that has been created by exploiting distributional methods. ItEM has been built by means of a process that articulated in three steps: the collection of a set of highly emotive words, their distributional expansion and the validation of the complete system. Since corpus-based methods reflect the type of the corpus from which they are built, in order to create a reliable lexicon we collected a new Italian corpus, namely FB-NEWS15. This collection has been created by crawling the Facebook pages of the most important Italian newspapers, which typically include a small number of posts written by the journalists and a very high number of comments inspired by long discussions among readers about such news.

Finally, we described some experiments concerning the sentiment polarity classification of tweets. We started from a system based on supervised learning that was originally developed for the Evalita 2014 SENTIment POLarity Classification task [14] and subsequently explored the possibility to enrich this system by exploiting lexical emotive features derived from social media texts. The experiment demonstrated that the emotive connotation of the words in a tweet may be exploited to determine its general polarity, especially if we consider the recognition of positive tweets.

6.2 Evaluation and results

Although research has tackled many Sentiment Analysis subproblems and proposed a large number of solutions, none of such subproblems has been solved satisfactorily and none of these approaches became a standard [109]. In addition, moving to Emotion Detection, the problem become even more complex because on the one hand it is very difficult to establish what an emotion is, and on the other one, the lack of annotated resources comparable in source and size makes it difficult to establish a benchmark for the level of accuracy required from this kind of resources to be used in Sentiment Analysis systems.

Moreover, many researchers perform studies and publish material about their work, but very often it's difficult to replicate, validate, and improve the proposed methods and results. This is crucial in dealing with the evaluation of systems that work on subjective information. One of the most challenging issues in this work has been the evaluation of the resources created using distributional methods. In our opinion, by making the research available for others, its impact can potentially increase. For these reasons, we plan to make publicly available selected material from this thesis, including the small resources used to build the extended ones (i. e. the results of the feature elicitation and the words annotated using crowdsourcing).

As a first output of this research, we evaluated the emotive lexicon ItEM. We performed two experiments aimed to establish (i) if the top ranked emotive words are recognized as emotive words by human annotators; (ii) if, given a random word, the emotive values assigned to that word by the system are in line with the degree of emotivity assigned by human raters. Our system reaches an overall precision of 68% on the top 50 candidates and 43% of F1-score on a random sample of elements (including neutral).

It is clear that additional value can be achieved if research methods are applied in real-life situations such as monitoring of sentiment and emotions in social media platforms like Twitter, Facebook or Instagram. By partecipating to SENTIPOLC [14], we tested (and in part confirmed) the hypothesis that the developed sentiment analysis methodology mostly based on lexical resources is useful in real-world applications. However, Twitter tweets often contain non-standard language and they are limited to 140 characters, and these restrictions can pose difficulties for lexicon-based systems.

Compared to other social network messages, tweets seem to be the less likely to contain emotive lexical items, so that we expect our system to perform better in other social media like Facebook.

6.3 Lessons Learned

We performed a number of experiments to select the most appropriate approach to build an emotive lexicon, including weighting parameters, context selection, evaluation etc. Some of the experiments provided acceptable performance results and some of them were less satisfying.

What we learned during this process is that there are some good and bad practices in the tuning of the parameters that influence the results. In the following list we summarize practical issues concerning to the construction of a distributional model for emotions.

Choosing a taxonomy for emotions: The longstanding dispute concerning whether emotions are better conceptualized in a discrete taxonomy, produced a long list of different theories on emotion classification. As in other works [122, 123], we borrowed our emotions inventory from Plutchik [161], who distinguishes eight *basic* human emotions: JOY, SADNESS, FEAR, DISGUST, SURPRISE, ANGER, TRUST and ANTICIPATION. Positive characteristics of this classification include the relative low number of distinctions encoded, as well as its being balanced with respect to positive and negative feelings. In addition, this taxonomy is a superset of the Ekman [52]'s one, which has been one of the most influencing taxonomies in the fields of Computational Linguistics and Facial Recognition.

Building Vector Space Models: The performance of a VSM is highly influenced by the hyperparameters setup chosen in the implementation phase, such as the corpus, the definition of the context, the weighting schemes and the techniques used to smooth the matrix. In order to build the space, we extracted the list of the 30,000 most frequent nouns, verbs and adjectives from general purpose corpora [10, 11] and from the social media corpus FB-NEWS15. The lemmas belonging to that list were then used as target and contexts in a square matrix of co-occurrences extracted within a five word window (± 2 words, centered on the target lemma). Raw frequencies were weighted using PPMI that is typically credited for granting the best performances across several tasks [27, 58, 207]. Moreover, the experiments showed [152] that context selection is useful to handle with Emotive VSMs.

Evaluation: We choose to take advantage of crowdsourcing to evaluate our lexicon. First, we collected a set of highly emotive words by asking 60 native speakers to provide 15 words (5 for each PoS of interest) for each emotion in our taxonomy. After the distributional expansion using VSMs, we used crowdsourcing to confirm that the top ranked emotive words were recognized as emotive words by human annotators.

Crowdsourcing: In order to prepare the crowdsourcing experiments presented in this thesis we run several experiments to ensure that the task was understood and perceived as meaningful. We learned that to design a crowdsourcing task the simpler solution is the better, so that it can be useful to: (i) focus on one word per question; (ii) assign the complete annotation of a word to the annotator (i. e. the annotator decides whether a word expresses [EMOTION 1, EMOTION 2, ..., EMOTION N]); (iii) add null options (i. e. to avoid random answers); (iv) select the proper number of subjects; (v) don't mix the tasks (i. e. don't evaluate polarity and emotions at the same moment).

6.4 Building an Emotive Lexicon for a new language: a (simple) recipe

We now briefly touch upon a recipe to create an Emotive lexicon for a new language. The recipe ingredients can be combined to create an original emotive lexicon. Certain items are necessary to start the process, but multiple ingredients can be used and handled to increase the effectiveness of the result.

6.4.1 Recipe Ingredients

- A PoS-Tagged corpus C;
- A taxonomy T_E of emotions;
- A small set of seed words S highly associated with the emotions in T_E (i.e. for each emotion E in T there are $\{w1, w2, ..., wn\}$ words associated wih E);

6.4.2 Recipe Procedure

1. Define the words to be included in the lexicon (i.e. the top-X words of C sorted by frequency);
2. Generate a matrix of co-occurrences. A row in a co-occurrences matrix corresponds to a word: a certain word occurs in a particular context a certain number of times (frequency).
3. Adjust the weights of the words in the matrix, because common words will have high frequencies, yet they are less informative than rare words [207].
4. Smooth the matrix to reduce the noise (i. e. using SVD [100]).
5. Isolate the vectors of the seed words and create the centroid vectors from them according with T_E.
6. Measure the similarity (e.g. the cosine similarity) between the target words and the centroid vectors to obtain the degree of association between the word and the emotions in T_E.

6.4.3 Recipe customization

As for recipe ingredients, the selection of the corpus affects the final lexicon, so that it is important to choose the proper one. As an alternative, a new corpus can be collected, for example from social media. Moreover, selecting the right taxonomy of emotions could be a hard task. Since for each emotion it is necessary to collect highly distinctive seed words, it's wise to use a taxomomy with a relative small set of emotions, with clear and accepted definition.

As for the procedure, at step 1, it is possible to define the targets of the matrix starting from white lists, or other criteria. At steps 2-4, the definition of the context (2), the weighting schemes (3) and the techniques used to smooth the matrix (4) greatly influence the final result. At step 6, different similarity metrics can be applied, such as the scalar product (dot product), or the euclidean similarity.

6.5 Open issues and further research

As widely discussed throughout these pages, a strong point of this distributional approach to create emotive lexicons is its own modularity. That is, in its possibility to modify one of its sub-modules in order to adapt the instrument to any task or domain.

A first important parameter for the creation of a corpus-based lexicon is of course the corpus from which the statistics are extracted. This point has been explored by using a corpus of Facebook posts instead of a general purpose corpus to induce the lexicon. An interesting follow-up study would be the analysis of the differences among the Social Networks, for example by extending FB-NEWS15 with the same news taken from other social media like Twitter. Similarly, it would be interesting to see if we can we improve the results in tweet polarity classification by extracting the emotive features from a corpus made entirely of tweets downloaded in real-time. Another option we didn't explore could be to replace the original seeds with existing lexical resources [194] annotated for emotions.

Moreover, recent researches showed that word embeddings are able to find related concepts and analogies, and they can become very useful if used as features in standard machine learning algorithms [189, 199]. It would then be interesting to evaluate the effects of representing word vectors by exploiting deep learning instead of explicit count-based models during the bootstrapping phase.

Another challenge associated with polarity and emotion analysis is the use of context information to determine the real emotional impact of a word in a particular sentence. To recognize emotions in text, in fact, it is crucial to go beyond the word level, because emotions are often hidden behind combinations of words [44, 127]. For the future, we plan to study how emotions are expressed and perceived through collocations, idioms and metaphorical expressions.

Moreover, the meaning of a sentence cannot be properly described as the sum of the affective values carried by the open-class lexical items [9, 120]. Combinatorial phenomena such as modification, negation and recursive construction radically influence the sentiment content of a sentence. To handle with these lexical phenomena, future work will be addressed to study effective solutions for vector compositionality.

Finally, future work should construct benchmark datasets to allow the methods (or their improved versions) to be compared each other and with the state of the art.

Bibliography

[1] A. Ahmad. Is twitter a useful tool for journalists? *Journal of Media Practice*, 11(2):145–155, 2010.

[2] E. C. Ovesdotter Alm. *Affect in text and speech*. ProQuest, 2008.

[3] M. B. Arnold. Emotion and personality. *Columbia University Press*, 1960.

[4] G. Attardi. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 166–170, New York City, New York (USA), 2006. Association for Computational Linguistics.

[5] G. Attardi and F. Dell'Orletta. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 261–264, Boulder, Colorado (USA), 2009. Association for Computational Linguistics.

[6] G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. Accurate dependency parsing with a stacked multilayer perceptron. In *EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy), 2009. Springer.

[7] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), 2010.

[8] F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, and V. Patti. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In *Proceedings of EVALITA 2016 – Evaluation of NLP and Speech Tools for Italian*, Napoli (Italy), 2016. aAcademia University Press.

[9] M. Baroni. Composition in distributional semantics. *Language and Linguistics Compass*, 7(10):511–522, 2013.

[10] M. Baroni, S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, and M. Mazzoleni. Introducing the la repubblica corpus: A large, annotated, tei(xml)-compliant corpus of newspaper italian. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1171–1174, Lisbon (Portugal), 2004. European Language Resource Association (ELRA).

[11] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–222, 2009.

[12] M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

[13] P. Basile and N. Novielli. UNIBA at EVALITA 2014-SENTIPOLC Task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, pages 58–63, Pisa (Italy), 2014.

[14] V. Basile, A. Bolioli, M. Nissim, V. Patti, and P. Rosso. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa (Italy), 2014.

[15] V. Basile and M. Nissim. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, 2013.

[16] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[17] S. Blair-Goldensohn, T. Neylon, K. Hannan, G. A. Reis, R. Mcdonald, and J. Reynar. Building a sentiment summarizer for local service reviews. In *NLP in the Information Explosion Era*, 2008.

[18] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[20] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 19–26, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[21] A. Blum, J. Lafferty, M. R. Rwebangira, and R. Reddy. Semi-supervised learning using randomized mincuts. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, page 13, Banff, Alberta (Canada), 2004. ACM.

[22] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[23] F. Bonin, F. Dell'Orletta, S. Montemagni, and G. Venturi. A contrastive approach to multi-word extraction from domain-specific corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta (Malta), 2010. European Language Resources Association (ELRA).

[24] C. Bosco, V. Patti, and A. Bolioli. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63, 2013.

[25] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings, 1999.

[26] T. Brants and A. Franz. Web 1t 5-gram version 1, 2006. LDC2006T13.

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

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

[29] E. Cambria and A. Hussain. Sentic computing techniques, tools, and applications 2nd edition, 2012.

[30] E. Cambria, A. Livingstone, and A. Hussain. The hourglass of emotions. In *Proceedings of the 2011 International Conference on Cognitive Behavioural Systems*, COST'11, pages 144–157. Springer Berlin Heidelberg, 2011.

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

[32] H. Chyi and D. L. Lasorsa. An explorative study on the market relation between online and print newspapers. *Journal of Media Economics*, 15(2):91–106, 2002.

[33] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[34] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[35] R. Cozma and K-J. Chen. What's in a tweet? foreign correspondents' use of social media. *Journalism Practice*, 7(1):33–46, 2013.

[36] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 1999.

[37] C. Darwin. *The Expressions of the Emotions in Man and Animals.* John Murray, 1872.

[38] A. Das and S. Bandyopadhyay. Towards the global sentiwordnet. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 799–808, Sendai (Japan), 2010.

[39] S. Dasgupta and V. Ng. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2 of *ACL '09*, pages 701–709, Suntec (Singapore), 2009. Association for Computational Linguistics.

[40] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528, Budapest (Hungary), 2003.

[41] F. Dell'Orletta. Ensemble system for part-of-speech tagging. In *EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy), 2009. Springer.

[42] F. Dell'orletta, G. Venturi, A. Cimino, and S. Montemagni. $t2k^2$: a system for automatically extracting and organizing knowledge from texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik (Iceland), 2014.

[43] F. Dell'Orletta, A. Lenci, S. Marchi, S. Montemagni, V. Pirrelli, and G. Venturi. Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio. In *Atti del Convegno Nazionale Associazione Italiana di Terminologia "I-TerAnDo", Università della Calabria*, pages 185–206, 2008.

[44] D. Derks, A. H. Fischer, and A. E. R. Bos. The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, 24(3):766–785, 2008.

[45] P. S. Dodds and C. M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010.

[46] L. Dong, F. Wei, C. Tan, D. Tang, , M. Zhou, and K. Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore (Maryland), 2014. Association for Computational Linguistics.

[47] E. C. Dragut, C. Yu, P. Sistla, and W. Meng. Construction of a sentimental word dictionary. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1761–1764, Toronto (Canada), 2010. ACM.

[48] S. T. Dumais. Enhancing performance in latent semantic indexing (lsi) retrieval. Technical Report, 1990. TM-ARH-017527.

[49] S. T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[50] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[51] J. Bernard Ed., editor. *The Macquarie Thesaurus*. Macquarie Library, 1986.

[52] P. Ekman, W. V. Friesen, and P. Ellsworth. What emotion categories or dimensions can observers judge from facial behavior? *Emotion in the Human Face*, pages 98–110, 1982.

[53] P. Ekman, W. V. Friesen, and P. Ellsworth. Affect theory. *Approaches to Emotion*, pages 163–195, 1984.

[54] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 617–624, Melbourne (Australia), 2005. ACM.

[55] A. Esuli and F. Sebastiani. Determining term subjectivity and term orientation for opinion mining. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, Trento (Italy), 2006. Association for Computational Linguistics.

[56] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 417–422, Genoa (Italy), 2006. European Language Resource Association (ELRA).

[57] S. Evert. *Corpora and collocations*, pages 1212–1248. Lüdeling, A. and Kytö, M., Berlin (Germany), 2008.

[58] C. Fabre and A. Lenci. Distributional Semantics Today Introduction to the special issue. *Traitement Automatique des Langues*, 52(2):7–20, 2015.

[59] I. Hernandez Farias, D. Buscaldi, and B. Priego Sanchez. IRADABE: Adapting English Lexicons to the Italian Sentiment Polarity Classification task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, pages 75–80, Pisa (Italy), 2014.

[60] C. Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, 1998.

[61] S. Feng, L. Zhang, B. Li, D. Wang, G. Yu, and K. Wong. Is twitter a better corpus for measuring sentiment similarity? In *EMNLP*, pages 897–902. ACL, 2013.

[62] K. T. Frantzi and S. Ananiadou. The c-value/nc-value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179, 1999.

[63] Gamon and Michael. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Geneva (Switzerland), 2004. Association for Computational Linguistics.

[64] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.

[65] J. A. Gray. *The Neuropsychology of Anxiety*. Oxford University Press, 1982.

[66] J. J. Gross. The future's so bright, i gotta wear shades. *Emotion Review*, 2(3):212–216, 2010.

[67] M. Guerini, L. Gatti, and M. Turchi. Sentiment analysis: How to derive prior polarities from sentiwordnet. *Computing Research Repository (CoRR)*, abs/1309.5843, 2013.

[68] M. Hagen, M. Potthast, M. Büchner, and B. Stein. Webis: An ensemble for twitter sentiment detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 582–589, Denver, Colorado (USA), 2015. Association for Computational Linguistics.

[69] H. Hamdan, P. Bello, and F. Bechet. Lsislif: Feature extraction and label weighting for sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 568–573, Denver, Colorado (USA), 2015. Association for Computational Linguistics.

[70] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. *Computing Research Repository (CoRR)*, abs/1606.02820, 2016.

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

[72] A. Hassan, V. Qazvinian, and D. Radev. What's with the attitude?: Identifying sentences with attitude in online discussions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1245–1255, Cambridge, Massachusetts (USA), 2010. Association for Computational Linguistics.

[73] V. Hatzivassiloglou and K.R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, EACL '97, pages 174–181, Madrid, Spain, 1997. Association for Computational Linguistics.

[74] Y. He. Learning sentiment classification model from labeled features. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1685–1688, Toronto (Canada), 2010.

[75] A. Hermida. #journalism. reconfiguring journalism research about twitter, one tweet at a time. *Digital Journalism*, 2013.

[76] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, Seattle, WA (USA), 2004. ACM.

[77] E. H. Huang, R. Socher, C. D. Manning A. Y., and Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Jeju Island (Korea), 2012. Association for Computational Linguistics.

[78] Emotion annotation and representation language (earl). http://emotion-research.net/projects/humaine/earl/proposal#DescriptiveSchemes, 2006. Accessed: 2016-06-14.

[79] C. E. Izard. *The Face of Emotion.* Appleton-Century-Crofts, 1971.

[80] P. Jaccard. *Nouvelles recherches sur la distribution florale*. Bulletin de la Société vaudoise des sciences naturelles. Impr. Réunies, 1908.

[81] R. Jackendoff. *The Architecture of the Language Faculty*, volume 28. MIT Press, Cambridge, MA, 1997.

[82] C. Jacquemin. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale*. Université de Nantes, 1997.

[83] W. James. What is an emotion? *Mind Association*, 9(34):188–205, 1884.

[84] T. Joachims. *Advances in Kernel Methods*, chapter Making Large-scale Support Vector Machine Learning Practical, pages 169–184. MIT Press, Cambridge, MA (USA), 1999.

[85] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[86] M. Joshi and C. Penstein-Rosé. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 313–316, Suntec (Singapore), 2009. Association for Computational Linguistics.

[87] D. Jurafsky and J. H. Martin. Speech and language processing (3rd ed. draft). Retrieved on September, 20 2016.

[88] N. Kaji and M. Kitsuregawa. Automatic construction of polarity-tagged corpus from html documents. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 452–459, Sydney (Australia), 2006. Association for Computational Linguistics.

[89] N. Kaji and M. Kitsuregawa. Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083, Prague (Czech Republic), 2007. Association for Computational Linguistics.

[90] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1115–1118, 2004.

[91] H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 355–363, Sydney (Australia), 2006. Association for Computational Linguistics.

[92] A. Kennedy and D. Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

[93] F. Keshtkar and D. Inkpen. A hierarchical approach to mood classification in blogs. *Natural Language Engineering*, 18(1):61–81, 2011.

[94] J. Kim, J. Li, and J. Lee. Discovering the discriminative views: Measuring term weights for sentiment analysis. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 253–261, Suntec (Singapore), 2009.

[95] S-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Geneva (Switzerland), 2004. Association for Computational Linguistics.

[96] S-M. Kim and E. Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 1–8, Sydney (Australia), 2006. Association for Computational Linguistics.

[97] S. Kiritchenko and S. Mohammad. Sentiment composition of words with opposing polarities. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California (USA), 2016.

[98] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin (Ireland), 2014. Association for Computational Linguistics and Dublin City University.

[99] H. Lakkaraju, R. Socher, and C. Manning. Aspect specific sentiment analysis using hierarchical deep learning. In *Proceedings of the NIPS Workshop on Deep Learning and Representation Learning*, 2014.

[100] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240, 1997.

[101] T. K. Landauer and N. Ross. *On the computational basis of learning and cognition: Arguments from LSA*, pages 43 – 84. Academic Press, 2002.

[102] R. S. Lazarus. On the primacy of cognition. *American Psychologist*, 39(2):124–129, 1984.

[103] R. S. Lazarus. *The cognition-emotion debate: A bit of history.* Guilford Press, 2000.

[104] A. Lenci, S. Montemagni, V. Pirrelli, and G. Venturi. Ontology learning from italian legal texts. In *Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, pages 75–94, Amsterdam (The Netherlands), 2009. IOS Press.

[105] H. Lin and L. Qiu. Two sites, two voices: Linguistic differences between facebook status updates and tweets. In P. L. Patrick Rau, editor, *Cross-Cultural Design. Cultural Differences in Everyday Life: 5th International Conference, CCD 2013, Held as Part of HCI International 2013*, volume 2, pages 432–440, Las Vegas (USA), 2013. Springer Berlin Heidelberg.

[106] K. H-Y. Lin, C. Yang, , and H-H. Chen. What emotions do news articles trigger in their readers? In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 733–734, Amsterdam (The Netherlands), 2007.

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

[108] B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[109] B. Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[110] H. Liu and P. Singh. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

[111] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, and V. Pirrelli. The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg (Sweden), 2014. Association for Computational Linguistics.

[112] J. Martineau and T. Finin. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media*, San Jose, CA, 2009. AAAI Press.

[113] D. G. Maynard. Term recognition using combined knowledge sources, 2000. PhD Thesis.

[114] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002. http://mallet.cs.umass.edu.

[115] W. McDougall. *An Introduction to Social Psychology*. Luce, 1926.

[116] Y. Mejova and P. Srinivasan. Exploring feature definition and selection for sentiment classifiers. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.

[117] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *Computing Research Repository (CoRR)*, abs/1301.3781, 2013.

[118] G. Mishne. Experiments with mood classification in blog posts. In *Proceedings of the 1st Workshop on Stylistic Analysis of Text For Information Access*, Brazil, 2005.

[119] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. In *Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006.

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

[121] S. M. Mohammad and S. Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

[122] S. M. Mohammad and P. D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, Los Angeles, California (USA), 2010. Association for Computational Linguistics.

[123] S. M. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[124] M. Montefinese, E. Ambrosini, B. Fairfield, and N. Mammarella. The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46(3):887–903, 2014.

[125] O. H. Mowrer. *Learning Theory and Behavior*. John Wiley, 1960.

[126] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, EMNLP '04, pages 412–418, Barcelona (Spain), 2004. Association for Computational Linguistics.

[127] M. Munezero, C. Suero Montero, E. Sutinen, and J. Pajunen. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111, 2014.

[128] R. Munro, S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily. Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 122–130, Los Angeles, California (USA), 2010. Association for Computational Linguistics.

[129] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California (USA), 2016. Association for Computational Linguistics.

[130] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia (USA), 2013. Association for Computational Linguistics.

[131] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Textual affect sensing for sociable and expressive online communication. In *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*, ACII '07, pages 218–229, Lisbon (Portugal), 2007. Springer-Verlag.

[132] N. Newman, W. H. Dutton, and G. Blank. Social media in the changing ecology of news: The fourth and fifth estate in britain. *Internet Science*, 7(1):6–22, 2012.

[133] V. Ng, S. Dasgupta, and S. M. Niaz Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 611–618, Sydney (Australia), 2006. Association for Computational Linguistics.

[134] F. Nielsen. Afinn, 2011.

[135] J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague (Czech Republic), 2007. Association for Computational Linguistics.

[136] Y. Niwa and Y. Nitta. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 304–309, Kyoto (Japan), 1994. Association for Computational Linguistics.

[137] K. Oatley and P. N. Jobnson-Laird. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1:29–50, 1987.

[138] Oriella. The new normal for news. have global media changed forever? *The 6th Annual Oriella Digital Journalism Survey*, 2013.

[139] A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychological Review*, 97(3):315–31, 1990.

[140] C. E. Osgood, G. Suci, and P. Tannenbaum. *The measurement of meaning*. University of Illinois Press, Urbana, IL, 1957.

[141] A. García Pablos, M. Cuadros, and G. Rigau. A comparison of domain-based word polarity estimation using different word embeddings. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož (Slovenia), 2016. European Language Resources Association (ELRA).

[142] G. Paltoglou and M. Thelwall. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1386–1395, Uppsala (Sweden), 2010. Association for Computational Linguistics.

[143] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Barcelona (Spain), 2004. Association for Computational Linguistics.

[144] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in Information Retrieval*, 2 (1-2):1–135, 2008.

[145] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 79–86, Philadelphia, Pennsylvania (USA), 2002. Association for Computational Linguistics.

[146] J. Panksepp. Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences*, 5(3):407–422, 1982.

[147] W. G. Parrott. *Emotions in Social Psychology: Essential Readings.* Psychology Press, 2001.

[148] L. C. Passaro, A. Bondielli, and A. Lenci. Fb-news15: A topic-annotated facebook corpus for emotion detection and sentiment analysis. In *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016*, Napoli (Italy), 2016.

[149] L. C. Passaro, G. E. Lebani, L. Pollacci, E. Chersoni, and A. Lenci. The coling lab system for sentiment polarity classification of tweets. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 87–92, Pisa (Italy), 2014.

[150] L. C. Passaro and A. Lenci. Extracting terms with extra. In *Proceedings of the EUROPHRAS 2015 – Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 188–196, Malaga (Spain), 2015.

[151] L. C. Passaro, L. Pollacci, and A. Lenci. Item: A vector space model to bootstrap an italian emotive lexicon. In *Proceedings of the second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 215–220, Trento (Italy), 2015. aAcademia University Press.

[152] Lucia C. Passaro and Alessandro Lenci. Evaluating context selection strategies to build emotive vector space models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož (Slovenia), 2016.

[153] M. T. Pazienza. A domain specific terminology extraction system. *International Journal of Terminology*, 5(2):183–201, 1999.

[154] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha (Qatar), 2014. Association for Computational Linguistics.

[155] E. Pianta, L. Bentivogli, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore (India), 2002.

[156] R. Picard. Blogs, tweets, social media, and the news business. *Nieman Reports*, 63(3):10–12, 2009.

[157] R. W. Picard. *Affective Computing*. MIT Press, Cambridge, MA (USA), 1997.

[158] J. C. Platt. *Advances in Kernel Methods*, chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pages 185–208. MIT Press, Cambridge, MA (USA), 1999.

[159] R. Plutchik. General psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience*, 1:3–33, 1980.

[160] R. Plutchik. On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion*, 9(2):197–200, 1985.

[161] R. Plutchik. *The psychology and biology of emotion*. Harper Collins, 1994.

[162] R. Plutchik. The nature of emotions. *American Scientist*, 89:344–350, 2001.

[163] T. Polajnar and S. Clark. Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 230–238, Gothenburg (Sweden), 2014. Association for Computational Linguistics.

[164] L. Polanyi and A. Zaenen. *Contextual Valence Shifters*, pages 1–10. Springer Netherlands, 2006.

[165] G. Qiu, B. Liu, J. Bu, and C. Chen. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1199–1204, Pasadena, California (USA), 2009. Morgan Kaufmann Publishers Inc.

[166] D. Rao and D. Ravichandran. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 675–682, Athens (Greece), 2009. Association for Computational Linguistics.

[167] E. Reid. Electronic chat: Social issues on internet relay chat. *Media Information Australia*, 67:62–70, 1993.

[168] L. Renzi, G. Salvi, and A. Cardinaletti. *Grande grammatica italiana di consultazione*. Number v. 1 in Strumenti (Bologna). Il Mulino, 2001.

[169] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics, 2003.

[170] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[171] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado (USA), 2015. Association for Computational Linguistics.

[172] S. Rothe, S. Ebert, and H. Schütze. Ultradense word embeddings by orthogonal transformation. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, San Diego California (USA), 2016.

[173] Klaus Rothenhäusler and Hinrich Schütze. Part of speech filtered word spaces. In *Sixth International and Interdisciplinary Conference on Modeling and Using Context*, Copenhagen (Denmark), 2007.

[174] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:161–1178, 1980.

[175] J. A. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, 115:102–141, 1994.

[176] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, 2003.

[177] H. Saggion and A. Funk. Extracting opinions and facts for business intelligence. *RNTI Journal*, E(17):119–146, 2009.

[178] M. Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54, 2008.

[179] M. Sahlgren and Stockholms universitet. Institutionen för lingvistik. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Spaces*. SICS dissertation series. Department of Linguistics, Stockholm University, 2006.

[180] Salton, G., Wong, A., Yang, and C. S. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[181] G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[182] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[183] K. L. Schmidt and J. F. Cohn. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology*, 116(s33):3, 2001.

[184] A. Severyn and A. Moschitti. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469, Denver, Colorado (USA), 2015. Association for Computational Linguistics.

[185] D. V. Shah, J. Cho, W. P. Eveland, and N. Kwak. Information and expression in a digital age. *Communication Research*, 32(10):531–565, 2005.

[186] M. L. Sheffer and B. Schultz. Paradigm shift or passing fad? twitter and sports journalism. *International journal of Sport Communication*, 3(4):472–484, 2010.

[187] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistic*, 22(1):1–38, 1996.

[188] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Honolulu (Hawaii), 2008. Association for Computational Linguistics.

[189] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, WA (USA), 2013. Association for Computational Linguistics.

[190] P. Stone. Thematic text analysis: New agendas for analyzing text content. *Text Analysis for the Social Sciences*, 1997.

[191] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[192] C. Strapparava and R. Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70–74, Prague (Czech Republic), 2007. Association for Computational Linguistics.

[193] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560, New York, NY, USA, 2008.

[194] C. Strapparava and A. Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1083–1086, Lisbon (Portugal), 2004. European Language Resource Association (ELRA).

[195] M. Taboada, C. Anthony, and K. Voll. Methods for creating semantic orientation dictionaries. In *Conference on Language Resources and Evaluation (LREC)*, pages 427–432, 2006.

[196] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, 2011.

[197] D. Tang, B. Qin, X. Feng, and T. Liu. Target-dependent sentiment classification with long short term memory. *Computing Research Repository (CoRR)*, abs/1512.01100, 2015.

[198] D. Tang, B. Qin, and T. Liu. Aspect level sentiment classification with deep memory network. *Computing Research Repository (CoRR)*, abs/1605.08900, 2016.

[199] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1555–1565, 2014.

[200] Richmond H. Thomason, editor. *Formal philosophy: Selected Papers of Richard Montague*. Yale University Press, 1974.

[201] R. Tokuhisa, K. Inui, and Y. Matsumoto. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22_{nd} International Conference on Computational Linguistics*, volume 1, pages 881–888, 2008.

[202] R. Tong. An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the SIGIR Workshop on Operational Text Classification*, pages 1–6, 2001.

[203] J. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Uppsala (Sweden), 2010. Association for Computational Linguistics.

[204] P. D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502, London, (UK), 2001. Springer-Verlag.

[205] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, Pennsylvania (USA), 2002. Association for Computational Linguistics.

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

[207] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

[208] A. Valitutti, C. Strapparava, and O. Stock. Developing affective lexical resources. *PsychNology Journal*, 2(1):61–83, 2004.

[209] A. Vanzo, D. Croce, and R. Basili. A context-based model for sentiment analysis in twitter. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, pages 2345–2354, Dublin (Ireland), 2014.

[210] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 777–785, Los Angeles, Californi (USA), 2010. Association for Computational Linguistics.

[211] S. Volkova, T. Wilson, and D. Yarowsky. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*, pages 505–510, Sofia (Bulgaria), 2013.

[212] J. Wagner, P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, and L. Tounsi. Dcu: Aspect-based polarity classification for semeval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 223–229, Dublin (Ireland), 2014. Association for Computational Linguistics and Dublin City University.

[213] H. Wang and J. A. Castanon. Sentiment expression via emoticons on social media. *Computing Research Repository (CoRR)*, 2015.

[214] D. A. Watson and A. Tellegen. Towards a consensual structure of mood. *Psychological Bulletin*, 98:219–235, 1985.

[215] J. B. Watson. *Behaviorism*. Chicago University Press, 1930.

[216] B. Weiner and S. Graham. An attributional approach to emotional development. *Emotion, Cognition and Behavior*, pages 167–91, 1984.

[217] J. Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740. AAAI Press, 2000.

[218] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Maryland, 1999. Association for Computational Linguistics.

[219] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Vancouver, British Columbia (Canada), 2005.

[220] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[221] C-H. Wu, Z-J. Chuang, and Y.-C. Lin. Emotion recognition from text using semantic labels and separable mixture models. In *ACM Transactions on Asian Language Information Processing*, pages 165–183, 2006.

[222] C. Yang, K. H-Y. Lin, and H-H. Chen. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136, Prague (Czech Republic), 2007. Association for Computational Linguistics.

[223] R. B. Zajonc. On the primacy of affect. *American Psychologist*, 39(2):117–123, 1984.

[224] M. Zhitomirsky-Geffet and I. Dagan. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35 (3):435–461, 2009.

[225] S. Zhou, Q. Chen, and X. Wang. Active deep networks for semi-supervised sentiment classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1515–1523, Beijing (China), 2010. Association for Computational Linguistics.

[226] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation, 2002.