University of Pisa

Department of Computer Science
Ph.D. in Computer Science

Superdiversity

(Big) Data Analytics at the Crossroads of Geography, Language and Emotions

Ph.D. Thesis

Ph.D. Candidate
Laura Pollacci

Supervisors
Prof. Dino Pedreschi
Prof.ssa Monica Barni
Asst. Prof. Alina Sîrbu

27 September 2019

Abstract

In a series of articles, Vertovec focused on the changes and contexts that have affected migratory flows around the world. These demographic changes, which Vertovec defines *Superdiversity*, are the result of the globalisation and they outline a change in the overall level of migration patterns. Over time, the migration routes have increased both their diversity and complexity. The nature of immigration has brought with it a transformative "diversification of diversity". Strictly connected with ethnicity and Superdiversity studies, the phenomenon of human migration has been a constant during human history.

In the era of Big Data, every single user lives in a hyper-connected world. More than 75% of the world's population has a mobile phone, and over half of these are smartphones. The use of social media grows together with the number of connected people. In these *social* Big data, User-Generated Content incorporate a high number of discriminating information. Language, space and time are three of the best features that can be employed to detect Superdiversity. The strongest point of social Big Data is that they typically natively include various information about different dimensions.

Starting from these observations, in this thesis, we define a measure of Superdiversity, a *Superdiversity Index*, by adding the emotional dimension and placing it in the context of social Big Data. Our measure is based on an epidemic spreading algorithm that is able to automatically extend the dictionary used in lexicon-based sentiment analysis. It is easily applicable to various languages and suitable for Big Data. Our Superdiversity Index allows for comparing diversity from the point of view of the emotional content of language in different communities. An important characteristic of our Superdiversity Index is the high correlation with immigration rates. For this reason, we believe this can be used as an essential feature in a nowcasting model of migration stocks. Our framework can be applied with higher time and space resolution compared to official statistics. Moreover, we apply our method to a different context and data to measure the Superdiversity of the music world.

Acknowledgments

> "Il problema è che abbiamo paura, basta guardarci. Viviamo con l'incubo che da un momento all'altro tutto quello che abbiamo costruito possa distruggersi, con il terrore che il tram su cui siamo possa deragliare."
>
> La Paura, F. de Luigi

Firstly, I would like to express my sincere gratitude to my advisors Dino Pedreschi, Monica Barni, and Alina Sîrbu for the support of my Ph.D. study and related research, for their motivation, and knowledge. Their guidance helped me in all the time of research and writing of this thesis. In particular, I am very grateful to Dino Pedreschi, who first suggested me to do a Ph.D., and then helped and guided me until today. I am also especially indebted to Alina for her invaluable role as advisor, supporter, psychologist, and friend. I could not have imagined having better advisors and mentors for my Ph.D. study. Together with my advisors, I am very grateful to Fosca Giannotti, who gave me her precious support and believed in me even in the most challenging moments of this journey. Besides my advisors, I would like to thank the rest of my thesis committee and reviewers, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also go to my KDDLab colleagues for their acceptance, suggestions, and clarifications: Michela, Francesca, Letizia, Anna, Salvo, Daniele, Riccardo, Paolo, Lorenzo, Luca, Salvatore, Mirko, Roberto T., Roberto P., Valerio, and Ioanna. A special thank goes to (Saint) Giulio, and Vittorio. They helped me, welcomed me into their office, and endured me every day. I want to thank also all my Ph.D. colleagues. Many thanks to my - few but real - friends that tried to understand me even not knowing what it means to do a Ph.D.

The biggest thank goes to my family for their sincere love. In particular, my *mamy* has travelled this journey with me, supporting me in the darkest moments, and showing me her unconditioned love. Finally, I thank Gianmarco, that despite our differences, fights every day with me to build our future.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In a series of articles, Vertovec focused on the changes and contexts that have affected migratory flows around the world [232]. These demographic changes, which Vertovec defines as *Superdiversity*, are the result of globalisation and they outline a change in the overall level of migration patterns. Over time, the migration routes have increased their diversity and their complexity, more people are now moving from more places, through more places, to more places [233]. The nature of immigration has brought with it a transformative "diversification of diversity" [102, 142] not just in terms of bringing more ethnicities and countries of origin, but also concerning multiplication of significant variables that affect where, how and with whom people live [232].

The concept of Superdiversity aims to encompass the increasingly complex and less predictable set of relationships between ethnicity, citizenship, residence, origin, and language. In terms of identity, the multilingualism has reached the next level in the context of superdiversity. The term "2.0" is often used to refer to this next linguistic level. The basis of this idea is that the Internet has created a new democratic space of interchange and allowed the development of new identities. There is no doubt that the meaning of "2.0" relates to Web 2.0, the second revolutionary phase in the development of the World Wide Web. This phase is characterised by greater interactivity compared to the previous stages of the Internet, in which even the "not-specialists" are involved and participate actively.

In the multilingual Europe 2.0, ethnicity is no longer determined only by biological descent. Understanding ethnicity is tantamount to abandoning "the tendency to take discrete, sharply differentiated, internally homogeneous and externally bounded groups as basic constituents of social life, chief protagonists of social conflicts, and fundamental units of social analysis" [41]. Nowadays, the demographic changes resulting in Superdiversity urge us to revisit, deconstruct and reinvent many of our established assumptions about language, identity, ethnicity, space, culture, and communication [30].

The Superdiversity concept is inherently strictly connected with the phenomenon of human migration, which has always been a constant during human history. The study of migration involves several research fields, including anthropology, sociology, and statistics. We are now at the moment where other research fields, such as physics and computer science, are involved. At the same time, the rapid growth in data availability and data characteristics have led researchers to focus on types of data not typically used to study this phenomenon. Nowadays, both traditional and novel approaches and data types are being exploited and combined for understanding different questions on migration. Thanks to the availability of *social* Big Data it is becoming possible to study migrations in real time. In particular, it should be possible to forecast migration stocks and flows by defining measures from these novel and unconventional data sources. Moreover, Social Big Data offer new opportunities for observing integration and perception of migration, building thus new integration indices.

Despite the data availability and their characteristics, Superdiversity has not yet been practically measured. As a consequence, even if in migration studies several indexes are used for several purposes, a Superdiversity measure is still missing. *Social* Big Data, such as Online Social Networks (OSNs) and User-Generated Content (UGC), incorporate a high number of discriminating information. Language, space and time are important features in the detection of Superdiversity. In this sense, much of the data available on social platforms have these characteristics. For instance, tweets on Twitter have a linguistic dimension, represented by the text and its peculiarities, such as hashtags, and emoticons; a time dimension, in terms of date and time in which they were published; and a geographic dimension, information about geo-location of the user - manual or automatic in the account settings. In this thesis we aim to use these data to fill the gap in measuring Superdiversity, which should bring us closer to being able to really measure the diversity among different communities, as well as between different groups of people, e.g., musicians playing different music genres.

1.1 Research Outline

We believe that Big Data can allow us to answer several questions on human behaviour as well as open questions on Superdiversity and migration studies. Our framework is located at the crossroads of three major areas of research: *a)* Data Mining, *b)* Linguistic Analysis, and *c)* Sentiment Analysis. It combines *linguistic diversity*, *space* - understood in terms of the geographical location of production of contents -, *time* - the production time of the text, in a perspective of evolution and temporal mutation of content -, and, *emotional content*.

We aimed to find a model that can be used to describe the concept of diversity - and Superdiversity - both in geographically restricted contexts and in much broader geographical contexts. That is the motivation that led us to move on to the complex multi-faceted system that the Internet is, which allows us to investigate from small communities to entire nations and - potentially - to wander

around the world.

Our motivations are based on the gap between Superdiversity theory and the existence of a real measurement of Superdiversity, which we believe can be applied to different contexts. Thus, in this thesis, we propose an approach to shift from theory to the real application of the concept of Superdiversity, by adding the emotional dimension and placing it in the context of Big Data.

We start from the hypothesis that different communities associate different emotional valences to words due to cultural differences, and we believe we could use this variability to measure cultural diversity. Based on a similarity measure between "standard" and "community" use of a language, we define the first Superdiversity Index (SI) as the distance between the sentiment valence of words used by a community and the standard valences coming from a manually tagged dictionary. Thus, our SI allows for comparing diversity from the point of view of the emotional content of language in different communities.

Our SI is computed starting from a measure related to the use of the language, that we obtained by designing an epidemic spreading algorithm. This spreading algorithm can extend an initial dictionary used in lexicon-based sentiment analysis automatically but also helps to characterise the group where the observed document came from. Our framework can be applied with higher time and space resolution compared to official statistics published by government agencies and international organisations.

We assessed the validity of our model in a multi-scale system: for example, for the study of small groups of individuals in a confined geographical context, as well as to wider levels, both geographically – national and international contexts, and also in terms of community size. Overall, our research highlights the potential of the data generated by users in the definition of open source indicators for human diversity and behaviour. Our approach allows us to tackle many issues, including the linguistic homogeneity, the geographic distribution of languages, the identification of specific linguistic communities, and ethnic homogeneity.

An important characteristic of our SI is the high correlation with immigration rates. For this reason, we believe that it could be used as an important feature in a future nowcasting model of migration stocks. Moreover, we also apply our method to a context that has not been widely explored previously, i.e., the music panorama, with particular attention to the Italian one. In this research context there are still many open research questions, such as those linked to feelings and emotions. To the best of our knowledge, until now no research analysed and compared the emotional content of Italian music from a regional point of view. We believe that this context is suitable for applying our method to investigate lexical and emotional differences in Italian music production. Moreover, this analysis allows us to check the consistency of our index even in domains other than human migration.

During the development of our framework, since we measure the Superdiversity in different contexts, we noted the lack of a pipeline able to format, clean and geo-localise different types of

data. This has lead to the development of an NLP pipeline able to automatically format, clean and geo-localise different kinds of data coming from several resources.

1.1.1 Specific contributions

In the following, we summarise the contributions we propose in this thesis.

- A Superdiversity Index (SI) as a measure of Superdiversity extracted from language and sentiment. The SI measures the distance between the sentiment valence of words used by a community and the standard valences from a manually tagged dictionary. The index can be computed with various space and time resolutions.
- An analysis of Superdiversity on Twitter. We evaluate Superdiversity in various communities in Italy and the UK and show that it correlates very well with immigration rates. Due to its strong correlation with migration rates, we believe the Superdiversity Index can form a strong basis for a nowcasting model of migration.
- Development of a sentiment analysis algorithm able to take into account lexical, emotional and geographical dimensions. The algorithm can extend the initial dictionary starting from a corpus of data. This dictionary also helps to characterise groups where the document came from. It enables an increase in the number of messages that can be classified, maintaining a good classification performance. The algorithm is also extensible to many languages.
- An NLP Pipeline for pre-processing tweets and music lyrics to format, clean and geolocalise text. This allows to automatically process different kinds of data gathered from different sources.
- Analysis of the global-wide music context that takes into account also the Tuscan emerging youth bands, in which we found a fractal structure. The Italian music context was deeply analised and Superdiversity was measured through the creation of lexical and melodic profiles.
- Creation and publication of the first Italian Music Dataset composed of both famous and emerging Italian musicians. The dataset contains music-related information, such as tracks' titles, as well as geographic information, such as the place where the artist came from.

1.2 Thesis Organisation

The thesis is organised in three main parts which are divided in turn in several sections.

Part I presents the related state-of-the-art in the research fields covered by this thesis. Thus, Chapter 2 provides concepts, basic notions, and literature related to Sentiment Analysis. Chapter 3 introduces the Human Migration research field through three different migratory phases and

presenting both traditional and unconventional data sources and approaches. Finally, Chapter 4 provides a brief state-of-the-art of the research work connected with the music context.

After having introduced all primary concepts, in the second part, we describe the data and resources we used and the preprocessing phase in Chapter 5. Then, in Chapter 6 we present our lexicon-based epidemic model for Twitter Sentiment Analysis (TSA) and its evaluation. In Chapter 7 we define and describe our Superdiversity Index together with the evaluation of the Superdiversity in communities in Italy and the UK.

The third part focuses on the music context. We present datasets, musical features and the preprocessing phase in the first part of Chapter 8. In the second part of Chapter 8, from Section 8.2, we focus on the Italian music panorama with particular attention to the Tuscan emerging youth bands. In Chapter 9 we analyse the worldwide music context, and we focus on the identification of the music fractal structure. Chapter 10 focuses on the Italian music context and describes the creation of lexical and musical profiles that can be used to measure the Italian Music Superdiversity.

Finally, Chapter 11 concludes this thesis by summarising the results we obtained and their contribution to the research field, and discussing some possible future plans.

1.3 Publications

- Laura Pollacci, Riccardo Guidotti, and Giulio Rossetti. Are we playing like music-stars? placing emerging artists on the italian music scene. In *9th international workshop on machine learning and music*, 2016
- Laura Pollacci, Alina Sîrbu, Fosca Giannotti, Dino Pedreschi, Claudio Lucchese, and Cristina I. Muntean. Sentiment spreading: an epidemic model for lexicon-based sentiment analysis on twitter. In *Conference of the Italian Association for Artificial Intelligence*, pages 114–127. Springer, 2017
- Laura Pollacci, Riccardo Guidotti, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. The fractal dimension of music: geography, popularity and sentiment analysis. In *International conference on smart objects and technologies for social good*, pages 183–194. Springer, 2017
- Laura Pollacci, Riccardo Guidotti, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. The italian music superdiversity. *Multimedia Tools and Applications*, pages 1–23, 2018

Submitted

- Alina Sîrbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisu Kim, Cristina Ioana Muntean, Luca Pappalardo, Andrea Passarella, Dino Pedreschi, Laura Pollacci and Rajesh Sharma. Human migration: A Big Data perspective. Submitted to *Springer Journal of Data Science and Analytics*.

In preparation

- Laura Pollacci, Alina Sîrbu, Fosca Giannotti, Dino Pedreschi. Measuring the *Salad Bowl*: Superdiversity on Twitter. To be submitted to *Springer Data Mining and Knowledge Discovery*.

Part I

Setting the Stage

Chapter 2

Sentiment Analysis and Social Media

Several studies have shown that an explicit call for a welfare assessment generates biased estimates of the variable of interest. However, indicators of individual and social well-being are usually obtained from surveys based on self-evaluation. On the contrary, social networks and web forums offer a large and continuously updated source of spontaneous self-evaluations, needs, perceptions and expressions of moods. Therefore a new stream of studies is trying to extract reliable information from these sources. To perform this task, fields such as Computational Linguistics and Natural Language Processing (NLP) have focused on textual analysis. In the last few years, the Sentiment Analysis field has attracted much interest. This discipline addresses the automatic detection and classification of feelings and emotions. Sentiment Analysis (SA) is a methodology that allows determining the intensity - positive or negative - of the sentiment expressed in an item. To do this task, it exploits some Computational Linguistics approaches, such as an NLP pipeline and textual analysis. Sentiment Analysis is based on searching keywords to identify, for each polarised term, attributes - positive, negative, neutral - "such that, once aggregated distributions of these terms, it becomes possible to extract the opinion associated with each term key" [51]. Hence it is possible to determine and extract quantitative information from texts on the subjective assessments of their content. To understand human behaviour and aspects of communication and social interaction, feelings are crucial. This research field is widely dedicated to the systematic extraction of Web users mood from the texts they post on Internet platforms, such as blogs, forums, social networks, i.e., Twitter or Facebook [51]. A standard task in Sentiment Analysis is classifying the polarity of a given text - document, sentence, or feature/aspect level. The goal of these tasks is to understand the opinion expressed inside the mood and its polarity. The literature shows a different grading scale. There are binary scales - positive *versus* negative -, ternary scales - positive, neutral, and negative -, continuous scales

that provide intermediate levels, i.e., *High_Positive, Positive, Neutral, Negative, High_Negative*. As underlined in [81], within SA and OM, several sub-tasks can be identified. All tasks foresee the use of a tagged text according to expressed opinion. Among them we can distinguish:

- *determining SO-polarity*, such as detect if a given text does not express an emotive opinion or expresses an opinion. These tasks use a binary classification based on Subjective and Objective classes [174, 255].
- *determining PN-polarity*, such as decide if a Subjective text expresses a Positive or a Negative opinion [174, 226].
- *determining the strength of text PN-polarity*, such as identify whether the Positive opinion expressed by a text is *Weakly_Positive, Mildly_Positive*, or *Strongly_Positive* [175, 245].

Existing approaches can be grouped into two main categories: *machine learning* methods and *rule based* or rather, *lexical methods*. Machine learning methods are data-driven approaches based on the use of learning algorithms. They use labelled corpora to extract sentiment information. Such techniques require creating a model by training the classifier with labelled examples. This means that first we need to gather dataset with examples of the examined classes. Then features are extracted from the examples and finally the algorithm can be trained based on the examples.

This approach courts machine learning techniques such as Latent Semantic Analysis (LSA), Support Vector Machines (SVM), the so-called "bag of words" approach (BOW) and Pointwise Mutual Information (PMI). Lexical methods use rules of thumb or heuristics to determine sentiment polarity. These techniques employ dictionaries of annotated words with their semantic polarity and sentiment strength. These dictionaries are used to calculate the polarity score and the document sentiment. Usually, these methods give high precision, but low recall and they are often "specialised" according to text types. Besides the thematic just mentioned, in this Chapter, we provide a comprehensive state-of-the-art of the Sentiment Analysis research field.

The rest of the Chapter is structured as follows. Section 2.1 provides the basic concepts relevant to the field. After that, Chapter 2.2 describes a sub-field of Sentiment Analysis, thus the Twitter Sentiment Analysis. Levels of sentiment classification and details of Sentiment classification are discussed in Section 2.3 and in Section 2.4, respectively. Section 2.5 provides and overview of the polarity lexicon induction, while Section 2.6 discusses available resources in Sentiment Analysis and their characteristics. The Section 2.7 addresses and describes motivations and both research and economical areas where prediction with social media may be made. Finally, Section 2.8 concludes the Chapter by providing and discussing the major challenges in Sentiment Analysis.

2.1 Sentiment Analysis: The Concept

The search field of Opinion Mining and Sentiment Analysis aims at the automatic detection and classification of feelings and emotions. The Merriam-Webster dictionary[1] defines the term *sentiment* as "an attitude, thought, or judgment prompted by feeling", as well as "a specific view or notion: opinion" and "emotion". Moreover, the concept of *opinion* is defined as "a view, judgment or appraisal formed in mind about a particular matter". Beyond formal definitions, "opinions are usually subjective expressions that describe people's sentiments, feelings toward entities, events and their properties" [132].

Most of the early research on textual information processing has been focused on information retrieval, text classification, text clustering and text mining [95]. The lack of Sentiment Analysis studies in the past was due to the few amounts of opinioned text available before the advent of the World Wide Web and the Web 2.0. As underlined by Liu et al. [132], before the rise of the Internet, individuals usually asked their acquaintances for opinions for making a decision, as well as companies conducted opinion polls and surveys to gather feedbacks by consumers about their products and services. After the rise of the Internet and the advent of the so-called social era, people are encouraged to express their opinions. This has lead to a proliferation of posts expressing opinions. Despite their nature, social networks have radically changed how we communicate. We search, share information, present our point of view, make judgements and impressions differently. The impersonal nature of communication, ironically, increases the potential of the relations and the ability of expression. Subjects, hidden behind the keyboard of a PC have no qualms about expressing opinions that face to face could inhibit. We refer to the *screen mediation* phenomenon [148], long debated in Psychology in recent years. For this reason, content from social networks has provided an unrivalled boost in the study of human behaviour in terms of feelings. A huge amount of social media including news, forums, product reviews and blogs contain various sentiment-based sentences.

A sentiment can be defined as "a personal belief or judgement that is not founded on proof or certainty[2]". Sentiment expressions describe the mood or the opinion of the writer - i.e., happiness or sadness - towards some specific entity. The importance of quantifying nature and intensity of emotional states at the level of populations is evident [69]. We aim to know how, when, and why individuals express feelings, in order to improve public policies, products, and services. The study of users mood need to deeply understand the nature of social, economic phenomena and the patterns in human behaviour so, there is a great area of interest in various fields of study. Information provided by Internet users posting reviews online represents the major driver for the Sentiment Analysis

[1] The Merriam-Webster dictionary site: https://www.merriam-webster.com/
[2] WordNet 2.1 definition.

research field. Pang and Lee [176] underline three additional factors leading to the huge burst of research of Sentiment Analysis. These include:

a) the rise of machine learning methods in NLP and Information Retrieval (IR) fields.

b) the availability of datasets for training machine learning algorithms, due to the rise of review aggregation websites.

c) the overcoming of the challenges and development of intelligence applications related to this area.

2.1.1 Definitions

Let d be an opinionated document - i.e., a product review - composed of a list of sentences $s_1, ..., s_n$, according with Liu et al. [134], the basic components of an opinion expressed in d are:

- *Entity*: can be an event, a product, a person, a topic or a service on which opinion is expressed - the opinion target. Entities are composed of sub-components that may have a set of attributes. For instance, a mobile phone includes a battery that can be characterised based its size and weight. Commonly both components and their attributes are referred to as aspects.

- *Opinion holder*: the person or the company holding a certain opinion on a particular entity. For instance, reviews' opinion holders are usually the authors of the reviews, while news articles opinion holders are explicitly indicated [24].

- *Opinion*: the attitude, a view or appraisal about an item made by an opinion holder. An opinion has an orientation that can be positive, negative or neutral, where a neutral orientation is interpreted as no opinion, as news. The orientation is usually named sentiment orientation, semantic orientation [226], or polarity.

Given the components of the opinion, an opinion is defined as a quintuple $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ [132], where e_i is an *entity*, a_{ij} is an *aspect* of e_i and oo_{ijkl} is the opinion orientation of a_{ij} expressed by the holder h_k during time period t_l [38]. oo_{ijkl} can be represented in several ways, but generally refers to positive, negative and neutral categories or to different strength or intensity scales. If the opinion refers to wholes entities, a_{ij} is named GENERAL. On the contrary, an opinion can refer only to a part of the entity. Indeed, several opposed opinions referring to the same entity could be found in the same opinionated document. This scenario is dealt with by discovering all opinion quintuples $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ from a collection D of opinionated documents through aspect-based or feature-based opinion mining methods [134, 38].

2.2 Twitter Sentiment Analysis

The Sentiment Analysis field is a growing area of Natural Language Processing in which research ranges from the document-level classification [176] to the polarity-identification of words [81, 98]. Due to the character limitations of tweets, classifying the sentiment - or the polarity - of Twitter messages is a task similar to sentence-level Sentiment Analysis [117, 255]. Social media platforms typically show an informal and peculiar language. In addition to this, messages on Twitter include various specific features making the Twitter Sentiment Analysis a very different task than generic SA [121]. Thus, Twitter more than other microblogging platforms poses newer and different challenges [1]. As is well-known, microblogging is a network service where users can share comments, images, videos and links to external websites that are visible to users subscribed to the service. Moreover, in contrast with traditional blogs, messages are short. For instance, from 2006 to 2017 Twitter had imposed a maximum of 140 characters. The character limit has been doubled from 2017. According to the last statistics, currently, Twitter has 330 million users who post more than 500 million messages per day[3]. The two major Sentiment Analysis tasks in Twitter Sentiment Analysis are polarity classification and opinion identification. According to [38], due to the short nature of Twitter posts, depending on the tasks, a sentence-level classification approach can be adopted by assuming that tweets express opinions about one single entity. Twitter provides a particularly easy process to access and download published posts, also by exploiting different search criteria. Due to this, it is considered one of the largest datasets of user-generated contents [87]. Twitter has developed a public API[4], called Twitter API, through which is straightforward to retrieve a huge amount of data by using easy queries. Twitter API rate limits are divided into 15-minute intervals and there are two initial buckets available for GET requests: 15 calls every 15 minutes, and 180 calls every 15 minutes[5].

As mentioned before, messages on Twitter are characterised by some specific features. In the following, we provide a detailed list of these Twitter-specific features.

- *Tweet*: In a simplified view, tweet is a message posted on Twitter. Tweets are the basic unit of the Twitter-world. A tweet object has several "root-level" attributes, consisting in the tweet's metadata. These attributes concern different tweet-related aspects and include attributes such as `id`, `created_at`, and `text`. Tweet objects are a kind of "parent" object to several "child" objects. Thus tweets' metadata are organised following a hierarchical structure. Tweet child objects include `user`, `entities`, and `extended_entities`. Furthermore, in the case of geolocalised tweets, these metadata also include the `place` child object. The basic

[3]Source JuliusDesign: https://urly.it/31f09, accessed January 2019.
[4]Twitter API. https://developer.twitter.com/en/docs/tweets/search/api-reference.html
[5]An exhaustive documentation of the API is provided in the Developer section of the platform at https:developer.twitter.com.

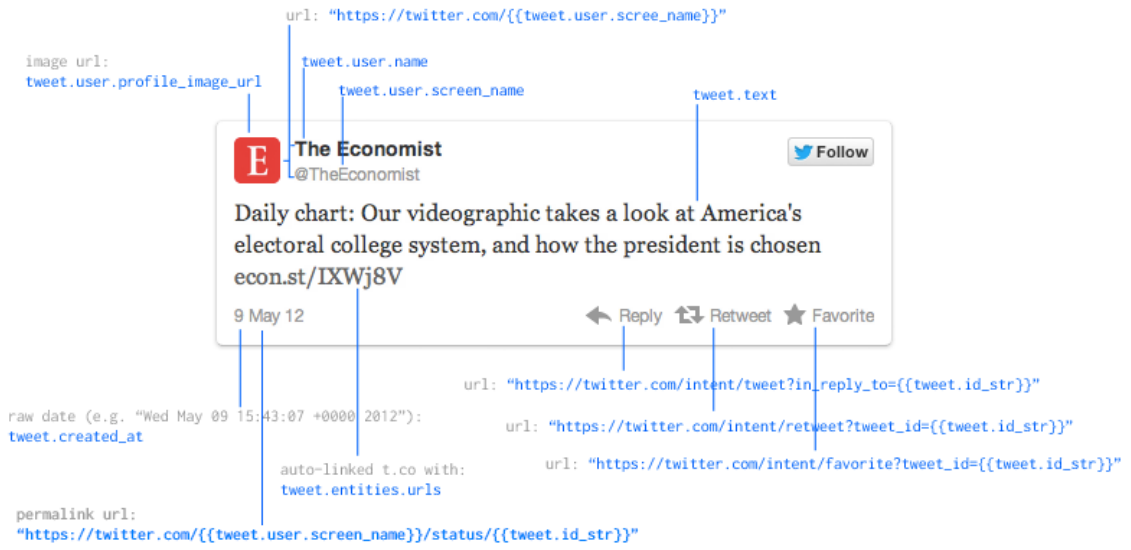


Figure 2.1: Anatomy of a tweet [212].

anatomy of a tweet is shown in Figure 2.1. The text of a tweet is composed by at most 280 characters, which besides text comprises both hashtags, usernames, and emoticons. Posts can also include links to external websites, news, photos, and videos.

- *Users & Usernames*: A user is an end-user that has to be registered to the platform to post tweets. During the registration phase, a user chooses his/her pseudonym as the username which is displayed when he/she posts messages.

- *Retweet*: A retweet is a tweet that in turn is redistributed by a user different by the initial writer. Retweet functionality is considered one of the most powerful tools for disseminating information [87]. Typically, when a user finds an interesting tweet, he/she can re-post it, eventually adding a personal contribution to the original one. Even without adjuncts, the post is marked as a retweet. The marker is automatically placed at the beginning of the tweet. It consists of the abbreviation *RT* followed by the mention of the original writer, i.e., *RT @username text*.

- *Reply*: A reply is a post used to answer to another tweet. Replies are typically employed to create conversations. Replies are marked by using the @ symbol followed by the username they refer to. Then the reply is placed next to the username that creates the reply.

- *Mention*: A mention results in the inclusion into the tweet of a username of a Twitter-user different from the writer. This reference can be placed everywhere in the body of the tweet.

To make a mention, the writer of the post must type the special character @ followed by the username they want to refer to - i.e., *@username*.

- *Follower*: A follower is a user who namely follows another users tweets and activity. The process of following, similar to the friendship on Facebook, is how users stay connected among themselves. The practice of following means that users receive updates from those they follow and send updates to those who follow them.

- *Hashtag*: A hashtag is a sort of keyword indicating the topic - or topics - whom the tweet refers to. Hashtags are composed by the special character # followed by the topic name, i.e., *#topic*. Tags can be freely created by users and can be used as query keywords to get all the tweets with the same hashtag, and thus referred to the same topic. As stated by Twitter developers, a hashtag cannot include punctuation marks as well as spaces. This means that hashtags may consist of one or more words not separated by spaces, i.e., *#savetheworld*. All the other characters are allowed, as numbers, i.e., *#schoolstrike4climate*. To make it easier to understand hashtags composed by various words, typically users capitalise initial characters of every single word, such as $\#WeLovePasta$ or $\#iHeartAwards2019$. When a hashtag is frequently used, it is considered as a trending topic. Hashtags represent one of the most peculiar traits of tweets and are the most important lexical expression of users' creativity.

2.2.1 Challenges in Twitter Sentiment Analysis

Besides challenges related to traditional Sentiment Analysis, detecting sentiments from Twitter involves several specific issues. As a consequence, the TSA is a non-trivial task which considerably differs from SA of more conventional texts as blogs [87]. Twitter-specific features bring to light emerging lexical trends as well as important limitations. Moreover, it is important to note that these characteristics have to be included in an informal, dynamic and evolving panorama. In the following, we describe the most relevant challenges of the TSA research field.

- *Text Length*: One of the most well-known unique characteristics of tweets is their length, which can be at most of 280 characters. As a consequence, this is the ground difference between standard SA and TSA. The study proposed by Bermingham and Smeaton [22] focus challenges involved by the short length of tweets and examine if this makes it more difficult to accomplish SA tasks. To answer their questions, the authors compared Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) classifiers' performance on both blogs and Twitter data. Provided results show that SVM performs better than MNB on blogs. Vice versa, MNB performs better than SVM on microblogs. More importantly, the authors stated that classifying short texts, i.e., tweets, is a more easy task than classifying longer documents.

- *Fancy Language*: The Twitter-language is one of the most specific languages amongst all

the microblogging platforms. Besides its informality and its length limitation, tweets' language shows several peculiarities which cannot be found in other text types, such as blogs or news. Amongst the various textual peculiarities, the most common include abbreviations and acronyms, emphatic lengthening and emphatic upper-casing, neologisms and widespread use of slang and emoticons. Related to these textual characteristics, the work presented in [39] focuses on frequency and impact in TSA of the emphatic lengthening phenomenon. Even though the study is dated if considering the rapidly evolving of the Twitter language, the authors already in 2011 affirmed that emphatic lengthening is very frequent on Twitter.

- *Data Sparsity*: As a consequence of non-standard and fanciful language, tweets contains much noise as well as misspellings. To this are added typing errors, mistakes in hashtags creation, local dialects variation, and so forth. The phenomenon is known as data sparsity and strongly affects all lexical tasks, including SA-related tasks. According to [201] and [87], one of the main reason of data sparsity on Twitter relates to the high percentage of tweets' terms that occur few than ten times in the entire corpus. The work presented by Saif et al. [201] focuses on reducing data sparseness of tweets. The authors proposed two sets of features to alleviate the data sparsity problem in Twitter sentiment classification, namely semantic features and sentiment-topic features. Their results show both methods improve the baseline Naïve Bayes model using unigrams only. Furthermore, by using sentiment-topic features, they obtain better results than using semantic features with fewer features.

- *Topic Relevance*: is a less explored TSA task. Most of the works done on tweets aim to polarity-classify them instead of considering the topical relevance. The research contributions which explore this particular field typically inspect the presence - or absence - of given words as indicators of the tweets relevance towards a specific topic. In other cases, hashtags have been used as topic relevance indicators. In the SA context, the topical relevance seems to be easier in short texts such as tweets. Proposed techniques seem partially right as in most of the cases the sentiment will target a specific topic [87].

- *Multilingual Content*: Tweets are written in several languages, also mixed in the same post. In these cases, language detection becomes difficult. The scenario seems is more critical as a result of the tweets' short length. To manage issues related to multilingual corpora, several works have been developed. For instance, [166] developed a language-independent classifier. The classifier is evaluated over a four-language Twitter dataset. Provided results show that the proposed method performs effectively in multi-language contexts without needing extra process.

- *Multimodal Content*: Besides the wide usage of multi-language contents, tweets are also characterised by a broad embedding of additional contents. These contents are in various kinds including images and videos. Both images and video analysis could be valuable for TSA since

they can provide useful information in determining the opinion holder as well as for the entity extraction. Nowadays, features extraction in multimodal contents for SA has been little explored [87].

- *Negation*: In SA, the presence, as well as the absence of negations, plays an important role. The correct negation handling is not trivial and remains challenging. Moreover, also the negation placement along with its action radius - i.e., words which are affected by the negation - is relevant. The importance of negation is due to its power in flipping polarity of sentences or words. Most of the existing works use simple techniques for handling negations. A common way consists in reversing the sentence's polarity when negation is found. Another wide used method foresees to reverse only the polarity of the word preceded by the negation. The contribution proposed in [119] is focused on negation handling and its effects. The authors developed two separate lexicons. One is composed of terms usually appearing in contexts with negations. The other one contains terms that usually appears in contexts without negations. Their analysis shows that negation of positive terms leads to negative sentiment. On the contrary, in cases of negative terms, the polarity remains unchanged in the negated context.

- *Stop Words*: The so-called "stop words" usually refers to particularly common words having a low discrimination power. These words are typically filtered out before or after processing of NLP [194]. Most of the words typically fall in particular parts-of-speech, as determiners, coordinating conjunctions, interjections, articles, and prepositions. However, a non-specific set of stop words may also include some verbs, such as "to be" and "to have", and modal and possessive verbs. Despite the fact that stop words are typically meaningless words of a given language, there is no single universal list. Thus, not all tools even use such lists. Some frameworks avoid removing stop words to support phrase search. Depending on the language, various lists can be freely found, like the one proposed by Stanford CoreNLP[6]. Domain-specific corpora require domain-specific word lists. For instance, in [188] the authors detected and removed general Italian stop words as well as music domain-specific stop words in all lexical variations (i.e., *strofa, ritornello, rit*). Moreover, pre-compiled stop words lists are not suitable for Twitter tasks. For instance, words as *like* or *unlike* is generally included in pre-compiled stop words lists, as in Stanford CoreNLP, Snowball[7], and Terrier[8] lists. In TSA tasks, this term has an important sentiment discrimination power. Include or exclude these words may strongly affect the SA - and thus the TSA - performance. The lack of universal pre-compiled lists has led the development of some works focused on building stop-words lists for Twitter. For instance, in [119] is presented an analysis of the impact of removing stop words on TSA

[6]Stanford CoreNLP stop word list: https://urly.it/31fba
[7]Stop word list published with the Snowball Stemmer: http://snowball.tartarus.org/algorithms/english/stop.txt
[8]stop word list published with the Terrier package: https://urly.it/31fc3.

effectiveness. By exploiting six different Twitter dataset, the authors applied six different stop-word identification methods. The analysis is performed in terms of classification performance, variations in data scarcity and size of the classifiers feature space.

- *Tokenization*: Tokenization is the process aiming at demarcating and classifying chunks of a string of input characters. To put it simply, it is the process of segmenting an input text into words or sentences. Given a character sequence, tokenisation is the task of partition it into basics linguistic units, called *tokens*. Since these basic units are often essential for most of NLP tasks, their effective identification is crucial. Errors in tokenisation may lead to critical errors in all later stages of text processing. Contrary to expectations, the process is non-trivial and involves several issues. The tokenisation typically occurs at the word level, but defining what is meant by a "word" could be difficult. Low-level tokenisers rely on simple heuristics, like the following:

  – Punctuation and white-spaces can be both included or excluded in the output list of tokens.

  – All contiguous strings of alphabetic characters represent the same token - likewise with numbers.

  – Tokens are separated by white-space characters, including blanks, line break, and punctuation characters.

  Most of the languages that use the Latin alphabet - as well as most programming languages - typically exploit white spaces to delimit different words. With languages that use inter-word spaces, a low-level tokenisation approach seems fairly straightforward. However, tokens do not even correspond to words since white spaces may result in different meanings. Expressions like "New York" and "rock 'n' roll" are individual expressions even if contain multiple words and blanks. On the other part, words like "I'm" need to be separated into "I" and "am". The task becomes more difficult for *scriptio-continua* languages in which words have no boundaries, such as Ancient Greek, Chinese, and Thai. Agglutinative and fusional languages, such as Dutch, German, Japanese and Korean, also make the tokenisation task harder to accomplish. In addition to the above, another challenge for tokenisation is the so-called "dirty text", or "noisy text". In general, it is not safe to make assumptions on the source text's correctness and integrity. Automatically extracted texts may contain several issues, including spelling errors, unexpected characters, and inaccurately compounded tokens. Moreover, if source texts are stored in a fixed-fields database having multiple lines per object, fields could need to be reassembled. However, the fields may have inconsistently been trimmed. Due to all these issues, some works focus on developing domain-specific or context-specific tokenisers. For instance, [172] proposes a Twitter-specific tokeniser. Other well-known tokenisers include Apache Open

NLP[9], Stanford Tokenizer[10], LinguA[11][64].

Amongst the all above-mentioned challenges, some - such as negations, and stop words - need to be managed in all SA tasks, independently of the kinds of input textual data. Moreover, in the TSA scenario, all these challenges should be taken into account to develop effective methods.

2.2.2 Feature Selection for Twitter Sentiment Analysis

Twitter messages imply several challenges related to features identification for effective polarity classification. Feature selection and their combination play a key role in detecting the sentiments of texts. In the domain of online reviews and news articles, various types of textual features have been considered, including part-of-speech tags. In the microblogs domain, textual features can be classified into four classes, namely *Syntactic, Semantic, Stylistic,* and *Twitter-specific*. While syntactic, semantic, and stylistic features are just well-known in SA literature, the Twitter-specific ones have lead to new challenges. Nevertheless, TSA feature selection is typically based on existing approaches evaluated over other domains. For instance, in [119] sentiment terms are identified by using the Pointwise Mutual Information (PMI) measure. Another suitable measure is the Chi-squared test, as in [5]. Besides established feature selection criteria, several works also focus on their impact, as in [1, 173, 121].

Further to inherent issues related to the standard feature selection, these need also to be domain-dependent to be effective. Due to this, several works focus on the usefulness of different features in the Twitter context. In [1], the authors propose a feature-based model which amongst others include both part-of-speech and lexicon features. Their results suggest that most performing features combination is represented by POS and words' polarity. The impact of different features in TSA is also examined in [121]. The work focuses on both semantic and stylistic features, including abbreviations, emoticons, and intensifiers. Best performances are achieved by words' polarity-related features and n-grams. Interestingly, in contrast with the study carried out by Agarwal et al. [1], Kouloumpis et al. [121] stated that POS had a negative impact on TSA.

Typically, the feature selection process starts by segregating words and features. Then, all the various feature selection techniques are applied to the input text. Finally, the most performing and informative features are selected. The major drawback of this method regards the series of tests that need to be conducted to handle all possible phenomena, such as negation and irony/sarcasm. To address these limitations, several algorithms have been developed. In particular, researchers are interested in developing methods able to learn representations of the data to extract information needed to build effective classifiers [20]. Recent research focuses on deep-learning methods based

[9]Apache Open NLP package comprises SimpleTokenizer, WhitespaceTokenizer, and TokenizerME. Apache Open NLP: https://opennlp.apache.org/docs/1.8.2/apidocs/opennlp-tools/overview-summary.html
[10]Stanford Tokenizer: https://nlp.stanford.edu/software/tokenizer.shtml
[11]LinguA: http://www.italianlp.it/demo/linguistic-annotation-tool/

on word embeddings to understand both the structure and semantics of sentences. Briefly, words embeddings are learned representations from texts in which words showing similar meaning also show a similar representation. Word embeddings are a class of techniques where each observed word is represented as a real-valued vector in a predefined vector space. Once word embeddings have been trained, vectors can be used to extract relations between words, such as similarities.

In the following, we describe the most common features suitable for TSA.

- *Syntactic features*: Syntactic features are the most frequently applied features, together with semantic features. Amongst the others, this class include part-of-speech tags, uni/bi/n-grams, terms' frequencies, and dependency trees. To evaluate the impact of terms on Sentiment Analysis tasks, several approaches consider the presence or the absence of terms through a binary weighting score. Uni-grams, as well as bi-grams and n-grams, are frequently exploited to train classifiers and allowing comparisons between them. Other methods consider frequencies of terms and apply a more advanced weighting schema. Also, POS tags have been widely exploited since some classes are considered good indicators of personal judgments and opinions. However, POS effectiveness is still debated due to conflicting research results. While some results do not report improvements as in [88, 121], others report at least minor improvements, as in [173, 1]. Finally, dependency trees are used to relate words with other linguistic units by directed links. In other words, dependency trees represent syntactical relations of the terms within a sentence. Typically, verbs are the sentence's centre, and are connected with the other linguistic units through syntactical relations.

- *Semantic Features*: Together with syntactic features, semantic features are the most applied. Amongst them, the more frequently used are opinion and sentiment words as well as semantic concepts, and negations. Both opinion and sentiment words can be extracted both using manual or semi-automatic methods from opinion and sentiment lexicons, respectively. Opinion words refer to terms indicating an opinion, a judgment or a personal perspective. Sentiment words typically refer to terms bearers of positive or negative sentiments. In TSA context, several works focus on the developing of methods allowing to generate polarity scores without supervision. Some works provide approaches to identify semantic features for training supervised machine learning methods. Other research aims to analyse the usefulness and effectiveness of semantic concepts. For instance, in [202], the authors exploited relations between tweets' terms and entities to improve TSA performance. As mentioned in Section 2.2.1, negations are particularly crucial features in both SA and TSA. The strength of negations regards their power to flip the texts' polarity. Due to their key role, negations, as well as their impact, have been widely explored as in [173, 119].

- *Stylistic Features*: Similarly to Twitter-specific features, stylistic features refer to the non-standard writing style typical of microblogging platforms. Stylistic features include intensifiers,

abbreviations, elongated words, acronyms, slag, punctuation marks and emoticons[12]. These features are widely used in most of the microblogs. However, Twitter seems to be the platform in which these features appear more frequently. In TSA context, the most important stylistic features are emoticons. Presence or absence of emoticons and their impacts and usefulness have been widely examined in the literature. Several lists of emoticons can be freely retrieved online. Amongst the various sources, Wikipedia provides an exhaustive set of the most common emoticons with their icons, thus the keyboard combination of symbols, and their meaning. This source has been frequently exploited for emoticon by researchers, as in [1, 159, 178, 57]. Intensifiers also are commonly used on social media platforms. Similar to emoticons' scope, intensifiers are used to improve and increase the text emphasis. The most common intensifiers may include emphatic lengthening, emphatic uppercase, and repeated characters. Finally, the fancy usage of punctuation marks, such as exclamation and question marks, may result in a useful indicator of emphasis.

- *Twitter's Specific Features*: In addition to stylistic features in common with other microblogs, Twitter includes some unique features. Besides the well-known hashtags, also replies, retweets, mentions, usernames, followers, and URLs can be used as features. The impacts of these attributes, as well as their usefulness in SA, have been widely analysed. In particular, researchers focused on both their presence versus absence and their frequencies, as in [14]. For instance, in [156] the author showed that the *hashtagged* emotion words - such as *#joy, #sadness, #angry,* and *#surprised* - are good indicators that the tweet expresses the same emotion.

2.2.3 Evaluation Metrics for Twitter Sentiment Analysis

Since the goal of a typical TSA scenario is the classification of sentiments expressed in a tweet as positive or negative, TSA can be considered as a classification problem [87]. In the literature, the performance of sentiment classification tasks is generally evaluated through four indices, namely *accuracy, precision, recall,* and *F-score*. Consider wanting to evaluate a classifier performance in classifying tweets' text as positive or negative. To visualise performances is used a table called confusion matrix, or error matrix. In a confusion matrix, each row represents the instances in a predicted class while each column represents the instances in an actual class - or vice versa [190, 213][13]. A confusion matrix shows the numbers of *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), and *False Negatives* (FN) instances, as in Table 2.1. These metrics are used to compare the ground truth with the classifier's predictions. In more detail:

- TP represents the number of instances predicted as positives which are indeed positive;

[12]Emoticons are representations of facial expressions, composed by combinations of keyboard characters. They are typically used as reinforcement of the writer's feelings or intended tone.

[13]The term confusion matrix is commonly used in case of supervised learning algorithms, while the term matching matrix refers to tables used to evaluate unsupervised learning algorithms.

		True condition	
	Total population	Positive	Negative
Predicted condition	Positive	True Positive TP	False positive FP
	Negative	False Negative FN	True Negative TN

Table 2.1: Example of a Confusion Matrix

- TN represents the number of instances predicted as negatives which are indeed negative;
- FP is the number of instances incorrectly predicted as positive;
- FN is the number of instances incorrectly predicted as negative.

In the following, we describe the four indices used to evaluate algorithms performances.

- *Accuracy*: Accuracy is the most frequently used evaluation metric. It measures how often the method provides correct predictions. It is computed as the sum of all the correct predictions - thus both TP and TN - divided by the total number of predictions - both correct and incorrect. Thus,

$$Accuracy = \frac{\sum true\ positive + \sum true\ negative}{\sum total\ population} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.1}$$

- *Precision*: Precision refers to the effectiveness of the method in terms of exactness. It is computed as the ratio of instances correctly predicted as positive divided by the total number of instances predicted as positive - meaning both correct and incorrect. Thus,

$$Precision = \frac{\sum true\ positive}{\sum predicted\ condition\ positive} = \frac{TP}{(TP + FP)} \tag{2.2}$$

- *Recall*: Recall, also sometimes called *sensitivity*, refers to the true positive rate, meaning the fraction of positive instances that are predicted as positive. Thus,

$$Recall = \frac{\sum true\ positive}{\sum condition\ positive} = \frac{TP}{(TP + FN)} \tag{2.3}$$

- *F-score*: The F-score is a metric that combines both recall and precision. It is also known as harmonic F-score, F1-score, or F-measure accuracy. Thus,

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.4}$$

2.3 Levels of Sentiment Classification

One of the most common tasks in SA consists of documents' sentiment classification. This level assumes that the targeted document expresses opinions on a single entity by one opinion holder [165]. Furthermore, the sentiment classification can be performed at a single sentence level, taking the name of sentence-level sentiment classification (Wilson, Wiebe, and Hoff- Mann, 2005). Finally, also aspects can be classified into the entity- or aspects-level. In this Section, we define these three levels of Sentiment Analysis and their contexts of usage.

According to Pang & Lee and Liu et al. [176, 135], based on the granularity levels, Sentiment Analysis has been mainly investigated at three different levels: *document-, sentence-*, and *aspect-level*.

2.3.1 Document-Level

The document-level SA consists of determining whether an opinionated document expresses an overall positive or negative opinion. For instance, determining the overall polarity of a customer review on a certain product. This level is the most coarse-grained since it assumes that a document expresses a single opinion on a single object. Generically, the result of Sentiment Analysis falls in two or three categories as positive, negative and neutral. Nevertheless, it is easy to find a few different opinions in a single document, so document-level Sentiment Analysis could not be suitable for all document types or could result in inaccurate values.

2.3.2 Sentence-Level

The sentence level of SA aims to determine whether each sentence expresses a neutral, positive, or negative opinion [85]. Since sentences can be viewed as a kind of short documents [135], no major differences occur between document- and sentence-level Sentiment Analysis. Sentence-level usually implies two sub-tasks [135]. The first is determining if a sentence is subjective or objective. Then, if the sentence is a subjective sentence, the second sub-task consist in determining whether it expresses a positive or a negative opinion. This task is related to the subjectivity classification [244] aiming in distinguishing subjective sentences that express sentiments from objective sentences that express factual information [193]. Subjectivity detection will be treated in Section 2.4.5. For now, we limit to point out that, according to Liu et al. [135], subjectivity is not equivalent to sentiment since also objective sentences can involve sentiments. For instance: "We married only one year ago, and now we are thinking to divorce". Those objective sentences are implying opinions and also belong to one subset of opinionated sentences. Compound sentences may be comparative or express different aspects of the entity. In these cases, sentence-level Sentiment Analysis is not applicable [165]. Classifying opinions at the document- or sentence-level is useful in many cases. However, these levels do not provide needed details for some tasks since they do not identify sentiment targets

[135].

Considering the document-level classification, a positive document about an item may not mean that the author has positive opinions on all the aspects of the item. The sentence-level classification of sentiment may be seen as an intermediate step [193]. It provides more fine-grained information on what entities of the object the opinions are on.

2.3.3 Aspect-Level

Aspect-level is the most fine-grained level of classification. It is also referred to as entity level or feature level by Hu and Liu [103], Pang and Lee [176], Steinberger et al. [214]. Instead of considering the constituents of documents, the aspect level of sentiment focuses on opinions itself. This level is the most suitable for determining opinions' polarity as well as identifying the opinion targets [214]. The aspect-level is based on the ground idea that each opinion is composed of sentiment - that can be positive or negative - and the target of the opinion. Opinion targets are described through entities and their different aspects. In consequence, this level of analysis aims to discover sentiments on entities and their aspects. Thus, according to Qin [193], is possible to state that like the sentence-level, also the aspect-level can involve two sub-tasks [135]: *aspect extraction* and *aspect sentiment classification*. The aspect extraction can be compared with an information extraction task aiming to extract the aspects that opinions are on. For instance, the sentence "*Ubisoft games have good quality, but their price is too high*" evaluates two different aspects of the same item - *Ubisoft games*. The sentiment on games' quality is positive, but the sentiment regarding their price is negative. In this example, "*games' quality*" and "*games' price*" are the aspects of the entity, *Ubisoft games*. A ground method [137, 27] for extracting aspects is to find frequent nouns or noun phrases - thus the aspects - and then classify the text containing the aspects as positive, negative or neutral. The aspect-level Sentiment Analysis is more challenging than both the document-level and sentence-level classifications [193]. Most of the studies [143, 131, 239] assume predefined aspects through keywords, while others [67] assume that aspects are known before the analysis.

2.4 Sentiment Classification

Traditionally, Sentiment Analysis was considered a binary classification of opinion [63, 177]. However, Esuli and Sebastiani [80] suggest that the sentiment classification issue can be divided into three specific sub-tasks:

- *determining subjectivity*, as determining whether a given text has a factual nature without expressing an opinion on it or on its subject matter - be it positive or negative. This type of task belongs to the Subjective Detection [174, 255] (see Section 2.4.5);
- *determining orientation* or *polarity*, as determining whether a given subjective text expresses

an opinion on its subject matter - be it positive or negative [174, 226];

- *determining the strength of orientation*, as determining whether an opinion expressed by a text is weakly, mildly or strongly positive or negative [248].

Research in Sentiment Analysis is divided into two main approaches [144, 53], namely Semantic Orientation Approach and Machine Learning Approach. In more detail, the semantic orientation approach is also named as the lexicon-based approach and is based on words and phrases as indicators of the semantic orientation. Usually, the overall polarity of the text is computed as the average sum of polarities of these indicators [176, 98]. The machine learning approach is based on machine learning techniques and on the choice of the best features to be considered to classify the polarity of a given text [177].

In the following, we review Semantic orientation Approaches and Machine Learning Approaches by investigating classes of methods belonging to these two categories, such as Lexicon-based and Corpus-based Approaches and Feature Selection and Machine Learning Algorithms, as well as Hybrid Methods.

2.4.1 Machine Learning Approaches

Machine-learning (ML) approaches are an application of Artificial Intelligence (AI) which provide systems with the ability to learn from experience and data automatically. Machine-learning research field focuses on the development of systems able to access data and use it to learn for themselves. While a supervised learning algorithm learns to map the input examples to the expected target, an unsupervised ML algorithm is capable of generalising the learned knowledge after the implementation of the training process. In the SA context, the ground idea is to train functions which allow determining the sentiment orientation of unknown document by using a corpus of documents tagged with sentiments. Most of the methods dealing with social media Sentiment Analysis, such as Twitter Sentiment Analysis (TSA), employs ML classifiers trained on several features. In particular, ML algorithms for TSA typically exploits Twitter-specific features, including hashtags and emoticons - i.e., presence vs absence, or type. In the TSA literature, the most applied ML classifiers include Support Vector Machines (SVM), Naïve Bayes (NB) and Multinomial Naïve Bayes (MNB), Maximum Entropy (MaxEnt), Random Forest (RF), and Logistic Regression (LR). In the following, we briefly review some TSA-related works exploiting these approaches.

Go et al. [88] carried out one of the first studies dealing with TSA. The authors classify the tweets as positive or negative by handling the classification problem as a binary classification. Tweets are gathered following the method proposed by Read [196] which uses emoticons as labels. Tweets are then tagged by a machine-learning classifier built by exploiting distant supervision. As Pang et al. [177], the authors examined NB, MaxEnt, and SVM classifiers. Moreover, they used bigrams, unigrams, and POS tags as features. Results show that NB with bigrams as features achieve the best

accuracy (82.7%). Emoticons are used as labels also by Pak and Paroubek [173]. Despite this, the authors handle the classification as a multiclass classification task tagging tweets as positive, negative, or neutral. Performance of MNB, Conditional Random Field (CRF), and SVM are compared using different features, such as bigrams and n-grams. The performances obtained by the authors show that the best combination is MNB together with n-grams and POS tags.

In [14] a two-step classifier is presented to deal with the TSA problem. The first step aims in determining if a tweet is opinionated, while the second classifies a tweet as positive or negative. The authors' result shows that SVM classifier obtains the best accuracy for both subjectivity detection (81.9%) and polarity detection (81.3%). SVM classifiers for addressing TSA are also used in [11] and [159]. In [11], the authors employ an SVM classifier trained on more than ten features. The best results are achieved by combining NLP- and Twitter-specific features. In [159], the authors represent tweets as a feature vector. Each vector includes several features, including n-grams, POS, emoticons, lexicon features, and negations. Authors' results show that the performance achieved by the SVM classifier trained using features are better than the baseline trained on unigrams.

The authors of [8] present a three-step cascaded classifier framework. The first step aims in identifying tweets of the topic of interest. The second step determines opinionated tweets, while the third finally annotate tweets with sentiment polarity. The authors computed the TSA performance of several well-known methods as well as new algorithms, including k-Nearest Neighbours algorithm (kNN), NB, weighted SVM, and Dictionary Learning. Their results show that classification performances are improved in a low-dimensional space.

According to [87], it seems that the machine-learning approach is the most popular on TSA tasks. Generally, approaches employ traditional machine-learning methods trained on a set of features. As shown above, features can include unigrams, n-grams, lexicon features, and POS tags. The feature selection plays an essential role in the effectiveness of the supervised methods. To this end, several works have been developed to identify features that achieve best performances. Despite being widely used, machine-learning methods have some limitations. Since their performance depends on the size of the training data, they typically require a large amount of annotated tweets. However, annotate large dataset is very expensive. As pointed out in [87], distant supervision can represent an alternative though the annotation quality is typically low and can negatively impact classifiers' performances. Another machine-learning approaches limitation is that they are domain-dependent. This means that a classifier can perform very well on the domain on which it is trained, but it may need to be retrained to perform well on a different domain. Finally, as mentioned before, machine-learning approaches depend on the set of selected features. Thus phenomena such as negation detection may not be captured. As a consequence, performances could be negatively influenced.

2.4.2 Semantic Orientation Approaches

The so-called semantic orientation of a word is defined by Lehrer et al. [130] as the feature indicating the direction of the deviation from the norm of its semantic group or lexical field. Depending on scholars it is also named *polarity* or *valence*. Word's semantic orientation can assume various directions - such as positive, negative or neutral - as well as intensity - such as weak, mild or strong [227]. A word having positive semantic orientation denotes a desirable state (i.e., *fantastic*, *beautiful*), while a word with negative semantic orientation refers to undesirable states (i.e., *anger*, *disgust*) [99]. Several studies [98, 226, 244] have already proven the power of polarised words, especially adjectives, as indicators of subjectivity. Following this approach, usually, the overall semantic orientation of a whole text is computed based on the sum of polarised indicators. The semantic orientation approach fits into unsupervised learning since it does not require a previous training phase to mining data. Most of the researchers approach the unsupervised sentiment classification by exploiting lexical resources available [179]. For instance, Esuli and Sebastiani [80] proposed a semi-supervised learning method which starts by expanding an initial set of seed-words using WordNet. Moreover, Chunxu Wu [251] proposed an approach to address scenarios where texts are insufficient for determining the orientation of opinion. The method exploits other feedbacks discussing a similar topic and used semantic similarity measures to check the orientation of opinion. Finally, in [53], Chaovalit and Zhou compared performances of Semantic Orientation approach and other machine learning approaches (see Section 2.4.1) on movie reviews. Machine learning approaches provide accurate results but they require an expensive training phase. On the contrary, the semantic orientation approach typically provides less accurate results but is suitable for real-time applications. However, the performance of semantic orientation approach is strictly dependent on the performance of the POS tagger used.

Lexicon-based Approaches

Lexicon-based methods typically exploit annotated lists of words. Words' annotations consist of a polarity dimension to obtain an overall polarity score of a given text. In the following, we describe and review existing lexical knowledge about sentiment classification.

Several research fields lack training data. When training data are scarce, applying supervised models become quite challenging or even impractical. In these contexts, lexicon-based approaches can turn out to be useful since they do not require training data. Lexicon-based methods have already been extensively applied on most common texts such as product reviews, forums, documents and blogs [216, 226, 87]. Nevertheless, in social network Sentiment Analysis - as TSA - these approaches are still less used than machine-learning methods. Motivations are due to the inherent nature of these texts [87]. Indeed, they typically contain several textual peculiarities, non-standard textual register as well as a dynamic nature characterised by new expressions and symbol combinations. It is possible to identify two main types of lexicon-based approach: *a*) unsupervised models which do not require training data, and *b*) mixed models combining labelled texts and lexical knowledge

[87]. To compute texts polarity, several ways can be explored. One of the most simple approaches is based only on the lexicon knowledge. These methods allow aggregating the polarity orientation of each known term contained in the observed text [99]. For instance, based on words' labels, it is possible to obtain the overall polarity score by computing the difference between positive and negative scores of words in a given text. A more sophisticated aggregation-based approach could also exploit linguistic knowledge. For example, specific rules for negations and opinion shifters can be taken into account [216, 221]. Particularly suited for social media, SentiStrength [221] is one of the most famous lexicon-based algorithms. The model identifies sentiment strength of informal text exploiting a human-tagged lexicon. In conjunction with a sentiment lexicon of about 700 words, SentiStrength uses a set of negations and boosting words, as well as emoticons to compute polarity scores. Initially tested on comments on MySpace, the algorithm was updated [220] introducing new sentiment-labelled words and idioms lists. The polarity orientation of two words and co-occurrence measures are the basis of the well-known model proposed by Turney [226]. The hypothesis of the approach is that a lexical rule could be useful to identify opinionated phrases. In more detail, given a phrase p, if p contains a sequence of adjective adj or an adverb adv, such as $< adj, adv >$, p probably expresses an opinion. Due to this, by exploiting a part-of-speech tagger, words' lexical categories are identified. Common POS categories are verb, noun, adjective, adverb, pronoun, preposition, conjunction and interjection. Then, all sentences satisfying the co-occurrence pattern are extracted. The sentiment of each phrase is then computed by using the point-wise mutual information (PMI) [58].

$$PMI_{(w_1,w_2)} = log\left(\frac{Pr(w_1 \wedge w_2)}{Pr(w_1)Pr(w_2)}\right) \tag{2.5}$$

The PMI provides a measure of statistical independence between two words. Thus, to compute the phrase's semantic orientation, the PMI value is computed against a positive and a negative word (w_{pos} and w_{neg} respectively). As shown in Formula 2.6, to obtain the so-called PMI-SO, the difference between the first and the second value is computed:

$$PMI - SO_p = PMI(p, w_{pos}) - PMI(p, w_{neg}) \tag{2.6}$$

PMI values' probabilities are calculated by using the frequency value which in turn is computed as the number of hits returned by a given query. In particular, a first query is composed by the phrase and a given positive word while a second is composed by the phrase and a given negative word. The overall PMI-SO of a document is computed as the average PMI-SO of the phrases which compose the document. Thus, if the overall PMI-SO value is positive, the document is tagged as positive, negative otherwise. A three-step model is proposed by Ortega et al. in [170] for the SemEval-2013[14] competition. As in [226], the first step consists in a pre-processing phase, while the second step

[14]SemEval-2013 web site: https://www.cs.york.ac.uk/semeval-2013/

relies on polarity detection. The final step regards a rule-based polarity classification. The polarity detection accomplished in the second step, as well as the classification, are based on WordNet and SentiWordNet. Amongst others, the SemEval-2013 dataset was also used by Reckman et al. [197] to develop a multi-rule-based system called Teragram. In more detail, each rule is handwritten and is considered as a pattern. The proposed method is particularly suited for TSA and ranks between the top-performing systems in SemEval-2013 competition. Finally, a lexicon-based approach was also presented by Saif et al. [203]. The authors have proposed SentiCircles, a model developed for TSA and evaluated over different datasets such as OMD, HCR, and STS-Gold [207, 211, 159]. SentiCircles, based on co-occurrences in different contexts, update both scores and polarity of words of a sentiment lexicon. The method was extensively tested proving its effectiveness and outperforming the methods based on MPQA and SentiWordNet. Besides the described approaches, several works extend sets of sentiment words exploiting semantic relationships. In a typical scenario, a small amount of sentiment words is identified and manually annotated and then is extended by adding words with similar semantics [117]. The problem with these approaches typically relies on the expansion of the initial set of seed that is restricted and dependant on the initial set of words. To avoid this issue, Feng et al. [84] proposed to use a connotation lexicon to enclose subtle dimensions of a words sentiment [87]. After the definition of a set of seed words, the authors apply a graph-based algorithm based on HITS and PageRank. This step allowed them to built a connotation lexicon which also includes connotative predicates.

To classify the sentiment of documents, hybrid models combine both opinionated words and sentiment-tagged documents. The most straightforward scenario foresees the use of emotive words to compute aggregated features in supervised classification schemes. For instance, a basic feature could consist of the number of positive and negative words found. In hybrid scenarios, lexicon-based features often provide exhaustive generalisation properties since they allow to include lexical information that relies on words not necessarily contained in training data. In [145] a generative naive Bayes model based on a polarity lexicon in presented. The authors developed an effective framework able to incorporate lexical knowledge in supervised learning for text categorisation. Moreover, they apply the framework to accomplish sentiment classification tasks. In [208], both documents and words are jointly represented through a bipartite graph composed of both labelled and unlabelled nodes. Then, using a sort of grid of regularised least squares, sentiment score is propagated from labelled nodes to unlabelled.

2.4.3 Hybrid Approaches (Machine Learning & Lexicon-Based)

Several researchers have combined machine-learning and lexicon-based approaches. These approaches are commonly defined as Hybrid. In [86], dynamic artificial neural networks are combined with n-grams. Firstly, tweets containing emoticons or words lexically related - i.e., synonyms - with Love or

Hate terms are used as features to build two different classifiers, an SVM and a Dynamic Architecture for Artificial Neural Networks (DAN2). Then, classifiers are tested and compared on a corpus of tweets having Justin Bieber as the subject. Provided results show that the DAN2 performance outperforms SVM. A hybrid unsupervised method to address entity-based TSA is presented in [260]. To summarise, given a sentence s containing the user-given entity, opinion words in the sentence are first identified by matching with the words in the opinion lexicon. Then, an orientation score for the entity e is computed. In more detail, first, the authors selected a set of five entities, namely *Obama, Harry Potter, Tangled, iPad,* and *Packers*. Then, using entities as queries' keywords, five diverse Twitter data sets are built. Based on proximity to words from a sentiment lexicon gathered from [66], semantic orientation scores are assigned to words themselves. Scores range from +1 - positive - to -1 - negative. Then, all the scores are summed up as follows:

$$score(f) = \sum w_i : w_i \in s \wedge w_i \in V \frac{w_i * SO}{dis(w_i, f)} \tag{2.7}$$

where w_i is an opinion word, V is the opinion lexicon, s is the sentence that contains the feature f, and $dis(w_i, f)$ is the distance between feature f and opinion word w_i in the sentence s. Thus, $w_i * SO$ is the semantic orientation score of the word w_i. The multiplicative inverse in the formula is used to give low weights to opinion words that are far away from the feature f. Through the Chi-square test on the results of the lexicon-based method, additional opinionated indicators - i.e., words and tokens - are identified. These are in turn used to identify additional opinionated tweets. Finally, a binary sentiment classifier is then trained to assign sentiment polarities to the opinionated tweets identified after the Chi-square test phase. In [115] a lexicon-based method is combined with a classifier to enhance the classification accuracy. Exploiting the MapReduce framework, a co-occurrence matrix based on bigram sentences is built. By using the cosine similarity between words, only links having high cosine score are maintained. Then, a sentiment classifier is combined with scores obtained by using a lexicon-based approach. The proposed machine learning algorithm uses an Online Logistic Regression algorithm from the Apache Mahout framework[15]. The choice is due to both the fast training time and the ability to adjust the model with new data. Provided results show how this hybrid approach outperformed the lexicon-based classifier which is based on emotive words or sentences. Also in [123], the semantic orientation of specific parts of speech is calculated based on a log-linear classifier combined with a dictionary-based method. The authors focus on the semantic orientation of adjectives, verbs, and adverbs. Then, tweets' overall sentiment is derived through a simple linear equation. Finally, in [113] a hybrid TSA model was included in a three-step framework called TOM. The first step of TOM regards the data acquisition and preprocessing phases. After that, preprocessed tweets are treated through a sentiment classifier. The Polarity Classification Algorithm (PCA) sentiment classifier detected sentiment on a tweet based on three

[15]Apache Mahout, a scalable machine learning library: https://mahout.apache.org/

sub-classifiers, such as Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC), and SentiWordNet Classifier (SWNC). These respectively use a set of emoticons, a list of sentiment words, and SentiWordNet dictionary. Obtained results show that overall hybrid classification managed to outperform the performance of using any of the EEC, IPC, and SWNC classifiers.

2.4.4 Comparing Semantic Orientation and ML Approaches

In Sentiment Analysis as well as Twitter Sentiment Analysis, machine-learning approaches seem to be the most popular. Moreover, the majority of contributions employ traditional and standard machine-learning methods typically trained on a set of features. However, as mentioned above, some researchers attempt to improve classification performances by combining several classifiers. These contributions show that classifiers ensembles tend to perform better than using a single classifier.

In the ML panorama, supervised machine learning techniques tend to achieve better performance than unsupervised methods. Supervised methods require huge amounts of labelled training data which are expensive to collect. On the contrary, unlabelled data are more easy to gather and less costly. Moreover, unsupervised methods are particularly suited for domains which lack labelled training data. However, machine learning methods involve some limitations. First, the performances are training-data-dependent. As a consequence, to obtain high performances, they require a large amount of labelled data. This could lead to an increase in costs since the annotation process is expensive. Other approaches to address annotation include distant supervision and label propagation. However, in particular with using distant supervision, annotation quality is low. Another drawback of machine learning approaches relies on their domain-dependence. This means a classifier could work well on the same domain to the one it is trained. However, the same classifier may decrease in performance if applied to other domains. From a machine learning point of view, most researchers agree that SVM outperforms other algorithms in both costs and accuracy. However, while the performance of algorithms such as SVM and Naïve Bayes are comparable, the performance of the Decision Tree algorithm is far below the others. Finally, it is crucial to take into account that ML is not able to capture some phenomena, such as negation detection.

Besides the huge amount of ML contributions, some works applied lexicon-based techniques. The major strength of these methods is that they do not involve training or labelled data. However, they typically employ dictionaries or lists of words. Is important to note that word-lists are static. Moreover, in these approaches, a given word is considered only if it is contained in the list. That implies that word-lists need to be often updated to be effective. In the TSA, this becomes more crucial. On Twitter, both contents and language change continuously and also evolve. Another major limit relies on words' contexts. Lexicons are context-independent. This means that these lexicons do not consider words' sentiments by the context in which they are involved. As a classic example, we refer to the word "small". In sentences as "*My new smartphone is small and fits well into my purse*", the word *small* expresses a positive opinion. On the contrary, by considering a

sentence like "*The buttons on my old phone were very/too small*", the sentiment carried by the term *small* is negative [87]. Finally, we also underline that in multi-language contexts, applying sentiment lexicons often implies the translation of the entire initial bag of words or the retrieval of multiple seed lexicons, such as one for each language. These techniques typically impose expensive preprocessing phase to allow to merge sources, when those are available[16].

Bearing in mind all the limitations of both machine learning and lexicon-based approaches, some hybrid methods have been proposed. The major benefit of hybrid methods is that they attempt to offset lexicon-based limitations with ML approaches and vice versa. For instance, training data manual labelling can be avoided by exploiting a lexicon-based method [87].

2.4.5 Subjectivity Detection

Subjectivity detection is an essential subtask of the Sentiment Analysis research field [54]. This is primarily due to the dichotomous optimisation of most of polarity detection tools. Polarity detection tools are typically developed to distinguish between positive and negative text. Thus, subjectivity detection becomes crucial to ensure that factual information is filtered and only opinionated texts are transmitted to the polarity classifier [136]. It should be noted that in many cases a document in a collection contains only factual information. Moreover, an opinionated document may contain both opinionated and non-opinionated sentences. Previous studies revealed that subjective content represents only 60% of documents as product reviews. Moreover, subjective content provides similar polarity results as full-text classification [34]. For instance, product reviews on Amazon include several neutral reviews. This makes the text interpretation more difficult. Also, sentiment classification tasks typically assume that observed documents are opinionated. The process of labelling sentences as subjective or objective is challenging also for human annotators [19]. This is due to the inherent ambiguity of texts and the combination of both subjective and objective frameworks in the same text or even sentence. Moreover, issues also refer to the huge computational costs of existing n-gram methods which are based on the syntactical representation of text, such as word-sense disambiguation or part-of-speech tagging. For instance, in [228], the authors noted that subjective sentences indicating a purchase interest often contain modal verbs.

Subjectivity detection is thus related to determining whether a sentence is subjective or neutral [246]. The issue can be addressed through supervised learning approaches. For instance, in [255] and in [244] a subjectivity classifier is trained on a corpus of journal articles by using Naive Bayes. According to [38], in TSA context, the subjectivity detection task seems to be a more challenging task than polarity classification. However, some researchers jointly address polarity and subjectivity detection. In these cases, the joint task can be dealing with a three-class classification (neutral,

[16]Note that sentiment lexicons are typically available in English and few other major International languages. For more details on the most famous available lexicons see Section 2.6.

positive, and negative). Lastly, it is crucial to note that an accurate, neutral class detection becomes fundamental in SA classification tasks. In fact, according to [120], while neutral training data improves the classification of both positive and negative documents, the knowledge learned by only positive and negative documents may not be generalised to neutral contents.

2.5 Polarity Lexicon Induction

In previous sections, we underline how lexical knowledge applied to Sentiment Analysis can improve polarity classification tasks. As mentioned before, a polarity lexicon - also called opinion or sentiment lexicon - consist of a list of words labelled with sentiment orientations. This latter can be found in several forms, such as lexical tag, numerical measure or binary scale. Lexicons can be manually or automatically created, even according to the domain they refer. Manually created lexicons are built by gathering words or lemmas from several sources and determining their sentiment orientation by human-taggers. Generally, manually annotating a lexicon is an expensive, time-consuming and labour intensive task. Therefore, the labelling phase is often conducted through crowdsourcing platforms, such as the already mentioned Amazon Mechanical Turk platform. Automatically-created lexicons have the advantage of being less costly and faster in the building. However, they are typically less accurate than manually-created ones. Two different types of sources can be used to build a lexicon automatically: *a*) document collections, and *b*) semantic networks.

In the following two subsections, we present the works focusing on sentiment lexicon induction from document collections and semantic network resources. Moreover, we describe the most popular lexicons in the Sentiment Analysis state of the art.

2.5.1 Corpus-based Approaches

Corpus-based approaches typically exploit lexical patterns as co-occurrence or syntactic patterns to induce lexicons. In these contexts, lexicons are composed of words or lemmas extracted from collections of unstructured text documents [38]. Generally, corpus-based methods imply data-driven approaches. That allows access to sentiment labels and to contexts which could be useful in ML algorithms. Indeed, a corpus inherently also includes domain-specific features which can be exploited in algorithms to identify sentiments labels also depending on a given context. In [98], the semantic orientation of a set of adjectives is used to discover new adjectives and their related semantic orientation. Using a log-linear regression, the authors demonstrate that conjunctions placed between two adjectives could provide information about the adjectives' semantic orientation themselves. For instance, given two adjectives, if they are connected by the conjunction "and", they will tend to have the same semantic orientation. On the contrary, if the adjectives are connected by the conjunction "but", they will tend to have opposite semantic orientation. Through this method, the authors can extract domain-dependent information as well as variations occurring when the corpus is changed.

Several works have been developed by researchers for lexicon induction, especially starting from Twitter, as in [4]. Most of them are based on the association between words and message-level sentiment labels, as in [119, 159, 262]. In [4], a general method is presented to extract large-scale domain-specific sentiment lexicons with fine-grained annotations from Twitter data. The model addresses the task as a regression problem, where terms are represented as word embeddings. Manually annotated lexicons are used for training the regression model to predict both intensity and polarity of both words and sentences in the Twitter corpus. In [262], domain-specific lexicons are built on an emoticon-annotation approach. Similarly, in [159] and [119], emoticons and hashtags associated with emotions are used to provide sentiment labels to tweets. Then tagged data are exploited to create two different emotion lexicons which in turn are used for tweet-level polarity classification. Moreover, in [158], a Twitter corpus - namely Hashtag Emotion Corpus - is extracted based on specific hashtags related to emotions. In particular, the authors selected six hashtags corresponding to the six Ekman basic emotions: *#happy, #disgust, #fear, #sadness, #anger* and *#surprise*. A Twitter-specific emotion-association lexicon is then created by relying on the Hashtag Emotion Corpus. All unigrams and bigrams have been mapped as strength association scores of the six observed emotions. Finally, based on PMI formula, the *Strength of Association* (SoA) between an n-gram w and an emotion e is calculated as

$$SoA_{(w,e)} = PMI_{(w,e)} - PMI_{(w,\neg e)} \tag{2.8}$$

where PMI is calculated as

$$PMI_{(w,e)} = log_2 \frac{freq_{(w,e)} * N}{freq_{(w)} * freq_{(e)}} \tag{2.9}$$

where $freq(w, e)$ is the number of times w occurs in a sentence with label e; $freq(w)$ and $freq(e)$ are the frequencies of w and e in the labelled corpus, and N is the number of words in the dataset.

$$PMI_{(w,\neg e)} = log_2 \frac{freq_{(w,\neg e)} * N}{freq_{(w)} * freq_{(\neg e)}} \tag{2.10}$$

Where $freq(w, \neg e)$ is the number of times w occurs in a sentence that does not have the label e; $freq(\neg e)$ is the number of sentences that do not have the label e. Thus, Formula 2.8 can be simplified as follow:

$$SoA_{(w,e)} = log_2 \frac{freq_{(w,e)} * freq_{(\neg e)}}{freq_{(e)} * freq_{(w,\neg e)}} \tag{2.11}$$

As a consequence, if an n-gram often occurs both in a sentence with a particular emotion label and in a sentence that does not have that label, then that n-gram-emotion pair will have an SoA score that is greater than zero.

2.5.2 Semantic Networks

A semantic network - or frame network - can be formally defined as a knowledge base that represents semantic relations between concepts in a network. In other words, it is a network representing semantic relations between concepts. In the lexicon induction scenario, the easiest approach foresees to expand an initial seed lexicon of labelled sentiment words by using lexical relations, such as synonyms and antonyms [117, 103]. As a consequence, these approaches are based on the hypothesis that synonyms have the same polarity, while antonyms have the opposite polarity. In [80] a method is proposed for determining the orientation of terms by using a semi-supervised classifier exploiting a seed of labelled words. The ground hypothesis is that terms having similar semantic orientation tend to show similar *glosses*, i.e., the definitions that these terms are given in on-line dictionaries, and on the use of the resulting term representations for semi-supervised term classification [80]. For instance, lemmas such as *honest* and *intrepid* have both glosses containing appreciative expressions. On the contrary, the glosses of *disturbing* and *superfluous* both contain disparaging expressions. A similar approach has been used to create SentiWordNet [81, 9]. In SentiWordNet, a publicly available lexical resource, each WordNet synset is labelled with a semantic orientation ranging from 0 to 1 according to the standard emotional classes, thus positive, negative or neutral[17]. WordNet is also used in [112] to create a graph. In detail, WordNet adjectives are represented as vertices, while relations between synonyms are displayed as links. Following Charles Osgood's Theory of Semantic Differentiation[18] [171], the polarity orientation of a term is typically defined by its evaluative dimension. Given that, the authors of [112] have associated an Evaluative function (EVA) to each term. The EVA function measures the relative distance from the two reference seed words "good" and "bad" as

$$EVA_{(w)} = \frac{d(w, bad) - d(w, good)}{d(good, bad)} \tag{2.12}$$

As a consequence, each word is assigned to an EVA function awarding a value ranging from -1, representing words on the "bad" side of the lexicon, to 1, representing words on the "good" side of the lexicon. Finally, we report the example used by the authors

$$EVA_{(honest)} = \frac{d(honest, bad) - d(honest, good)}{d(good, bad)} \tag{2.13}$$

[17] More details about SentiWordNet are provided in Section 2.6.

[18] Charles Osgood's Theory of Semantic Differentiation [171] highlights "latent cognitive structures" referred to three different dimensions. Each dimension corresponds to a factor reflecting the subjective attitude to the entity investigated. The three factors of the emotive meaning are the evaluative factor (e.g., good-bad); the potency factor (e.g., strong-weak); and the activity factor (e.g., active-passive).

Exploiting the same approach, the authors define also measures for other Osgood's dimensions, such as Potency (POT) and Activity (ACT). The potency factor of w is defined as

$$POT_{(w)} = \frac{d(w, weak) - d(w, strong)}{d(strong, weak)} \tag{2.14}$$

while the activity factor of w is defined as

$$ACT_{(w)} = \frac{d(w, passive) - d(w, active)}{d(active, passive)} \tag{2.15}$$

Thus, this method allows the authors to define measures for any two connected words in WordNet. As SentiWordNet, also SenticNet is a well-known lexical resource for Sentiment Analysis built on semantic networks. SenticNet is based on the *sentic* computing paradigm, a multi-disciplinary approach to Sentiment Analysis at the crossroads between affective computing and common-sense computing [43]. Among SenticNet's updates and versions [19], the first two have been built applying graph-mining and dimensionality-reduction techniques [38].

ConceptNet[20] is a semantic network of commonsense knowledge in which assertion are composed of two concepts connected by a relation. In other words, ConceptNet is a multilingual knowledge base, representing both words and sentences and the common-sense relationships between them. Data in ConceptNet is collected from several resources, including crowd-sourced resources - such as Wiktionary[21] and Open Mind Common Sense[22] -, domain-specific and researcher-created contents - such as WordNet and JMDict[23] -, games with a purpose - such as Verbosity [234] and nadya.jp. For instance, given the word *graph*, one can find *graph IsMadeOf a set of vertices and a set of edges*, or *graph IsA visual communication; a diagram; graphical*. Besides the two mentioned above - *IsA* and *IsMadeOf* -, the framework makes available 33 different types of relations, including *PartOf*, *UsedFor*, *CapableOf*. The resource has been widely exploited for lexicon induction, as in [224, 250, 242]. In [224], ConceptNet's concepts are extracted and tagged with a sentiment score by using iterative regressions. Sentiment iterations are propagated by random walks. The model proposed by Tsai et al. [224] has been improved in [250]. The improvement regards the addition of a bias correction step after the random walk process. The correction step has been introduced to reduce the volatility of polarities. To find the best combination of sentiment and relations from ConceptNet, the authors use the sequential forward search. Moreover, in [242], ConceptNet is used to avoid the typical lack of contextual information of opinion lexicons. The proposed model focuses on the contextualisation

[19] More details about SenticNet are provided in Section 2.6.
[20] ConceptNet: http://conceptnet5.media.mit.edu/
[21] Wiktionary: https://www.wiktionary.org/
[22] Open Mind Common Sense: https://www.media.mit.edu/projects/open-mind-common-sense/overview/
[23] JMDict: https://www.edrdg.org/jmdict/j_jmdict.html

and enrichment of large semantic knowledge bases through three steps. In detail, *a*) identify ambiguous sentiment terms, *b*) provide context information extracted from a domain-specific training corpus, and *c*) associate the contextual information with structured background knowledge sources, such as ConceptNet and WordNet.

Despite the effort of researchers, lexicons built through semantic networks show two major weaknesses. The first relies on the inability to capture emotive information about concepts or words not included in the observed semantic network. The second is strongly connected with the usage of terms. Semantic networks like WordNet and ConceptNet are based on formal English. As a consequence, resources expanded from these networks typically lack in informal expressions. This may lead to limitations when lexicons expanded via a semantic network are applied to social platforms contexts such as Twitter.

2.6 Lexical Resources in Sentiment Analysis

Part of the work carried out for this thesis concerned the use of some lexical resources, e.g., the Affective Norms for English Words lexicon [37] and SentiWordNet [81]. For this reason, below we present the most popular lexical resources.

ANEW. The Affective Norms for English Words lexicon (ANEW) is an emotive lexicon proposed by Bradley and Lang [37]. The lexicon provides a set of normative emotional ratings for 1,034 English terms collected from human subjects. Provided ratings are computed according to three psychological reactions to the observed word. In more detail, the lexicon provides emotional ratings in terms of *valence*, namely the level of pleasantness, *dominance*, thus the degree of control, and *arousal*, namely the intensity of emotion. *Pleasure* represents positive versus negative emotions. *Arousal* finds the two extremes of the scale of values in "calm" and "exciting". The *dominance* dimension determines if the subject feels in control of the situation or not. Each dimension is represented as a number between 0 and 10. Moreover, the lexicon also provides ratings depending on taggers' gender. Amongst all values, the valence dimension, which ranges in the scale from pleasant to unpleasant, is the most used for polarity calculation.

SENTIWORDNET. SentiWordNet [81] is a publicly available lexical resource in which words are associated with emotional values describing how objective, positive, and negative they are. The resource is based on a well-known lexical database for English, WordNet. This latter resource is a set of words clustered into groups of synonyms called synsets [151]. In more detail, SentiWordNet exploits the glosses of WordNet synsets as semantic representations of the synsets themselves and classifies each synset *(s)* into three categories, such as *Pos(s)*, *Neg(s)* and *Obj(s)*. The resource has been updated over time [9]:

1. SentiWordNet 1.0 was presented in [81]. The resource consists of an annotation of the older WordNet 2.0 and was publicly made available for research purposes;
2. SentiWordNet 1.1 was only discussed in a technical report [82] and was never made public;
3. SentiWordNet 2.0 was never made public and is only discussed in the author's PhD thesis [6];
4. SentiWordNet 3.0 [9], is actually the latest version and is based on the WordNet's 3.0 version.

EmoLex. The NRC word-emotion association Lexicon (EmoLex) [160] is a list of more than 14,000 distinct English words annotated according to both emotion and sentiment categories. In more detail, for each word w, the lexicon provides the w's associations with eight basic emotions coming from the Plutchick wheel of emotions [183] (*anger, fear, anticipation, trust, surprise, sadness, joy*, and *disgust*) and two sentiments, such as *negative* and *positive*. The annotations were manually done using the crowdsourcing Amazon Mechanical Turk platform[24]. All the provided categories are not mutually exclusive. Thus each word can be tagged with multiple emotions and polarities. Moreover, the lexicon also includes words not associated with any emotion or polarity category, namely neutral words.

MPQA Subjectivity Lexicon. The MPQA Subjectivity Lexicon (list of subjectivity clues) is a lexical resource proposed by Wilson et al. [247]. The resource is part of OpinionFinder [56], an automatic system for detecting subjective sentences in corpora. The resource is composed of a little less than 7,000 English words tagged according to traditional polarity classes (*positive, negative*, and *neutral*) by human annotators. The list also includes a small set of words tagged with both negative and positive value.

AFINN Lexicon. The AFINN lexicon is a list or around 2,500 terms manually tagged with an emotional value ranging from -5 (negative) to +5 (positive) by Finn Arup Nielsen. In more detail, positive words are scored from 1 to 5, while negative words from -1 to -5. Even though the lexicon was inspired by ANEW [37], it is particularly focused on the language used in microblogging platforms. Due to this, the set also includes obscene words, slang, acronyms and Web jargon.

SentiStrength. SentiStrength [221] is a lexicon-based Sentiment Analysis method that estimates the strength of positive and negative sentiment in short texts. The lexicon is composed by English words - both formal and deriving from social media - manually annotated with a score ranging between +5 (*positive*) and -5 (*negative*). In details, the positive score ranges between 1 (*not positive*) and 5 (*extremely positive*), while the negative one ranges between -1 (*not negative*) and -5 (*extremely negative*). Moreover, the resource allows to obtain binary (*positive* and *negative*), ternary

[24]Amazon Mechanical Turk platform: https://www.mturk.com/

(*positive, negative* and *neutral*) and single scale (-4 to +4) scores.

SenticNet. SenticNet is a concept-level semantic network for Sentiment Analysis. The resource provides both sentiment and semantic annotation for about 30,000 knowledge concepts. Moreover, it also includes a parser able to return two sentiment variables, as the polarity score and a *sentic* vector. The polarity score is a real value. The *sentic* vector is an emotion-oriented value related to four base-ground emotions: pleasantness, attention, sensitivity, and aptitude. These emotions have been identified according to the Hourglass model of emotions [44], a model inspired by Plutchiks theory on basic human emotions.

Sentiment140. The Sentiment140 Lexicon [215] is a lexicon provided by the NRC-Canada team, as the NRC-Hashtag Sentiment Lexicon resource. Instead of using hashtags as tweet labels, the authors exploit a 1.6 million corpus of tweets which include positive and negative emoticons. Then they apply the same procedure already used for building the Twitter Emotion Corpus [155].

NRC-Hashtag/TEC. The NRC-Hashtag Sentiment Lexicon (or Twitter Emotion Corpus, or TEC) [155] is an automatically created sentiment lexicon. The resource is built starting from a set of over 750,000 tweets containing emotional hashtags, such as `#good`, `#bad`, `#terrible`, and `#excellent`. Each tweet is labelled according to the polarity of hashtags inside it. Emotions categories refer to the eight basic emotions theorised by Plutchik in the Hourglass model of emotions. Then, using the pointwise mutual information (PMI) [58] measure between each word and the corresponding polarity label of the tweet, a sentiment score is computed for each word. The resource is composed by over 16,500 terms tagged with a real-valued score between 0 (not associated) to ∞ (maximally associated).

Harvard General Inquirer. The Harvard General Inquirer is a lexicon proposed by Stone et al. [111]. It provides about 1,900 positive terms and over 2,000 negative terms. Words are tagged according to a broad set of multiple dimensions such as polarity, emotions, and semantics. In more detail, the semantic category is, in turn, split according to Charles Osgood's semantic differential findings regarding primary language universals. The polarity classification has been further partitioned providing more focus than the traditional binary categories. Following this partition, the system identifies four additional dimensions, such as pleasure, pain, virtue and vice. The firsts two typically refers to positive and negative terms, while the two latter indicate strength and weakness respectively. Additional categories focus on overstatement and understatement, generally reflecting the presence or the lack of emotional expressiveness [25].

[25] A detailed and comprehensive list of provided tags can be found in Descriptions of Inquirer Categories and Use of Inquirer Dictionaries at https://urly.it/31b6g

2.7 Making Predictions using Social Media

In this section, we address and describe both motivations and areas where prediction with social media may be made.

According to Yu and Kak [256], currently, most predictions using social media can be made better by human agents such as specifically experts. However, the automatic prediction is motivated by several good reasons [256]. Firstly, automatic prediction requires fewer costs than human labour [35]. Secondly, human labour is intentionally or unintentionally, affected by bias, intents, and ideas of the labourer. As a consequence, human labour may be not purely based on objective probability [249]. Thirdly, the automatic prediction allows processing a more significant amount of data and providing faster - or real-time - results. Finally, people tend to undervalue high probabilities and overvalue small probabilities. As a consequence, events having high or small probabilities are worst predicted by people [249, 96].

According to Yu and Kak [256], not all topics are well suited for making predictions starting from social media data. In general, a topic well predictable with social media must meet some requirements. Firstly, the prediction topic must be related to a human event. This means that social media are not suitable to predict events whose development is independent of human actions, such as natural disasters. In more detail, on social media, users publish their feelings, opinions, and beliefs. Prediction methods extract, analyse and match data and finally make predictions according to the impact and influence of people to the observed topic. However, if the topic is a non-human-related event, such as an earthquake, gained data are not related to the development of that event. As a consequence, these social media data cannot be used to predict events independent to human actions, but at the most to understand people's opinions regarding these events.

Secondly, at least partly, the distribution of the composition of involved persons on social media should reflect as to that in the real world [108]. Since it is well-known that not all people in real-world use social media, these data could be treated as samples. Moreover, the sample could present a built-in bias since the sampling process is uncontrollable [256]. In these contexts, it is preferred to take into account proportions of biased samples to make sure they remain in an acceptable and reasonable range. Lastly, one must consider the nature of the topic. Indeed, there are some topics upon which is considered impolite to express opinions having a particular orientation. Therefore, the topic should be easily and freely discussed by people in public. Otherwise, social media contents could be biased [256, 108].

2.8 Sentiment Analysis Challenges

The Sentiment Analysis research field implies several challenges. Some of these have been already discussed in regards of TSA in Section 2.2. Beyond the Twitter Sentiment Analysis, common challenges in Sentiment Analysis are crucial for the effectiveness of methods. This has lead to

the development of works focusing only on challenges that still remain and unexplored questions, such as [157, 109]. For instance, in [109], the author presents two comparison research among forty-seven previous methods in Sentiment Analysis. The first comparison examines relationships between Sentiment Analysis challenges and review structure. The second comparison discusses the importance of solving these challenges to improve accuracy.

Besides works centred on challenges, many research papers highlight encountered issues and possible solutions to avoid them, as in [199] and in [163].

Most research agrees on the identification of some significant issues. A fundamental challenge is related to the correct interpretation of the context in which words are observed. The vast majority of SA tools still have difficulties in precisely evaluating the polarity of statements. For instance, this is particularly clear when one deals with sarcasm or irony. People can infer various useful information to identify ironic or sarcastic statements directly from the context in which these appear.

Moreover, voice intonation together with facial mimicry and body gestures may provide various information. Despite this, capturing sarcasm or irony can be hard also during face-to-face conversations. If also for humans detecting sarcasm and irony in written texts is harder than in face-to-face conversations, the task becomes enormously challenging for automatic tools. The domain-dependence is another essential factor to recognise the sentiment challenges, as suggested in [109].

Challenges are still multiple and may relate to the aim of the task that has to be done as well as specific phenomena. In the following, we first discuss specific phenomena that typically introduce issues: negated expressions - Section 2.8.1, degree adverbs, intensifiers and modals - Section 2.8.2, and figurative expressions - Section 2.8.3. Then, we examine challenges concerning the SA tasks' typology. For the purpose of this thesis, we limit our focus on challenges related to Multilingual Sentiment Analysis - Section 2.8.4.

2.8.1 Negated Expressions

In Section 2.2, we briefly discussed the effects that might be introduced by negations about the Twitter Sentiment Analysis. However, negations represent a crucial point in all Sentiment Analysis related contexts. The negation has been defined by Morante and Sporleder [162] as "a grammatical category that allows the changing of the truth value of a proposition". As an obvious consequence, an analysis of sentiments can be strongly empowered by understanding the impact of negation. Negation that is often expressed through linguistic entities as no, not, and never, can significantly affect the sentiment of its scope [157, 263]. Following the existing literature, we will refer to the negation unit as the negator, while the text that is affected by the negator is the argument.

Since earliest SA works, handling negation appeared a crucial point. The literature shows that

first works typically employ simple heuristics to manage negations' effects. According to Zhu et al. [263], negation models based on heuristics can be divided into two main classes, namely *Non-lexicalized assumptions and modelling*, and *Simple lexicalised assumptions*. The first class can be further divided following two main hypotheses, thus *Reversing* hypothesis and *Shifting* hypothesis.

- **Non-lexicalized assumptions and modelling**. This view states that the impact of negators is independent by the negator itself and by both semantics and syntax of the argument. In these models, which are defined in [263] by Equation 2.16, parameters are only based on the sentiment value of the arguments.

$$s(w_n, \vec{w}) := f(s(\vec{w})) \tag{2.16}$$

  where s is the argument, $(\vec{w})$ is the sentiment score of the argument, and $(w_n, \vec{w})$ is the sentiment score of the entire negated phrases.

  - *Reversing* Theory. The *Reversing* theory consists of changing the degree of the words' sentiment through a fixed constant. Thus, in this view, a negator simply reverses the sentiment score $(\vec{w})$ into $s(\vec{w})$.
  - *Shifting* Theory. The Shifting theory foresees the polarity flip of the words affected by negations. In other words, a basic shifting model depends on $s(\vec{w})$ only, which can be written as Equation 2.17.

$$f(s(\vec{w})) = s(\vec{w})sign(s(\vec{w}))C(2) \tag{2.17}$$

    where $sign(*)$ is the standard *sign* function which determines if a constant C should be added to or deducted from $s(wn)$. The constant is added to a negative $s(\vec{w})$ but deducted from a positive one [263].

- **Simple lexicalised assumptions**. As observed, both previous hypotheses rely on $s(\vec{w})$. However, the effectiveness of non-lexicalised heuristics is limited. Due to this, semantic or syntactic information derived from negators or arguments could be helpful. To this end, researchers have proposed to extend non-lexicalized heuristics by building models dependent by the negator[26], as shown in Equation 2.18.

$$s(w_n, \vec{w}) := f(w_n, s(\vec{w})) \tag{2.18}$$

The empirical study performed by Zhu et al. [263], has been demonstrated that these simple heuristics can be not sufficient in establishing the real impact of negation units on words. They show

[26] For a more comprehensive investigation on Negation-based modelling see [263].

that negators often mutate positive words into negative, and also make negative words less negative, but not positive. Starting from these observations, the authors propose to use an embeddings-based recursive neural network to capture the impact of negators better. Another proposed solution [119] foresee the building of separate words' sentiment lexicons to allow the observation of terms in affirmative and in negated contexts.

To summarise, despite the researchers' effort, several aspects of negation are still not understood. Open issues include the relationship between negators and contexts, the possibility to rank negators depending on their impact on their arguments' sentiment, and the relationship between negators and their usage in different communities and cultures.

2.8.2 Degree Adverbs, Intensifiers, and Modals Verbs

Degree adverbs - such as *just*,*moderately*, and *slightly* - as well as intensifiers - such as *very* and *too* - can impact the sentiment of the sentence's predicate. In more detail, degree adverbs quantify the extent or amount of the predicate, while intensifiers add emotionality to the predicate they modify, without changing the propositional content. Due to some shared characteristics, degree adverbs and intensifiers can be occasionally inter-changed. This led to the impossibility of building comprehensive lists of them also for linguists. Since both affect the sentence's predicate, they also can impact on the entire sentence's sentiment. Due to this, various works have been focused on the interaction between these linguistic units and the predicate. Most of the works focus on identifying sentiment words by bootstrapping over patterns involving degree adverbs and intensifiers [157]. As a consequence, manage these phenomena remain challenging, and several questions are open still now. Open issues include the identification of regular patterns about their impact on sentiments, the possibility to rank them by their impact, and the identification of the contexts where the modifier's impact may change[27].

Modals verbs[28] are used to convey the degree of confidence, possibility, probability and permission to the predicate. As a consequence, if the sentences' predicate is sentiment-bearing, modal verbs may change or vary the whole sentences' sentiment. In other words, the sentiment of a modal verb conjunct with a predicate could diverge from the sentiment of a predicate observed alone. Contrary to the case of the intensifiers and the degree adverbs, the automatic determination of modals' impact on sentiments has been sparsely investigated [157].

[27] For a more comprehensive view of issues related with degree adverbs and intensifiers see [118]
[28] English modal verbs include *can, could, may, might, shall, should, will, would* and *must*.

2.8.3 Figurative Expressions

Literal and figurative language is a dichotomy distinction referred to the fields of language analysis. The literal language uses words exactly according to their meaning. The figurative - or non-literal - language uses words by the common knowledge deviates from their exact definitions. Thus, by definition, figurative expressions are not compositional. Consequently, the meaning of a figurative expression cannot be entirely inferred by individual meanings of its components. Given the difficulties caused by the impossibility to derive the correct meaning of these lexical expressions, there is a growing interest in detecting figurative language. In particular, several researchers focus on irony and sarcasm. The task is such crucial that some competitions directly focus on these themes, such as the Task 11 of the 2015 SemEval competition on Sentiment Analysis of Figurative Language in Twitter[29]. Also in SemEval 2014, the Task 9[30], namely Sentiment Analysis in Twitter, provided an additional test set for sarcastic tweets[31]. Results provided by SemEval participants showed that performances had dropped by about 25 to 70 per cent. These out-comings testify the need to apply dedicated solutions to models to manage sarcastic texts. As underlined by Saif M. Mohammad [157], while sarcasm and irony have seen a growing interest from researchers, the literature lack in works focusing automatic sentiment detection in rhetorical questions and hyperbole, as well as other creative uses of language.

2.8.4 Challenges in Multilingual Sentiment Analysis

The literature regarding Multilingual Sentiment Analysis suggests that the most used approach foresees the mapping of sentiment resources from English into other languages. This method is particularly widespread mainly due to the lack of resources for most languages. In this thesis, we will present some analysis falling in the Multilingual Sentiment Analysis field. Following previous existing works, we address multilanguage contexts by translating English resources into the Italian language. Our motivations and solutions, which will more deeply be discussed following, are driven by the need to align and compare results obtained for two different languages, and thus to start from similar starting points - for instance, statistically comparable seed-lexicons - in both cases of study.

The mapping of English resources has been used to automatically generate a Romanian subjectivity lexicon to classify Romanian texts in [149]. In [237], Chinese customer reviews are translated into English to allow rule-based annotation. In [40], the authors adapted an existing English semantic orientation system to Spanish and also compared several alternative approaches. The authors stated

[29] SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter : `http://alt.qcri.org/semeval2015/task11/`

[30] SemEval- 2014 Task 9: Sentiment Analysis in Twitter: `http://alt.qcri.org/semeval2014/task9/`

[31] The 2014 test dataset combines five test sets: Twitter-2013, SMS-2013, Twitter-2014, *Twitter-sarcasm-2014*, and LiveJournal-2014.

that the Spanish Semantic Orientation calculator obtain worst performances than the English Semantic Orientation calculator also due to both a significant semantic loss and the small preliminary dictionary. In [206], the authors presented a Spanish resource for a multilingual sentiment analysis tool. Moreover, the authors empirically explored the impact of machine translation on sentiment analysis performances. German, Russian and Spanish corpora have been translated to English. The SA performance decrease in the worst case remained within 5%.

Despite the effort spent, some research areas of Multilingual Sentiment Analysis remain less explored. Open issues relate with how sentiment modifiers - i.e., modal verbs and negators - differ across languages, with preserving the degree of sentiment in texts after the translation, and how figurative language can be translated without losing the figurative meaning.

Chapter 3

Human Migration

In this Chapter, we present the basic context of the human migrations study field, and we describe an overview of various means of analysing migration stocks and flows. In particular, we analyse the current state-of-the-art concerning human migration studies starting from primary definitions and notions. Then, the role of Big Data as an alternative data source is discussed, considering its possible advantages and disadvantages.

3.1 Introduction to Human Migration

Humans have been on the move since the beginning of time. Recent history has proven to be one of the periods most affected by the migration phenomenon than at any time before in human history. Motivations and consequences of the migratory phenomenon have been the focus of a wide amount of studies carried out by scholars belonging to various research fields. To date, several definitions have been formalised to capture the migration essence. In general, *human migration* consists of any movement by human beings from a place to another driven by the intention of temporarily or permanently settling in the new location. It typically involves - or can involve - movements over long distances by an individual or large groups. Consequently, a *migrant* can be defined as an individual who permanently or temporarily lives in a country differing from the country of birth.

For the purpose of this thesis, we concentrate on three macro-groups of migrants - *immigrants*, *emigrants*, and *returner migrants* - that will be treated in detail in Section 3.1.1. However, in order to provide a quite comprehensive overview of the complexity of the human migration phenomenon, in following we report some of the possible classification related to both migrants and types of migrations.

Migrants can be defined and classified based on several factors including the motivations that have driven the journey, i.e., *migrant worker* and *refugee*, or destination places, i.e., *international migrant*. The United Nations Convention on the Rights of Migrants has defined a migrant worker

as a "person who is to be engaged, is engaged or has been engaged in a remunerated activity in a State of which he or she is not a national". Despite that, there is no formal legal definition of an international migrant, is commonly accepted that "an international migrant is someone who changes his or her country of usual residence, irrespective of the reason for migration or legal status" [1]. Refugees are defined by the United Nations High Commissioner for Refugees as people "who are outside their country of origin for reasons of feared persecution, conflict, generalised violence, or other circumstances that have seriously disturbed public order and, as a result, require international protection".

A further distinction relates to migration's duration. A short-term or temporary migration typically refers to journeys with a length between three and 12 months. On the contrary, long-term or permanent migration concerns changes in the country of residence of a duration of at least one year. Similarly, the Special Rapporteur of the Commission on Human Rights has established a formal classification for migrants, refugees, and stateless people[2]. Observing these definitions is easy to deduce how difficult is to formally and universally define what constitutes a migrant or a refugee.

Besides the definitions of people who move, also the movements itself can be classified based on their types. Briefly, types of human migration include:

- Internal migration: moving within the same state, country, or continent.
- External migration: moving from to a different state, country, or continent.
- Emigration: leaving one country to move to another.
- Immigration: moving into a new country.
- Return migration: moving back to the origin place.
- Seasonal migration: repetitive moving based on the season or in response to climate conditions or labour.

Migration is driven by factors frequently referred to as *push* and *pull* factors. These factors are in large degree self-explanatory. Push factors are those which forces people to leave their homeland. On the contrary, pull factors are the offers which attract people to move to a country, such as opportunities not available in migrants' homeland. Is interesting to note that push factors were substantially unchanged since humans have begun moving from a place to another over the world. People have been motivated to seek new residences as a result of poverty, drastic climate change, civil war, wars between nations, discrimination, political, gender, ethnic, racial and religious persecution, famine, territorials annexations, and imperial expansions.

[1] United Nations Department of Economic and Social Affairs.
[2] For further information refer to https://www.ohchr.org/en/issues/migration/srmigrants/pages/srmigrantsindex.aspx

As well as push factors, also pull factors are still unchanged. They include the entire set of opportunities, benefits, and improvements which contribute to migrants' well-being and that cannot be obtained in their own homeland.

3.1.1 Immigrants, emigrants and return migrants

Following the related literature, three macro-groups of migrants can be identified, plus the already mentioned refugees.

Emigrants The migration phenomenon starts from out-migration, thus when people leave a country to reach and settle in a new abroad location. Obtaininig information about these people, the *emigrants*, is quite difficult. Reasons are various. First, based on country policies, people are not always obliged to declare their departure. Moreover, also when the departure registration is compulsory, people can leave their country in an illegal manner escaping from the controls. Secondly, after departure information on emigrants are no longer included in origin country registers. Due to these reasons, information about emigrants is typically obtained from the destination countries.

Immigrants As a result of international migration, foreign-origin communities arise in destination countries. Despite citizenship and ethnicity, the country of birth and parents provide the basic factors in determining the foreign-origin or immigrant population. Typically, in the foreign-origin population are included both first and second immigrants generation [83]. The main difference between the two generations consists of the birth country. The first generation relates to people born abroad and then moved to a new destination country of residence. The second generation relates to immigrants' children that are born in the country of residence but having at least one parents born abroad.

Return migrants Migration is strongly characterised by the repetitiveness of events. For instance, surveys have proven that in Estonia people change their residence on average 5 - 6 times during their life and this behaviour tends to increase in younger generations [192]. In the international migration phenomenon context, the repetitiveness of events leads to *return migration*, *consecutive migration* and *circular migration*. Respectively, return migration occurs when the out-migration is followed by returning to the origin country; consecutive migration implies that from the destination country people move to other countries; and, circular migration provides that occur repeated departures from and returns to the origin country.

3.1.2 The migratory process

The migratory process can be analysed through three different main phases, namely the *journey*, the *stay* and the *return*. Respectively, the journey relates to the analysis of migrants' stocks and flows; the stay concerns migrants' integration and the migration impact on both communities and

countries involved; finally, the return refers to the study of both migrants and factors involved in the migrants' return to the own origin homeland.

The journey. Most of the information about migration flows and stocks derive from administrative data and official statistics, thus from surveys and national censuses. The migratory phenomenon intrinsically concerns at least two countries. Due to this, its study implies the use of cross-referenced data coming from among various nations. Data are often inconsistent across databases and shows low spatially coverage and poor time resolution. Given the recent wide availability of Big Data, researchers have begun to employ these new data in understanding and estimating the migration. Thanks to Big Data characteristics it may be possible to develop methods able to analyse high-resolution real-time data and to extract new indexes for estimating, nowcasting, and evaluating stocks and flows.

The stay. Migration has a multitude of important effects both on the migrant population, and on the receiving and source communities, including cultural diversity, economic changes, social interaction. Migration's impact on the host society as well as on the hosted community could result in both long- and short-term changes. The study of the stay phase highly correlates with migrants' integration in the host country. Several indicators can be used to measure immigrants' integration rate including social ties, financial situation, and the labour market. As for the journey, the study of the stay phase is based on official datasets which when available typically have low resolution. Again, the rise in availability of Big Data has led to new opportunities in observing migrants' integration and related consequences through unconventional data with high resolution and large spatial cover. Given that the integration affects several aspects in both the host and hosted society, various unconventional types of data can be used to measure the phenomenon. For instance, call data records can help in tracking the population after disasters occur. Changes in purchase habits and economic status can be analysed through retail data. Social network analysis can allow computing novel integration measures or estimating social evaluation.

The return. Origin communities are one of the main actors involved in the migration phenomenon since they face the effects of migrants' departures. However, source communities are also involved in the opposite phenomenon occurring when migrants return to the homeland.

3.2 The role of Big Data

Data volume and sources have reached exponential growth leading to data generation at 2.5 Exabytes (1 Exabyte = 1.000.000 Terabytes) per day[3]. This high amount of data come from everywhere:

[3]IBM, Big Data and Analytics, 2015. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

mobile phone GPS signals, climate sensors, traffic and flight routes, social media platforms, purchase transactional records, digital pictures and videos (YouTube users upload every minute 72 hours of new video contents), and many others. These data are *Big Data* [76]. "Big Data" refers to the enormous amounts of unstructured data produced by applications falling in wide and heterogeneous application scenarios. Nowadays, Big data represent a novel possibility to provide an alternative to traditional solutions [18]. The term does not rely only on storing or accessing data but also include the set of methods and solutions designed to analyse them. The concept of Big Data is described through the *3V model*. The model, theorised by Laney [71] in 2011, is defined as: "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". The definition was updated by Gartner [25] in 2012: "Big data is high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimisation". Both definitions point out on the three basic features of Big Data: *Volume*, *Variety*, and *Velocity*. However, big data practitioners have again extended the 3V model into a 4V model by including a new "V" for *Value* [97]. Furthermore, the 4V model can be even enriched by the *Veracity* concept, rising to a 5Vs model. Summarising, the set of V-models defines what is (and what is not) Big-Data-related. The five concepts can be described as follows [97, 71, 18]:

- Volume refers to a large amount of data from several data sources, including digital devices. Benefits coming from these large datasets implies several challenges, for instance, obtain reliable knowledge - see Value feature.
- Velocity: refers to the speed of data broadcast. Data are continuously uploaded and enriched by complementary data collections, and different kinds of streamed data from multiple sources. In this context, this feature has driven the development of new methods able to suitable processing and analysing data.
- Variety refers to different types of data collected from several sources, including mobile phones, social networks, and sensors, such as texts, audio, images, and data logs. These data can be unstructured or structured, such as data from relational databases.
- Value: refers to the process of mining reliable and valuable information from large sets of data, and it is commonly referred to as Big Data analytics. The Value feature is the most important characteristic of most of the Big-Data-related application.
- Veracity: refers to the reliability, accuracy, and fairness of information. At the basis of the information management, there are the core doctrines of data quality, metadata management, and data governance, together with privacy and legal concerns.

Recently, analytics over Big Data repositories has received great attention from the research communities. Without a doubt, the scientific computing is the most significant application context

in which Big Data are exploited [62]. Traditional methods, tools, and frameworks have become inefficient in providing effective solutions for managing the data growth, as well as for processing this increasing amount of data. As result, one of the most popular and crucial topic in computing research relies on handling and extracting useful knowledge from these data sources [62, 252].

The journey Given the high number of constraints and issues related to traditional data sources, in the last decade, we have witnessed an increasing Big Data use in Human Migration studies. In this context, works exploiting Big Data[4] have been carried out using mobile phone data.

Currently, phone operators collect very detailed information about human mobility, even if those are limited in spatial scale - national -[210, 116]. Other Big Data examples include e-mail communication logs and Skype data.

Mobile phone data have been employed to describe internal migration in developing countries [32] and to track population whereabouts after disasters occur [21]. Blumenstock et al. [32] show how call record data from Rwanda can be used to compute aggregate levels of migration captured in a standard household survey. Phone data have also been used to measure patterns of mobility. Authors conclude that call record data allow observing individuals migrations, their destination and frequent returns in either origin or destination place.

As well as phone data records, human migration has been studied by using e-mail communication logs. In [240], geographic locations of anonymised users log into Yahoo! services are used to estimate country-to-country flows and migration patterns between countries. The authors develop a protocol to distinguish people that in a one-year period had spent more than three months abroad - defined "migrants" - from users who spent less than a month in a country different from their country of residence - "tourists". To predict migration and tourism flows, phone calls data are crossed with data regarding colonial ties, geographic locations, visas, and economic development. At the global level, results show migration patterns driven by socio-cultural dimensions, such as geography, language, and economy. Together with traditional ones, new routes are also traced. Geolocation data allow the authors to characterise the *pendularity* of migration flows, i.e., the extent to which migrants travel back and forth between their countries of origin and destination. Pendularity levels are spatially dependent and are higher among closer countries. High levels are observed within the European Economic Area. In this case, Big Data is used in conjunction with traditional data in a crossed study including sociology, demography, and geography.

Yahoo! e-mail messages can also be used to estimate human mobility by inferring both age and gender-specific migration rates [258]. Out-comings show increasing mobility and a faster pace of female mobility. Again, the authors conclude outlining the Big Data potential in demographic research, especially for developing countries.

Big Data has attracted researchers attention due to their capability to avoid various limitations

[4]Researches conducted by employing Social Big Data will be discussed separately in Section 3.2.

imposed by traditional data. Together with the easy retrieving and the high resolution, they allow real-time analysis. Besides Social ones, which will be discussed in Section 3.2, data like those gathered from Google Trend Index (GTI)[5], have been recently employed by Böhme et al. [33] to forecast the immigrants' journey. GTIS are cross-validated with data of the Gallup World Poll survey which is focused on peoples' plans of moving. The comparison seems to testify that the GTI data could effectively nowcast trend of migration intention.

However, also Big Data has some restrictions. In particular digital traces resulting from interactions of people and Information and Communications Technology (ICT) infrastructures implies some issues that need to be taken into account [125].

A well-known concern relates to privacy issues linked to the disclosure of personal information online. For migrants, privacy problems can become more critical, especially in cases where they have travelled and entered their destination country illegally. Together with privacy, another main problem related to selection bias is the sample's representativeness. For instance, as is widely known, digitalisation penetration rates vary depending on several factors, including the countries and users' age and gender. As a consequence, some specific groups of users could be difficult to observe. On the other hand, as previously explained, Big Data allows overstepping several traditional data limitations. Their principal benefits related to the low expense and effort needed for retrieving, and their potential real-time collection. Also, while considering potential selection bias, Big Data provides fine-grained worldwide coverage.

Social Big Data *Social Big Data* comes from joining the efforts of two previous domains: social media and Big Data. Social Big Data are based on the analysis of the wide amount of data coming from multiple distributed sources focused on social media. Social Big Data analysis [141, 45] is inherently interdisciplinary and includes various research areas such as Computer Science - for instance, through Natural Language Processing (NLP), Machine Learning (ML), Data Mining, Sentiment Analysis (SA) and Opinion Mining (OP), Information Retrieval (IR), and Network Analysis - Linguistic and Philology, Economy and Psychology, amongst other. As a consequence, their employment covers an almost unlimited number of domains. Data, as well as resources, show a wide range of different characteristics and includes GPS signals, supermarket transactions, and healthcare data.

According to Khan et al. [114], Social Media has become the most relevant and representative data source for Big Data. Thanks to their ubiquity, data can be collected from a practically limitless number of websites and applications, such as Twitter, Facebook, LinkedIn, Instagram, and Flickr. Starting from 2013, the number of tweets each minute has increased by 58%, leading to more than 455,000 tweets per minute in 2017[6]. The era of Big Data is underway [36]. Just like the

[5]https://trends.google.com/trends/
[6]Domo's Data Never Sleeps 5.0 reporthttps://urly.it/3f-v

Web rapidly evolves, users are evolving too. We are in the era of social connectedness, where people interact, share, and collaborate through social networks, online communities, blogs, and other online collaborative platforms. Datasets gathered from social media contain various type of information about users. As compared to traditional data sources, they are potentially unlimited in terms of sample size since they cover a large set of the worldwide population. Online Social Networks (OSNs) can allow researchers to analyse temporal variations easily, and to now-cast migration phenomena.

Besides the advantages, the literature has highlighted also potential issues. Together with privacy and ethics concerns, the selection bias represents a critical point. As underlined by Boyd et al. [36], Twitter does not represent "all people" since Twitter users are a particular sub-set of the entire world population. As a result, the population using Twitter may not be representative of the global population. Moreover, we cannot assume that *accounts* and *users* are equivalent. Beyond the so-called *Twitterbots*[7], users can access social media via the web without ever registering an account, an account can be shared by multiple individuals, while some users subscribe to multiple accounts.

The selection bias has become the focus of various works, in particular, the one concerning Twitter [254, 152]. Results are consistent among different works. For the United Kingdom and the United States, results highlight predominant rates of young adults with higher percentages of males and a higher Twitter's penetrations in densely populated urban areas. Moreover, studies report that the Twitter population is a non-representative and non-random sample of the offline population and of the race and ethnicity distribution.

An interesting study of the Pew Research Center [181] has inspected trends in social media usage over demographic groups. Even if young adults are the most likely to use social media - about 90% -, over-sixties has more than tripled since 2010. Previously attested socio-economic differences seem to decrease. In the past, higher-income households and higher educated people were more likely to use social media. Today 56% of people living in the lowest-income households are using social media. Similar results are obtained observing educational levels. Surprisingly, no differences in gender, racial and ethnic rates are underlined. However, still, there are different levels of social media penetration between rural and urban areas.

Even though selection bias must be considered, various research attested the importance of these unconventional data sources in understanding migration patterns.

Geolocalised Twitter data have been used by Zagheni et al. [257], to analyse trends in mobility and migration flows in OECD countries. The work considers only users that have at least three geolocated tweets for each period of four months. Users' country of residence is estimated for each period as the "modal" country. That is the country where most of the tweets were posted. For each user, if two consecutive periods shares the same modal country, authors assume that the user did not move over eight months. If two consecutive periods show two different modal countries, i.e. C_1

[7] A Twitterbot - sometimes spelt "Twitter bot" - is a software program which produces automated posts without directly involving a person on Twitter.

in the first period and C_2 in the second, they estimate that the user has moved from country C_1 to country C_2 over the eight months considered. Bias was addressed through a difference-in-differences approach. Results they obtain outline an increasing out-migration flows from European countries heavily penalised by the economic crisis, such as Greece, and Ireland. Others European countries, such as Italy, show that out-mobility rates are decreasing, at least in relative terms. A decline in out-migration rates from Mexico to other countries is observed. This last result is consistent with recent estimates of the Pew Research Center for the period 2005-2010[8]. Moreover, their Twitter data show that this trend persists. Official statistics would show this information only with a considerable delay.

Most of the research employing Social Big Data has been cross-validated with official statistics. Results almost always agree on the potential of Twitter as a proxy for human migration and mobility. Furthermore, geolocalised Twitter data allow observing mobility trends and localise flows and stocks of migrants before the official statistics are published.

Measuring Integration: Unconventional Data Sources Nowadays, mobile phones are omnipresent. In most developed countries their coverage reaches 100% of citizens, and also in developing countries, mobile phones are becoming more common. Due to this increasing ubiquity, mobile phones are ever more used as data sources for a wide range of research lines. For instance, they have been used as traffic or mobility sensors [164], as population density estimators [65], detectors of populations reactions to emergencies [10], and as communication hubs, among others.

Most research on social interactions was typically made by using surveys having low coverage - around 1,000 people [31] - and which provide potentially biased data, due to the subjective nature of answers of people interviewed. On the contrary, CDR inherently includes objective data regarding communications between millions of people at a time. A CDR typically includes the call's timestamp, both the caller ID and the called phone number, and the mobile tower used to route the call. These data could be enriched by geographical information also and may be matched with demographic data containing information, such as age and gender. Such personal data combination leads to particularly rich datasets that can be used as information sources for researchers.

While CDR data have been widely exploited for various mobility purposes, CDR-based studies aiming to analyse the international migration phenomenon are few. Even if motivations are various the most crucial relate to cover and privacy issues. Firstly, due to the difference in phone operators among nations, CDR datasets typically cover one single country. Second, due to privacy reasons, CDR datasets do not include information regarding customers' nationality. Thus, cover constraints and lack of fundamental information make the study of international migration quite difficult. However, in the literature there are few exceptions, some of which are represented by challenges, such as

[8] https://urly.it/3f_b

the "Data for Refugees Turkey" (D4R Turkey) challenge[9]. Recently, Rein et al., [3] exploits roaming CDR datasets to differentiate four different travel behaviours. Starting from Vertovec's theory [231], the authors focus on the concept of *transnationalism*. By meaning transnationalism as "operating actively in two or more countries", a transnational person is defined as an individual who generally stays in Estonia but also regularly visits one or several other countries. Travel behaviour parameters and social profiles allow authors to identify transnationals, cross-border commuters, foreign workers and tourists.

Retail data can be used to observe migrants' integration dynamically. Retail data gathered from supermarkets allows inspecting migrants' purchases for understanding if they are gaining habits of the destination country. Immigrants' degree of integration can be estimated by analysing food consumption baskets, and their variation during the immigrants stay in the host country. Moreover, the degree of integration can be estimated by considering economic aspects as well as the timing and how immigrant customers change types of purchases. Analysing retail data over time allows observing the migration phenomenon from both an individual and a collective point of view. Furthermore, these analyses can result in better understanding of migrants' habits and the developing of novel well-being policies.

In [92] and [93], data from a large Italian retail supermarket chain are used to assess the purchases habits distance between immigrants' and natives. The observed dataset is composed of prices, promotional sales, and products purchased by customers. In addition, the dataset includes for each customer, the country of birth and the date in which the supermarket's fidelity card was obtained. Despite the majority of customers are Italians, the dataset also includes a 7% of foreign-born customers. By observing the numbers of customers joining the fidelity club in France, Romania, and Albania, the authors found results in line with immigration trends from European official statistics. Thus, these data could be representative of the migrant population. The authors apply two different approaches: a top-down and a bottom-up approach. The top-down approach allows comparing the amount spent by foreign-born with the amount spent by Italian customers on a specific period. By applying the top-down approach, authors can estimate the immigrants' habits distance with natives' customs. However, identifying products reflecting the distance with natives' norms is not an easy task. Conversely, the bottom-up approach starts from migrants' baskets composition for developing an indicator of shopping distance concerning the natives' typical baskets. Applying the algorithm developed in [94], firstly, the authors identify a representative basket for each customer. Then, by re-clustering representative baskets for countries, they build representative national collective baskets. Finally, they apply a set-based distance measure and develop the indicator of shopping divergence/convergence.

[9] "Data for Refugees Turkey" is a Big Data challenge, whereby Türk Telekom opens a large dataset of anonymised mobile phone usage to researcher groups for providing better living conditions to Syrian refugees in Turkey. In this case, customers' status, i.e., to be a refugee is made available.

In [169], is presented a work based on Twitter data analyses trends in global human migration. The authors describe human migration as a three-factor model. On the basis of origin country, destination country, and time of migration movements patterns belonging to different groups are inferred by temporal windows ranging from 1 to 5 months. For instance, observing a one-month window, the authors found patterns related to tourists. Three-months windows relate to Erasmus students, while five-months windows relate to long-term migration.

As can be observed, most of the recent literature is based on Twitter data. However, some works have exploited other social network platforms, such as Google+ [147] and Facebook [101, 74]. Messias et al. have presented the first Google+ data based-study on migration clusters, i.e. groups of countries in which individuals have lived sequentially. The authors have developed a model to predict the prevalence of a certain triad. Countries more likely to be clustered are identified by using a set of features such as a shared language or colonial ties. The authors underline that an analysis such this cannot be accomplished by using surveys data. Surveys do not provide the information they have used, and, even when are collected they regard only to small regions of a country.

Anonymised data from Facebook's advertising platform have been used by Dubois et al. [74] to evaluate the cultural assimilation of Arabic-speaking migrants in Germany. The authors base migrants' groups comparisons on an assimilation score. The score is built on the interests' similarity between profiles belonging to different migrants' groups. These groups are modelled on the basis of the immigrants' origin country. Results show that the score varies among subgroups populated by younger and more educated men. Recently, also Zagheni et al. [259] focus on the use of Facebook's advertising platform data. The work aims to monitor stocks of migrants in the United States. The authors assume that besides migration studies, the proposed approach may be applied to demographic indicators too. Moreover, they underline the importance of Facebook's advertising platform as a sampling tool to find and recruit specific populations with targeted ads.

Language in OSNs Migration flows and stocks have been extensively studied starting from a wide range of available data ranging from labour market data to the number of people living outside their origin or residency country [126]. Another possible type of data allows to study the migratory phenomena is represented by the language the users' use on OSNs. Indeed, the language could represent a direct link between individuals, origin country, and nationality. Moreover, the language allows observing the way in which linguistic characteristics of people varies when the contact with other cultures occur.

The language allows us to express needs, feelings and achieve our communication goals. When over the time, the society changes and grows more complex, the language must evolve and adapt itself to the new needs of its population. As a consequence, this evolution leads to changes, creations, and vanishings of expressions, dialects and even whole languages [89]. Over the past two decades, globalization has driven social, cultural and linguistic changes panorama in societies all over the world. Thanks to the influence of pioneering works of linguistic anthropologists, mixing, mobility

patterns and historical framework became key issues in the study of the languages and of the language groups [29]. Over time, linguists and sociologists analysed variation and changes in both oral [124] and written [16] language by exploiting surveys, corpora, and records [89]. In the last decade, the pervasive use of online social networking and micro-blogging services led to the availability of freely-made contents never seen before. This unprecedented wealth of written data allows us to obtain a view of language evolution both from geographical and the time point of view [168].

The literature regarding the language in social networks applied to migration studies is wide and involves several research fields, including but not limited to mobility patterns, migrations stocks and flows, Well-Being and Sentiment Analysis. Even though some works focused more on metadata instead of the real data contents, the text bears a wealth of information, starting from the language in which is written [126]. The language is the most characteristic trait of human communication, but it can assume various heterogeneous forms [89]. In particular, dialects can be defined as linguistic varieties differing lexically, phonologically, and grammatically in regions geographically separated [52]. Despite its crucial importance and the high effort spent, several aspects about the manner in which the language spatially varies are still unknown. Detecting and analysing variation in language is the core of research fields such as socio-variational linguistics and Dialectology. Nevertheless, since contents in online social networks are generated from users located all over the world, the language on social platforms testify geographic variations [122]. Identifying these linguistic variations is a challenging task due to the various forms the variation can assume, such as lexical, syntactic and semantic. The language has been also investigated in the spatial distribution as well as the spatial extension of dialects. In this context, most of the works [13, 73, 77, 78] focus on the lexical variation detection in certain geographic regions. For instance, Ibrahim et al. [110] have combined different datasets - Hotel reservation, TV program comments, tweets, and product reviews - to present a semi-supervised sentiment analysis approach for standard Arabic and Egyptian dialectal Arabic. The authors also build a support vector machine (SVM) classifier based on both linguistical and syntactical features.

However, works such as the one carried out by Kulkarni et al. [122] are not limited to specific regions. Kulkarni et al. presented GEODIST, a computational approach allowing to detect English linguistic variation and quantify its significance among geographic regions. The method tracks and detects significant linguistic shifts in word usage across geographical regions. GEODIST has been used to investigate the linguistic variation of Twitter data (a) between four English speaking countries and (b) between fifty states in the USA. The model is able to identify several changes including region-specific usages, and regional dialectical variations. Finally, the method has been applied to analyse distances between language dialects. Results of British and American English over a period of 100 years reveals that the semantic variation between dialects is decreasing probably due to cultural mixing and globalisation.

Although several studies have investigated the language dynamics on OSNs [15], the most of

these works focus on specific aspects of the combined study of language and geographical analysis in Twitter, thus a global picture is often neglected [154]. For instance, language-dependent differences have been explored from Twitter's users activity on the basis of posting and conversations patterns [241]; language, network properties and sentiments have been analysed for the top ten active countries on Twitter [184]; the impact of language, boundaries, geographic distance, and air travel frequency has been observed in the formation of social ties on Twitter [217]; and Doyle [73] have proposed a Bayesian method to build conditional probability distributions of the spatial extension of English dialects.

Despite the proved effort, many works are still limited in space or languages. A recurrent motivation is due to the kind of needed resources. For example, in Automated Text Categorisation, example sets are used to predict categories of unlabelled documents. In Sentiment Analysis, lexicon-based methods exploit labelled lexicons to assign polarised scores to words and texts. Building these resources is inherently language-dependent and requires remarkable human efforts, high costs and time. On consequence, efforts primarily focus on the most attested languages and active countries on social platforms, leaving aside the others. Only a few parts of the studies provide comprehensive worldwide-scales studies. For instance, Mocanu et al. [154] have characterised the worldwide linguistic geography finding a universal pattern describing users' activity across countries and a high correlation between the Gross Domestic Product (GDP) and the Twitter adoption. By aggregating OSNs data at different scales, the authors show the high heterogeneity of Twitter penetration and its relevant correlation with GDP. Moreover, they find that the statistical usage pattern of the social platform is independent of such factors such as country and language. Results of temporal variations of the language composition for countries can lead in observing travelling patterns and identifying in real time seasonal travelling and mobility patterns.

In [139], geolocated tweets have been exploited to analyse spatial-language interaction on Twitter. The study aims at identifying localised patterns in language among different countries. The authors mainly focus on relations between tweets language and their spatial distribution to investigate *a*) language diversity and its usage in Twitter communities geolocalised by country; and *b*) cultural groups' spatial distribution into different countries. This study allows identifying either dominant languages and the spatial distribution of minor languages in local communities. Secondly, it leads to evaluate the language dominance in both local communities and in the global Twitter community, giving information on language usage in Twitter data. Third, the authors show if the Twitter's community sample can be representative of the actual population by comparing languages' diversity measures with official data. Finally, they are able to localise different cultural groups by analysing the spatial distribution of languages inside countries. The latter scope could be useful to understand certain situation related to specific cultural groups such as Syrian refugees preferring to get closer to similar culture communities. For each local community identified, several statistical measures are computed to show language diversity inside the community. Findings show that statistical

measures taking into account the language distribution within the community are more coherent and robust in identifying the highest language diversity countries. Results also show that in 65% of cases, local Twitter communities are characterised by the relative local language. Thus, the English language cannot be considered a general language proxy independently from tweets' spatial distribution. Finally, to assess the Twitter population representativeness, tweets are grouped by language at different spatial-based scales. Results show that tweets in the dataset are posted in more than 55 distinct languages, as well as in different dialects for the same language, from 206 countries over the world.

In [89], Gonçalves et al. performed a large-scale analysis of language diatopic variation using geolocalised Twitter datasets. The authors focused on the Spanish language, building a corpus from which a list of concepts has been extracted to characterize Spanish varieties on a global scale. By using a cluster analysis, sets of macro-regions sharing common lexical properties have been defined. In more detail, the authors found that the Spanish language seems to be split into two super-dialects, namely an urban speech typically used in the major Spanish and American cities, and a suburban speech mostly attested in small towns and rural areas. The latter can be further clustered into smaller dialect sets regionally characterised.

Due to this lack of worldwide-scale coverage studies, the Cross-Language Text Categorization [17] (CLTC) is becoming a key research field. CLTC aims to exploit labeled samples of a source language to learn a classifier for a different target language. This approach aims to turn away the language-dependent issue by reducing the human effort. Over the time, several approaches have been proposed. Parallel corpora methods [75, 182, 253] are traditionally based on the bag-of-words model, which exploits lexical resources where synonyms in different languages share the same vectorial representation. Machine Translation (MT) methods [238] reduce all documents to a single language. Anyway, translating a huge number of documents involves high cost either in economical and time terms. Structural Correspondence Learning (SCL) [28] is applied to Cross-Language Text Categorization (CL-SCL) [191] to avoid the needs of MT tools. Indeed, the approach is based on a word-translator that allows building a set of word pairs, the so-called dubbed pivots. Pivots represent the ground knowledge to discover analogies between the source and target languages. Anyway, CL-SCL methods involve high computational costs imposed by optimisations and by the inclusion of Latent Semantic Analysis (LSA). Finally, in [79] is proposed a Distributional Correspondence Indexing (DCI) method based on feature profiles according to the features' distributional correspondence. This approach allows comparing documents in different languages since these have been indexed in a common vector space. Opposed to SCL, MT and parallel corpora approach, DCI method needs lower cost both in human effort and in computational.

Alongside the lack of works providing a global perspective, most of the studies focus on the identification and characterisation of linguistic islands or of certain communities and on the differences between them. More interestingly, some works focus on analysing the integration of immigrants or

on the patterns of language mobility.

An extensive study of immigrant integration using Twitter language detection is proposed in [126] by Lamanna et al. The authors introduce metrics of spatial integration to quantify the *Power of Integration of cities* - i.e. their capacity to integrate diverse cultures - and characterise the relations between different cultures. First, immigrants are characterised using their digital spatio-temporal communication patterns, defining the most probable native language and their residence place. In more detail, for each city, the authors define a spatial grid composed of square cells and define the home location of a user to be the grid cell where most of his or hers activity occurs between 8 pm and 8 am. The language is assigned to users based on those detected in their Twitter posts, for instance, a user tweeting in a different language from the local one or in English is assigned to the corresponding minority community. Then, using a modified entropy metric as a quantitative measure of the spatial integration of each community in cities, the authors carried out a spatial distribution analysis. To quantify the spatial integration of immigrant communities in cities, the authors build a Bipartite Spatial Integration Network H, where a language l is connected to the city c, where the corresponding immigrant community is detected. The degree of integration is computed as the edge weight $h_{l,c}$ based on a modified version of the Shannon entropy-like descriptors [243]. Similarities between cities' manner of integration are outlined by a clustering analysis based on the distribution of edge weight $h_{l,c}$ for each city c. The authors conclude by stating that despite the well-known bias, Twitter it is able to identify spatial patterns and mobility profiles.

In [161] Moise et al. have explored the use of Big Data analytics for tracking language mobility in geolocalised Twitter data. The work is split into two steps. In the first step, the authors explore the temporal dynamics of languages in three specific case studies. Results obtained by performing linear regressions show that Twitter mainly reflects language distribution in multi-linguistic countries such as Switzerland. Interestingly, some results show an example of the Twitter's bias. The most attested language in the dataset is English. However, English is followed by Portuguese, Spanish and French. The rank does not reflect the official estimates of language speakers in the world where Chinese leads the ranking, followed by English, Spanish, and Hindi. This disagreement cannot be ignored and serves as a reminder. The worldwide and the Twitter landscape are not equivalent, for instance, Twitter is not available in China. During the second step, the authors move from a temporal perspective to a spatial one. The spread of languages is observed through monthly snapshots resulting from a density-based clustering technique. Then, either the spatial and the temporal analysis inferences are used to investigate how language mobility over time is reflected in Twitter data. A neural network model is applied to emulate the mobility of a language through the time-lapse of the centers of the language mass. Finally, authors show that Twitter is a promising data source for discovering tourism trends and flows by exploring shifts in language composition.

Besides the "mere" identification of attitudes - positive, negative or neutral - toward certain topics, sentiment analysis techniques can also be applied to the human migration strand.

Geolocalised Twitter data have been used to generate sentiment maps [23, 153]. Dynamics of sentiments in New York have been analysed by Bertrand et al., while Mitchell et al. [153] have studied happiness dynamics at both urban and states level in the United States. The work aims (a) to measure the overall happiness of people and (b) to understand happiness variation of different cities. To accomplish the first task, the authors have applied the technique introduced by Dodds and Danforth [69] and improved in Dodds et al. [70]. Regarding the second aim, the authors inspect how the word usage correlates with happiness and both social and economic features. Similar approaches could be used to evaluate happiness in specific urban areas, i.e. districts densely populated by immigrants, to observe changes over time of people happiness before and after migrants' settle as well as to measure migrants integration.

Citizens' sentiments toward government policies have been analysed in [7]. By exploiting specific concepts and keywords, the authors have found that citizens' negative sentiments typically concerns the scarce agency communication on social media, backlogs in claims processing, and low awareness of services provided. This works highlights the possibility to use SA techniques to derive guidelines for the government.

The perception of refugees has been analysed by Coletto et al. [60]. The authors have employed the framework presented in [59] to investigate Twitter users' perceptions of Brexit and refugee crisis in Europe. As in our case, also this work is based on a cross analysis along multiple dimensions, such as space, time, and sentiment. Beyond the findings, this study testifies the wide range of study opportunities provided by the cross-analysis of different dimensions. Moreover, the study confirms how polarisation on one topic may help in understanding attitudes on other related topics.

The Return: migrants returning to the country of origin Migration is commonly seen as a permanent change in residence habits. However, when considered as a temporary phenomenon, several implications arise. We refer to the phenomenon of return migration, which is increasing in several countries. In human history, some migration routes have remained constant and densely traversed for long terms.

Among most-known waves of migration, population flows moving between Mexico and the US are probably the most significant and also famous. Newest trends in Mexican migration flows' direction seem to reverse the previous ones. Today, Mexicans returning from the United States are more than those migrating in the US [42].

As already underlined, migrations' drivers can be various. These also include economic aspirations which are led, i.e. by the migrants' aim to found a job, to enhance their economic status, to improve own education and skills, and to support own family economically. The results obtained by Bucheli et al. [42] regarding the Mexican case agree with almost other all contributions focusing on the valuation of benefits and drawbacks lead by the return of migrants. The authors show that when Mexicans return in their homeland tend to become part of the economically active population.

Besides the Mexican case, nowadays, return migration represents an increasing and ever more

current phenomenon in various countries, i.e., China [261], Jamaica [222], Tunisia [146], and Mali [55].

The increasing of return migrants has led to a rise in the developing of contributions which analyse either motivations driving the return and the impact of the return itself on both hosting and homeland communities. The most recent literature almost totally agrees in underlining the benefits led by returning migrants. These advantages concern a very wide range of fields and include the rise of business activity, and the wages increase [235, 236], the improvement of educational attainment and health conditions, the increase of electoral participation [55], and the decrease of violence [42].

As well as for the journey and the stay phases, also the study of the return of migrants has been widely investigated by using traditional data - i.e., household and migration surveys data [261] - but also Big Data [222] - i.e., money transactions gathered from banks.

In the return migration-related literature, most of the research has been focused on the "brain gain" provided by the return of high-skilled individuals, i.e., scientists returning in the country of birth. Scholars found that even if migration leads to a brain drain over the short-term, return migration can contribute to brain gain [68, 236]. Moreover, the most recent works demonstrate that return migrants contribute to the own community's long-term well-being independently of skills they have gained abroad [42].

Chapter 4

Introduction to the Music Context

Musics origins, like those of language, are hidden in the most ancient past of humanitys history. It is a crucial part of human civilisation for centuries, becoming a universal language. Every culture has given birth to its own music "style". However, as time goes by, the constant collapse of physical barriers, as well as the progressive reduction of geographical distances eased by the media and the World Wide Web, caused the overall globalisation of music. During the last decade, we have witnessed a constant growth of online streaming services that allow a broad open, "democracy" and omnipresent access to the widest choice of music ever seen. Emerging and niche artists just like famous ones can quickly gain global visibility. In this quickly evolving panorama, music threatens to lose its geographical-cultural characterisation. This multifaceted scenario has drawn the attention of the researchers. Indeed several works inspect music from a wide range of perspectives. The current music scene offers multiple research suggestions, i.e., geographical and cultural connotations, data collection from online services, specific music features inference, users behaviours, and tastes. Recently, there is high attention to Music Sentiment Analysis. This specific domain exploited several techniques and involved different kinds of features.

In the last decade, particular attention turned to corpus-based methods. Despite the fact that creating songs polarity tagged datasets is not an easy task [47], datasets of songs labelled with emotions, and polarity tagged lexicons are essentials prerequisite to compute those classification models. As suggested in [47], a music dataset should observe four main characteristics, such as (a) strong polarisation; (b) easily understandable labels taxonomy; (c) high coverage and large size (at least 1,000 lyrics); and (d) publicly available. For instance, All Music Guide (AMG)[1] [72] has invested both from the economic and from the human points of view considerable resources to annotate high-quality emotional music databases. Consequently, they are unlikely to share their data publicly. Besides, it is also quite impossible to manually tag large datasets. Indeed, freely

[1] All Music Guide website: `https://www.allmusic.com/`

available manually annotated datasets are generally small in size. In [223], the authors created and publicly shared a dataset of 593 songs all of which have been annotated employing six emotions by three experts. In [225] Turnbull et al. collected CAL500, a dataset of songs annotated by at least three annotators. CAL500 is composed of one song for 500 artists.

The literature underlines different solutions to avoid these problems. A first approach to collect emotion annotations is a survey. Surveys and therefore the crowdsourcing platforms, such as Amazon Mechanical Turk, represent a straightforward way to gather this kind of musical contents. A second approach to collect intelligence involves social tagging and online services, such as Spotify, Apple Music, Last.fm. In particular, Last.fm[2], is a music discovery website with a wide heterogeneous community of music listeners. Users can contribute to social unstructured text tags [127]. As opposed to AMG, some music platforms like Last.fm, Genius, AllMusic, and Spotify have made its data through public APIs. While Last.fm, as well as others like Spotify, represent a useful resource for researchers, [49] underlined several problems with social tags, such as their sparsity, the fake tagging, popularity bias, and lexical tags variations. Despite this, Last.fm has been broadly used [106, 128] in music sentiment analysis tasks, even though obtained datasets are not made public. Indeed, as already underlined in [47], as far as may be difficult to believe, still today no lyrics sentiment dataset fulfils all the four conditions mentioned before. In this context, the Italian music scenario is even more dramatic.

4.1 Music Analysis: A State-of-the-Art

From the beginning of the 21st century, the music scene is facing ever-increasing growth of attention from the scientific community, empowered by the permeation of the World Wide Web and the music-dedicated platforms into daily life. The research on the Italian music domain is sparse to nonexistent. To the best of our knowledge, the unique contribution focused on this specific domain is proposed in [180]. The authors present the Rapscape corpus, a POS-tagged and lemmatised lexical resource containing about 16,000 Italian rap songs grouped by artist. As declared by authors, the building of this resource aims to overcome the lack of resources about Italian rap music. The initial dataset is gathered by exploiting both Discogs[3] and Spotify APIs. Unfortunately, the article, the relating results, and in particular, the resource is not attested in the music analysis literature and is not publicly available or accessible. Indeed, we are aware of the work only thanks to the personal collaboration into the preprocessing phase of data used in the contribution.

Regarding the worldwide music analysis panorama, the literature on music analysis is noticeably large. Several works [26, 105, 107, 106, 185, 195, 198] have analysed data gathered by online services in order to analyse different phenomena related to the online music consumption, such as

[2]Last.fm website: `https://www.last.fm/`
[3]Discogs website: `https://www.discogs.com/`

model diffusion of new music genres/artists, behaviours and tastes of users, recommendation models, classification of semantic states inferred by lyrics, music classification, and so on. In particular, the music sentiment analysis, also called music mood recognition, exploits and combines several techniques, such as machine learning, and data mining to classify songs into polarity classes. The literature displays that several different kinds of features like melodic/audio, lyrics, or metadata can be involved. In [105], AllMusic metadata are used to create a categorical representation of music emotions. Their results show that many individual mood terms are highly synonymous or express different aspects of a more general mood class. This leads in some cases, to better identification of the underlying mood by decreasing mood vocabulary size. Authors also propose a five-class music categorisation and a set of not even thirty mood's popular terms, recommending to reduce mood lexicons in a set of classes rather than using immoderate individual mood terms. One of the two most attested trends in music emotion recognition is to use self-created datasets. In [106] Last.fm tags are used to build a 5,000 songs dataset tagged with 18 mood categories. Authors employ a binary approach for all the mood categories, independently if songs have or not category tags.

As underlined above, a second trend is gathering intelligence about music by human feedback employing surveys, in particular exploiting Amazon Mechanical Turk. In [129] the authors proposed an inquiry of the possibility to apply this service to music mood ground truth data creation. By comparing Mechanical Turk data with those of MIREX AMC 2007 task6, authors report a similar distribution on the MIREX clusters. Even though authors warn of problems which can diminish annotations qualities, such as spamming, they conclude that Mechanical Turk represents a valid approach option. In [140] a lyrics dataset based on Valence-Arousal model of Russell [200] is created employing (AMG) tags. Likewise [69, 187] and the work we present, ANEW is used as a lexical resource. Once classified AMG tags in the four Russells model quadrants using ANEW, songs are categorised using the obtained tags, and finally, annotations are evaluated by employing human evaluators.

As underlined in [47], this tagged dataset is one of the few public lyrics datasets. Finally, another attested tendency both in music sentiment analysis and in the analysis of the phenomena related to the music is to build multi-modal music datasets [150, 204] which merge and combine several kinds of features. In [133] a semi-supervised approach is used to study the problem of the music artist genre identification both from lyrics and melodic features, as acoustic ones. The similarity between the 45 analysed artists is identified by exploring AMG artists pages. Obtained results report that lyrics and sound performances are comparable. Authors of [204] presents Musiclef, a professionally created multi-modal dataset of about 1300 popular songs. Musiclef combines several features such as general metadata, Last.fm tags, audio features together with web pages and labels provided by expert annotators. In Musiclef songs are first labelled using a seed set of 188 terms, after reduced to 94. Nevertheless, this wide range of labels may seem redundant, superfluous and not reliable [47]. The approach proposed in [150] merges either textual and melodic features. The work exploits a 100

items dataset of popular songs annotated for emotions at line level using Amazon Mechanical Turk. The dataset is then used to explore the automatic recognition of emotions in songs. The results demonstrate that (a) emotion recognition can be performed using either melodic or textual features, and especially that (b) the joint usage of these two dimensions leads to improving significantly classifications based on only one dimension at a time. The obtained dataset is available for research upon request to the authors, but due to its small size cannot be used as experimentation set [47]. On the contrary, a rich set of lyrics like the one presented in [46] lacks in the human evaluation, and therefore it too cannot be used as a ground truth set.

Part II

Estimating Superdiversity through Twitter Sentiment Analysis

Chapter 5

Data, Lexical Resources and Preprocessing

In this chapter, we describe the datasets, the lexical resources, and the preprocessing phases we adopted to develop the sentiment lexicon-based model, which allows us to obtain the "used" emotional valence of words in the population of a region - see Chapter 6, and to apply it to estimate Superdiversity - see Chapter 7. The works share, at least in part, data, resources, and pre-processing steps. To avoid multiple repetitions and redundancies, we chose to insert all these information over the following sections. Then, in the relevant Chapters, we will just recall these descriptions outlining only eventual differences or additions. A similar procedure will be adopted for works referring to the music context - Chapters 9 and 10 - in Chapter 8.

5.1 Datasets

In this Section, we describe the datasets we employ to develop our sentiment epidemic algorithm. We exploit two untagged Twitter datasets which have been used to both build the lemmas' network and extend the initial seed dictionary. More precisely, we used the so-called *Untagged Twitter Dataset* to develop and evaluate the sentiment lexicon-based model. Then, to estimate the Superdiversity in the United Kingdom and Italy, we exploit a more extensive version of the dataset mentioned above, namely the *Untagged Extended Twitter Dataset*. Finally, a third dataset, namely *Tagged Twitter dataset*, has been used to evaluate the performance of our extended dictionary and for a Sentiment Analysis generic task.

5.1.1 Twitter Datasets

We employ different Twitter datasets - one tagged with sentiment classes, two untagged.

Untagged Twitter dataset. The *Untagged Twitter Dataset* we exploit is a subset of the one used in [61]. The dataset was originally collected by using the Twitter Streaming API under the Gardenhose agreement which grants access to 10% of all tweets. Our subset consists of just under one million and seven hundred geolocalised English Twitter messages retrieved for 6 days, from the 14th to the 20th of August 2015. All tweets were linguistically treated following the method described in Section 5.1.2. The dataset is used in two different ways. First, it is at the base of our method for extending the dictionary of terms tagged with sentiment valences. Second, a sentiment analysis evaluation-task is performed on it, and the resulting sentiment scores compared between our extended dictionary and a previously established dictionary from the literature.

Untagged Extended Twitter dataset. The *Untagged Extended Twitter Dataset* is a subset of the dataset used in [61]. The entire dataset was collected by using the Twitter Streaming API under the Gardenhose agreement, which grants access to 10% of all tweets. The subset we exploit is composed of just under 73,175,500 geolocalised Twitter messages gathered for three months, from the 1st August to the 31st October of 2015. This dataset is used to estimate the Superdiversity, thus the distance between the "standard" emotional valence of a set of words and the "used" valence in the population of a region, see Chapter 7. All selected tweets were linguistically treated following the method described in Section 5.1.2.

Tagged Twitter Dataset. The *Tagged Twitter Dataset* is built by merging three different Twitter corpora. At first, it is composed of 3,734 publicly available tweets and their sentiment classifications, retrieved from the following sources:

- The *Semeval 2013 Message Polarity Classification competition (task B)*[1]. The original dataset consisted of a 12-20K messages corpus on a range of topics, classified into positive, neutral and negative classes. We retrieved 2,752 such tweets that were still available on the Twitter platform. These were passed through the language detection algorithm provided by the Python package Langdetect, to ensure they were written in English, leaving us with 2,745 tweets in the final dataset - 428 negatives, 1,347 neutral and 970 positives.

- The *Semeval 2014 Message Polarity Classification competition (task B)*[2]. Similar to Semeval 2013, this corpus consisted originally in 10000 tweets, out of which we downloaded 687 English tweets (142 negative, 319 neutral and 226 positives).

- *Earth Hour 2015 corpus*[3]. This dataset contains 600 tweets annotated with Sentiment information - positive, negative, neutral - where each annotation is triply-annotated through a

[1]The Semeval 2013 Message Polarity Classification competition (task B), https://www.cs.york.ac.uk/semeval-2013/
[2]The Semeval 2014 Message Polarity Classification competition (task B), http://alt.qcri.org/semeval2014/
[3]The Earth Hour 2015 corpus: https://gate.ac.uk/projects/decarbonet/datasets.html

crowdsourcing campaign. Out of these, 295 tweets were still available for download through the Twitter platform - 30 negatives, 185 neutral and 80 positives.

We observe that a small proportion of tweets is labelled with a negative valence, with neutral tweets being most prevalent, in all three sub-datasets. This can be seen as a characteristic of Twitter messages in general. The Tagged Twitter Dataset is composed of 3,727 tweets, and it is used to evaluate the performance of our final extended dictionary. For this, we first proceeded to preprocess data. Then, the entire dataset is normalised and lemmatised through a general-purpose pipeline of linguistic annotation tools.

5.1.2 D4I Dataset

The *Data Challenge on Integration of Migrants in Cities* (D4I) dataset is a high-resolution immigration rates data for a data challenge[4]. Provided data are aggregated from the 2011 censuses from some selected European countries. More precisely, it contains the concentration of migrants in all cities of eight EU countries: Spain, Germany, Italy, France, Netherlands, Portugal, United Kingdom, and Ireland. Migrants are counted based on three different levels of aggregation: by country, continent and EU versus non-EU. For our analysis, we focused only on data for the UK and Italy and summing EU and non-EU immigrant counts to obtain total immigration levels.

5.2 Data Preprocessing

The aim of the preprocessing phase of tweets is their grammatical annotation. The first problem we decided to deal with was to obtain clean data which can be processed effectively by automatic methods. The annotation of linguistic constituents required the development of a dedicated rule-based cleaning procedure. This consisted of removing punctuation, links, and usernames, and normalising hashtags and emphasis words.

The output of this cleaning phase is linguistically standardised tweets that are subsequently treated by a general-purpose pipeline of linguistic annotation tools. Specifically, tweets are lemmatised and tagged with the Part-Of-Speech tagger TreeTagger, described in [205]. Once obtained POS-tags, we address the problem related to the number of noisy words. These words uselessly increase the data to be processed and may be wrongly identified. To this end, for each tweet, we selected only words belonging to specific grammatical classes: nouns, adjectives, and verbs. This allowed us to obtain only significant words from the sentiment and meaning point of view. We applied the preprocessing procedure to the Tagged and both the Untagged tweets described above. Hence the datasets contain cleared standardised texts composed of only nouns, verbs, and adjectives.

[4]Data Challenge on Integration of Migrants in Cities (D4I), `https://bluehub.jrc.ec.europa.eu/datachallenge/`

Algorithm 1 Location Matching Algorithm

procedure LocationMatching($UETD, D4I$)
  $A = []$ ▷ array of tweets with not matched origins in the D4I dataset
  $B = []$ ▷ array of tweets with matched origins in the D4I dataset
  for $t \in UETD$ **do**
    if $t_{origin} \in D4I(cities)$ **then** ▷ searches correspondences among cities into D4I
      if $t_{country} = country(D4I(t_{origin}))$ **then** ▷ checks country consistency
        $B((t)) \leftarrow [NUTS3(D4I(t_{origin})), NUTS2(D4I(t_{origin})), NUTS1(D4I(t_{origin}))]$
      else
        A **add** t
    else
      A **add** t
  return A, B

5.2.1 Geo-referencing Tweets & Language Selection

The first problem we faced during the preprocessing phase is the tweets' georeferencing. The task was partially accomplished by using the *metadata* of the tweets themselves. In other words, we initially check if information into the `place` field - thus `'country'` and `'name'` fields[5] - can be matched with a country and a city into the D4I dataset.

Let t be a tweet in the Untagged Extended Twitter Dataset $UETD$ (the same applies to the Untagged Twitter Dataset). Let be the tweet characterised by a `place` metadata field which includes the name of the origin city t_{origin}, and the country where the city is located $t_{country}$. Each location in the D4I dataset $D4I$ is characterised by its name *city*, its county, its region and country, and the three NUTS code associated with them, thus $NUTS3$, $NUTS2$, and $NUTS1$, respectively. The Algorithm 1 checks if the tweet's origin corresponds to a location in the D4I dataset. If the information matches, the algorithm assigns to the tweet the three NUTS levels. At the end of the process, the algorithm returns two arrays. The first is composed of tweets with not matched origins in the D4I dataset, thus A. The second array B is composed of tweets to which it has assigned the NUTS codes.

However, metadata allowed us to match only parts of tweets' origins with locations in the D4I dataset. To avoid loss of information, we perform a further specific procedure to geo-allocate all tweets, which is described in the following.

Once faced with the issue related to the tweets' *origin*, we focus on the *language selection* problem. Again, we exploit the tweets' metadata, and in particular the field `lang`.

[5]For the tweets structure refer to Figure 2.1

Algorithm 2 City-Tweet Allocation Algorithm

```
1: procedure LocationSearch(A, B, t)                          ▷ arrays returned by Alg. 1
2:     for t ∈ L do
3:         LocationFound ← False                               ▷ boolean control variable
4:         api_response ← MediaWiki(t, LocationFound)          ▷ calls MediaWiki API - Alg. 4
5:         if api_response[0] = False then
6:             api_response ← GoogleSearch(t, n, LocationFound)   ▷ calls GoogleSearch API - Alg 5
7:             if api_response[0] = False then
8:                 LocationFound ← api_response[0]
9:         if LocationFound = False then
10:            api_response ← GoogleMaps(t, LocationFound)     ▷ calls GoogleMaps API - Alg. 6
11:            if api_response[0] = False then
12:                api_response ← Geopy(t, LocationFound)      ▷ calls Geopy API - Alg. 7
13:        if LocationFound = True then
14:            B(t) ← api_response[1]                          ▷ update the B array
```

5.2.2 City-Tweet Allocation

We perform a further processing stage, which has as objective the allocation of tweets from each dataset to a city in the UK or Italy. This was required to match the D4I data to our results. Each tweet has associated a location, including the city or town it came from. Thus, for each tweet, we check if its origin is attested in D4I. If the city is found, we assign to the tweet the related NUTS[6] code. Otherwise, we perform a dedicated rule-based pipeline. This step is often required since in D4I locations are at the city level, while several tweet origins are at the district or town level.

The process applied to assign to each tweet the origin city is described by Algorithm 2. At the end of the process, all the retrieved cities are matched with those in the D4I dataset, and each tweet is labelled with NUTS codes at three levels, namely NUTS1, NUTS2, and NUTS3. Following this procedure, we are able to associate a NUTS code with over 94% of the UK tweets and to over 97.5% of the Italian tweets.

The City-Tweet Allocation Algorithm, first exploits the MediaWiki API[7] by using the tweets location as a key-word, as shown in Algorithm 4. The API calls, in turn, the Wikipedia pages' parser - Algorithm 3 - to automatically extract information.

When the MediaWiki API allows us to extract the city referred to the tweets origin from the Wikipedia pages info-box, we assign it to the tweet. If not, the location value is used as a Google Search API[8] parameter to extract URLs of the first five pages - Algorithm 5. From these, URLs

[6]The Classification of Territorial Units for Statistics (NUTS), in French "*Nomenclature des unités territoriales statistiques*", is a standard geocode referring the subdivisions of countries for statistical purposes. The standard is developed and regulated by the European Union.

[7]MediaWiki API main page: `https://www.mediawiki.org/wiki/API:Main_page`

[8]Google Custom Search API page: `https://developers.google.com/custom-search/`

Algorithm 3 Wikipedia Parser Function

function WIKIPEDIAPARSER($page, LocationFound, t$)
  $geo\_data = []$ ▷ empty array
  if $page \exists$ & $page \neq redirect\ page$ **then**
    PARSE $page$ ▷ parse MediaWiki page
    $infobox \leftarrow page$.GET($infobox$) ▷ gets Wikipedia info-box
    $LocationFound \leftarrow True$
    if $t_{lang} =$ "eng" **then**
      $geo\_data \leftarrow infobox$.GET($county, region, country$) ▷ fields' names changes according to the page's lang.
    else
      $geo\_data \leftarrow infobox$.GET($provincia, regione, stato$)
  return ($LocationFound$, geo_data)

Algorithm 4 MediaWiki API Function

function MEDIAWIKI($t, LocationFound$)
  $wiki\_url =$ MW_API.SITE(t_{lang}, "$wikipedia$") ▷ MediaWiki (MW) API call parameters
  $page \leftarrow$ MW_API($wiki\_url, t_{origin}$) ▷ call to MediaWiki (MW) API
  return WIKIPEDIAPARSER($page, LocationFound, t$) ▷ calls Wikipedia Parser function - Alg. 3

pointing to Wikipedia are selected and, as before, are used to extract the city referred to by the location from the info-box. If the combination of the already mentioned APIs does not allow us to retrieve a city attested in the D4I dataset, we exploit the Google Maps API[9] first and the Geopy[10] python libraries then - Algorithm 6 and Algorithm 7. These two are used after the first searching phase due to their rate and call limits, even though they are generally more accurate.

It is, however, worth pointing out that in both Twitter datasets there are several tweets geo-labelled with the region/county or country name. Also, in particular in the Italian dataset, many tweets are labelled with multi-language city names translations, such as "Nápoles", "Trentino-Alto Adigio", "Venecia", "São Gimian" and "São Remo", "Ancône", "Naturns" and "Florencia". Tweets belonging to the first case are discarded due to the multilevel geographical nature of our analysis. Country or region/county tags do not allow us to reach a fine-grained geographical level. Instead, tweets belonging to the second case are not ruled out *a priori*, but our pipeline rarely assigns them a NUTS code.

[9]Google Maps Platform: `https://cloud.google.com/maps-platform/`
[10]Geopy client for geocoding web services: `https://pypi.org/project/geopy/`

Algorithm 5 GoogleSearch API Function

function GOOGLESEARCH($t, n, LocationFound$) ▷ n = number of pages to get (five)
  $query\_string \leftarrow t_{origin} + t_{country}$ ▷ adds the country for a better search
  if $t_{lang} =$ "it" **then** ▷ checks the tweet's lang. to use the proper Wikipedia URL
    $wiki\_url \leftarrow$ "$https://it.wikipedia.org/wiki/$"
  else
    $wiki\_url \leftarrow$ "$https://en.wikipedia.org/wiki/$"
  $pages \leftarrow$ GS_API($query\_string, n$) ▷ call to GoogleSearch (GS) API
  for $page \in pages$ **do**
    if $wiki\_url \in page$ **then** ▷ looks for URLs pointing to Wikipedia
      return WIKIPEDIAPARSER($page, LocationFound, t$) ▷ calls Wikipedia Parser function - Alg. 3

Algorithm 6 GoogleMaps API Function

function GOOGLEMAPS($t, LocationFound$)
  $geo\_data = []$ ▷ empty array
  $query\_string = t_{origin} + t_{country}$
  $geo\_code \leftarrow$ GM_API.GET(GEOCODE($query\_string$)) ▷ calls GoogleMaps (GM) API
  if $geo\_code\,\exists$ & $geo\_code_{country} = t_{country}$ **then** ▷ checks country correctness
    $LocationFound \leftarrow True$
    $geo\_data \leftarrow geo\_code$.GET($county, region, country$)
  return ($LocationFound$, geo_data)

5.3 Lexical Resources

To develop our algorithm and validate its performances, we employed several resources of lemmas annotated with sentiment valences.

- ANEW. The Affective Norms for English Words lexicon (ANEW) is an established emotive lexicon proposed by Bradley and Lang [37]. The lexicon provides normative polarity ratings for 1,034 English words including verbs, nouns, and adjectives. Emotive ratings refer to three psychological reactions to a given word. Thus, ANEW is characterised along three dimensions: *valence*, as the level of pleasantness, *dominance*, as the degree of control, and *arousal*, as the intensity of emotion, as well as by word frequency. Scores range between 0 and 10. In the literature, the most used dimension is the valence, which ranges in the scale from pleasant to unpleasant.

- SENTIWORDNET. In SentiWordNet [81], words are associated with emotional values according to how objective, positive, and negative they are. The resource is based on the well-known database. By exploiting WordNet's glosses, SentiWordNet use synsets as semantic representations of the synsets themselves and classifies them into three polarity categories, such as *Pos(s)*, *Neg(s)* and *Obj(s)*.

Algorithm 7 Geopy API Function

```
1: function GEOPY(t, LocationFound)
2:     geo_data = []
3:     geo_code ← GP_API.GET(GEOCODE(t_origin))          ▷ calls Geopy (GP) API
4:     if geo_code ∃ & geo_code_country = t_coutry then   ▷ checks country correctness
5:         LocationFound ← True
6:         geo_data ← geo_code.GET(county, region, country)
7:     return (LocationFound, geo_data)
```

- FULL LIST OF BAD WORDS BANNED BY GOOGLE: Due to the wide usage of informal and vulgar language in tweets, we built in-house bad words lexicon. Terms have been retrieved from the *Full List of Bad Words Banned by Google* of the "What Do You Love" (WDYL) Google project[11]. Once obtained all the 550 swear and curse terms, we manually enriched each of them with a valence score. Since these terms are considered strongly negative, all of them are tagged with a 0.0 valence score.

- PAISA. The PAISÀ [138] corpus is a vast collection of about 380,000 Italian texts taken from the web, for a total of approximately two hundred and fifty million tokens. Together with lemmas, the corpus also provides their frequencies.

5.4 Lexical Resources Preprocessing

Words in ANEW are tagged over a continuous scale ranging from 1 to 10. On the contrary, lemmas in SW are labelled throught three categories, namely *Pos(s)*, *Neg(s)* and *Obj(s)*.

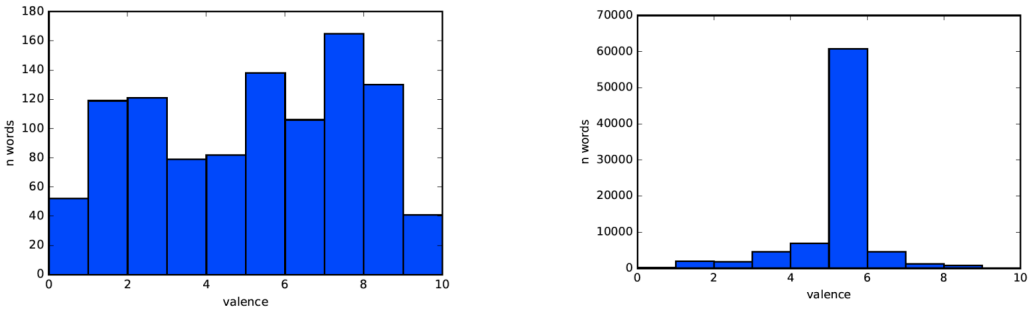
To be able to use valences derived by SentiWordNet together with ones from ANEW lexicon, we computed a unique polarity for each word in SW. Derive prior polarities from SentiWordNet is a well-know task in Sentiment Analysis literature. Several approaches can been applied[12]. Amongst the formulae proposed in [90], we adopt the difference between the positive and the negative score as an overall sentiment valence, properly scaled to interval $[0, 10]$ like ANEW.

Given a lemma-PoS with n senses `lemma#PoS#n`, every formula f is independently applied to all the `Pos(s)` and `Neg(s)`. This results in two scores, $f(posScore)$ and $f(negScore)$, for each lemma-PoS. A unique prior polarity for each lemma-PoS, $f(posScore)$ and $f(negScore)$, is derived as

$$f_d = f(posScore)f(negScore) \tag{5.1}$$

where f_d computes the difference between them. According to [167, 2, 91], we chose the most basic form of prior polarities. Thus we only consider the first (and thus most frequent) sense is considered

[11] The service is now inactive. The list can be downloaded from Free Web Header: https://urly.it/31fxt.
[12] An exhaustive description of available formulae is proposed in [90].



(a) ANEW (b) SW

Figure 5.1: Comparison between ANEW (a) and SentiWordNet (b) distributions of words' polarities.

for the given `lemma#pos`. This is equivalent to considering only the score for `lemma#PoS#1`.

Figure 5.1[13] compares polarities of words in ANEW - Figure 5.1a - with those in SW - Figure 5.1b. After calculating prior polarities for words in SentiWordNet, we observed that the distribution of valences is strongly heterogeneous. As shown in Figure 5.1b, neutral words are much more than positives and negatives. Therefore, we decided to balance classes. We start by identifying the least represented class, thus the negative class. Then, choosing the total number n of items in the negative class, we select from other classes the n most strongly polarised lemmas, for a total of 16,914 lemmas.

5.4.1 Conclusions

In this chapter, we presented datasets, resources, and the preprocessing phases we adopted to perform our work. The preprocessing phase we presented is composed of several steps which aim to geo-reference the tweets. The first step exploits tweets' metadata to match their locations with those into the D4I dataset. Since metadata allowed us to match locations only partially, we performed a dedicated rule-based pipeline. The pipeline exploits several API, such as MediaWiki API, Google Search API, Google Maps API, and the Geopy python library. These services are called in order one after the other until the tweet's location matches with one in the D4I dataset.

The phase described before allowed us to obtain all English tweets posted from the UK and all Italian tweets posted by Italy from the Untagged Extended Twitter Dataset. These data represent the starting point to identify Superdiversity in UK and Italy. As described in the following chapters, these tweets are then split depending on the NUTS layers. Each tweet subset is used to build a network of words, that is required to obtain the sentiment dictionary through the sentiment spreading process - see Chapter 6. Finally, the Pearson correlation between the standard and the

[13]Scales of bar plots are different to allow a proper visualisation since the datasets are very different in numbers of words they include.

actual valences of words in each dictionary obtained allows us to compute the Superdiversity Index - see Chapter 7.

Chapter 6

Sentiment Spreading

The challenges posed by the growing amount of Big Data availability has open novel and challenging scenario into the Sentiment Analysis research field. As a consequence, the most recent Sentiment Analysis literature is related to microblogging contents. The growing attention towards both Sentiment Analysis and social media platforms have led to novel research fields, such as the Twitter Sentiment Analysis. In particular, this field has attracted most of the researchers' attention. This is because texts to be analysed consist of very short messages - only 280 characters - generally lacking in structure and semantic context. The nature of these novel data imposes fast and effective methods to handle and explore the sentiment in these data efficiently. Lexicon-based methods, which use a predefined dictionary of terms tagged with sentiment valences to evaluate sentiment in longer sentences, can be a valid approach. However, the Twitter Sentiment Analysis, as well as standard Sentiment Analysis, still often remain a segregate research field.

This chapter focuses on the development of the lexicon-based epidemic model for Twitter Sentiment Analysis, that is the algorithm that provides us the data needed to identify Superdiversity. As mentioned before, to assess Superdiversity, we created the *Superdiversity Index* (SI), which estimates the distance between the "standard" emotional valences of words and the "used" ones by a target population. While the standard valences of words are derived from an established lexicon, the "used" valences need to be computed. To this purpose, we built the lexicon-based epidemic model for Twitter Sentiment Analysis. The model is a data-driven algorithm which automatically extends an initial seed dictionary and assigns an emotional valence to new terms following an epidemic based approach. The output of the algorithm consists of a domain-dependent sentiment dictionary. Moreover, since valences of words depend on the way language is used, the dictionary is population-dependent.

This chapter describes the development of the lexicon-based epidemic model for Twitter Sentiment Analysis. Thus, in the following sections, we first explain our algorithm, and then we describe

the steps it computes. Finally, we describe the algorithm's evaluation in Chapter 6.1.3, while in Chapter 6.1.5 we discuss about the results we obtained.

6.1 Lexicon-based Epidemic Model for Twitter Sentiment Analysis

Sentiment analysis has been an important research challenge in the last years, especially with the availability of large amounts of user-generated content from various microblogs. Although several methods have been introduced [176], issues still exist both when it comes to applying and evaluating new methods. In particular, for microblogging data such as Twitter, difficulties arise due to the length and structure of the messages. These are limited to 280 characters providing little semantic context. Moreover, they do not always abide by grammatical rules, which means preprocessing is not straightforward.

Additionally, the size of the data to be processed is very large, so methods need to be powerful and fast. As seen in Chapter 2.4.2, several approaches to TSA are lexicon-based. They use a dictionary of words that are either manually or automatically tagged with a sentiment valence to assign sentiment to tweets. These methods have the advantage of being easily translatable to other languages, i.e., by translating the small dictionary and they are fast enough to process large amounts of data in a short time, which is an important feature when it comes to Big Data processing. However, the main disadvantage of this approach is that, since tweets are short, no term may be found in the dictionary, reducing the number of tweets that can be classified.

The main idea behind our work is that different cultures assign different emotional valence to different words. By computing an index which we call *Superdiversity Index*, we can estimate the distance between the "standard" emotional valence of a set of words and the "used" valence in the population of a region. In particular, to derive the "standard" emotional valence, we exploit a manually tagged lexicon, namely ANEW[37]. The "used" valence is derived from a Twitter data corpus. To estimate sentiment valences of words on Twitter, we develop a data-driven algorithm which automatically extends an initial seed dictionary and assigns valence to new terms using an epidemic based approach - opinion dynamics like. The output of the model consists of an extended sentiment dictionary that can be used for tweets' classification tasks. Interestingly, this allows characterising groups where the tweets came from.

The sentiment algorithm extends an initial sentiment-tagged seed lexicon using a Twitter corpus. The base-concept is to built a model to enhance lexicon-based sentiment analysis on Twitter. Often, when processing Twitter data, the dictionaries are too restricted, and small amounts of tweets can be classified. To tackle this issue means to extend the labelled dictionary have been investigated. For instance, Velikovich et al. [229] build a network of words and phrases from a collection of Web documents. For each word, they first build a vector of co-occurrence with the other words. Then

they construct a word graph in which each word is a node, and edges are weighted with the cosine similarity over the co-occurrence vectors. Sentiment scores are propagated from seed words into other nodes. The algorithm computes both a positive and negative polarity for each node in the graph (pol_i^+ and pol_i^-). Positive and negative scores are equal to the sum over the maximum weighted path from every seed word to a specific node. The final polarity score for the word is calculated as $pol_i = pol_i^+ - \beta pol_i^-$, where β is a constant meant to account for the difference in overall mass of positive and negative flow in the graph. Compared to such approach, our method simplifies significantly both the graph building stage, and the sentiment propagation stage. In particular, we use the unweighted co-occurrence network and simple propagation based on ideas from contagion models. These simplifications are important to be able to process larger amounts of data.

Our sentiment algorithm uses the tweets to be classified to create a network of terms, which are then tagged with the sentiment using an epidemic-like process. In [211], the authors use label propagation within the Twitter follower graph to improve the polarity classification. Moreover, in [104], the authors propose a sociological approach to show that social theories such as Emotional Contagion and Sentiment Consistency could be useful for sentiment analysis. We also propose the use of ideas from contagion models. In our case, epidemic spreading and opinion dynamics are used to extend the dictionary used for sentiment analysis automatically. However, we do not base our spreading process on the follower network, but on an unweighted co-occurrence graph.

The process can be summarised in the following three main steps:

1. The epidemic model starts from a small seed lexicon with sentiment valences. It builds a network where nodes are terms from the Twitter corpus. An unweighted and undirected edge connects terms that co-occur in a tweet. Once the network is built, seed-nodes are tagged with the valence score derived from ANEW.

2. Sentiment valences diffuse from seed-nodes to the other words. Thus, nodes without valence take the valence of its neighbours.

3. After many iterations, the system converges to a stable valence which is assigned to terms.

The resulting extended dictionary:

- consists of large amounts of polarity-tagged terms. As a consequence, it allows enhanced sentiment analysis on Twitter.

- is domain-dependent and also depends on the way language is used. In particular, the new valences are population-dependent.

- can be used to compute the distance to a manually tagged lexicon. Moreover, the new valences are estimates for the real emotional content of the words in the population.

The key parts of the methodology are described in the following subsections.

6.1.1 Data, Preprocessing and Annotation

In this subsection, we describe the lexical resources we used to assign valences to nodes into the network. To create the expanded dictionary, we require a seed lexicon labelled according to mutually exclusive positive, negative, and neutral sentiment classes. Moreover, we exploit two datasets. The first is an *Untagged Twitter Dataset* which is employed to build the network of lemmas and to extend the initial seed dictionary. Thus, it is used to develop the algorithm. The second dataset, namely *Tagged Twitter Dataset*, will be used in the following to evaluate the performance of our extended dictionary, compared to an already established dictionary, in classifying tweet sentiment. Details regarding these two datasets are provided in Section 5.1.1.

To develop our algorithm and validate results we identified several resources of lemmas annotated with sentiment valences. The resources we exploit has been already presented in Section 2.6 and in Section 5.3. More precisely, we refer to

1. ANEW [37]: we select the *pleasure* dimension as a sentiment valence, as evaluated by both male and female subjects.
2. SentiWordNet [81]: we compute unique polarities for each word.
3. *Full List of Bad Words Banned by Google* of the "What Do You Love" (WDYL) Google project[1]: we manually enrich words with a 0.0 valence score.

The preprocessing phase we apply to the datasets mentioned above consists of three main steps, thus a cleaning phase, a pre-processing phase, and a phase concerning the noise reduction. Details regarding the entire pre-processing process are provided in Section 5.2.

6.1.2 Extending the Dictionary

We adopt an epidemics-based approach to extend the dictionary of terms used for lexicon-based Twitter Sentiment Analysis. Our method consists of two stages: building the network and sentiment spreading. In the following two subsections we separately describe these two stages.

Building the Network

To build the undirected and unweighted network, we start from the preprocessed Untagged Twitter Dataset. In the network, nodes are represented by lemmas found in the preprocessed Untagged Twitter Dataset. Two nodes are connected by an edge if there is at least one tweet where both lemmas appear. Hence, the network is an unweighted co-occurrence graph based on the target tweets to be classified. A large number of tweets are used to build this network. To obtain the

[1]The service is now inactive. The list can be downloaded from Free Web Header: https://urly.it/31fxt.

most possible meaningful and robust network, we chose to consider only tweets composed by more than three lemmas. We expect that lemmas with positive valence will be mostly connected to other positive lemmas, while those with negative valences will be connected among themselves. As previously explained in Section 2.8.1, handling negations is a challenging task. Our issue with negation mainly relies on the difficulty to understand which lemmas from the negated tweet can be considered connected in the network, and which not. Because of that, we chose to consider only tweets that are not containing a negation, i.e., "dont", "not".

Let t a tweet in the Untagged Twitter Dataset (UTD), and i a negation in a manually computed set of negations N. The tweet t will be discarded in our analysis if its length $len(t)$ is below a fixed threshold - we chose a tweets' minimum length of 3 - or if it contains a negation - Algorithm 8.

Algorithm 8 Negation and length control function

```
1:  procedure TweetCheck(UTD, N)
2:      minLength = 3                          ▷ tweets' length fixed threshold
3:      discard = False                        ▷ control variable
4:      for t ∈ UTD do
5:          if len(t) < minLength then
6:              discard ← True
7:          else
8:              while N(index(i)) < size(N) & discard ≠ True do
9:                  if i ∈ t then
10:                     discard ← True
11:                 i ← N(index(i)) + +
12:         if discard = True then
13:             delete UTD(index(t))
14:     return UTD
```

Sentiment Spreading

Once the network of lemmas is obtained, we start to add valences to each node in the network. We start from a seed dictionary, which is typically reduced in size. This seed allows us to assign valences to a reduced number of nodes in the network. This is the initial state of our epidemic process.

In the following section, we will show results obtained when the seed is 50% of the ANEW dictionary - while the other half is used to validate the results -, together with all lemmas in the SentiWordNet and Bad words lexicon.

Starting from the initial state, we follow a discrete time process. At each step sentiment valences spread through the network. At time t, for all nodes v_i which do not have any valence, the set of neighbouring nodes $N(v_i)$ is analysed, and v_i takes the valence that aggregates the distribution of valences in $N(v_i)$, through a function $F(N(v_i))$. Let V be the set of all nodes in the network N and S the set of initial seed nodes with their valences $val_s(v_i)$. The aim is to build a set V^* of nodes in V with valences $val(v_i)$ assigned. The process is similar to those seen in continuous opinion dynamics

Algorithm 9 Sentiment Spreading Algorithm

```
1: procedure SentimentSpreading(V, S, N)
2:     V* = ∅
3:     for v_i ∈ S do                                  ▷ Initialise valences with seed
4:         val(v_i) ← val_s(v_i)
5:         V* ← V* ∪ {v_i}
6:     repeat
7:         V*_old ← V*
8:         for v_i ∈ V − v* do
9:             v = F(N(v_i))          ▷ F(N(v_i)) aggregates the distribution of valences in N(v_i)
10:            if v ≠ NULL then
11:                val(v_i) ← v
12:                V* ← V* ∪ {v_i}
13:    until V* = V*_old
14:    return V*
```

models [209], where agents take into account the aggregated opinion of their entire neighbourhood when forming their own. The difference here is that once a valence is assigned, it is never modified. The procedure is defined in Algorithm 9.

To decide the aggregation procedure $F(N(v_i))$ we took into account several observations. In general, tweets appear to be very heterogeneous, most containing both positive and negative words. Hence a simple averaging of valences would most of the time result in neutral lemmas, although they contain meaningful sentiment. So, we decided to use instead the *mode* of the distribution of valences in the neighbourhood, which is a much more meaningful criterion in these conditions. However, the mode was only considered in special circumstances. We observed that in some cases the distribution of valences in the neighbourhood is very heterogeneous. That means the range of valences is very large, or the entropy of the distribution is very high. In this case, it is unclear what the valence of the new lemma should be, so we chose not to assign one at all. Again, this procedure was inspired by works from opinion dynamics, i.e., the *q-voter model* [48], taking into account the concept of social impact: agents are better able to influence their neighbours as a consensual group rather than isolated. Hence a heterogeneous group will not influence its neighbours. Here, this was implemented as thresholds on the range and entropy on the neighbouring valence distribution. Thus, a node will be infected with the aggregated valence of its neighbourhood only if the range and entropy are below these thresholds. Let $range^*$ and $entropy^*$ be the two thresholds. Then the procedure F taking as input $N(v_i)$, the neighbours of v_i, can be described by Algorithm 10.

To avoid outliers, we consider the range to be the difference between the 10th and the 90th percentile. Following this line, the two thresholds - range and entropy - become two parameters of our model, that need to be tuned to maximise performance.

Algorithm 10 Infection Function

```
1: procedure F(N(v_i), entropy*, range*)
2:     e = entropy(val(N(v_i)))          ▷ entropy of valences of nodes in N(v_i)
3:     r = range(val(N(v_i)))            ▷ range of valences of nodes in N(v_i)
4:     if e < entropy* & r < range* then
5:         return mode(val(N(V_i)))
6:     return NULL
```

6.1.3 Evaluation

In this subsection, we describe the criteria used to evaluate the effectiveness of the epidemic lexicon-based algorithm as well as the extended dictionary.

Two different analysis have been performed to evaluate the obtained extended dictionary:

1. valences' cross-validation with an established dictionary in the literature;
2. classification performance cross-validation with an established dictionary in the literature.

Valences' cross-validation

We first concentrate on the valences obtained accomplishing our sentiment spreading process. For this, we use cross-validation on the ANEW dictionary. Specifically, the ANEW dictionary is divided into two halves. One half is used as a seed dictionary during the epidemic process, together with SentiWordNet and the Bad words dictionary. The second half is used as a test dataset. This means that the valences obtained through our procedure are compared to the original valence in the ANEW dictionary. Furthermore, we want to obtain an indication of whether our process produces valid sentiment valences. To this end, to quantify the similarity, we use the Pearson correlation. Given a pair of random variables (X, Y), the Pearson correlation ρ is

$$\rho XY = \frac{cov(XY)}{\sigma_X \sigma_Y} \tag{6.1}$$

It is important to note that while a significant correlation between the modelled and real valences is desired, we do not expect to obtain very large such values. This is because the correlation highly depends on the corpus, i.e., how Twitter users use language, the topic. We expect that by changing the Twitter population, the correlation changes as well. We propose the correlation to be a means to quantify superdiversity on Twitter, and we believe this can be very useful in understanding the effects of population migration both on receiving an incoming population.

To extract optimal values for the entropy and range thresholds, we repeated the spreading procedure ten times for various thresholds. Figure 6.1 shows the average correlation obtained for each threshold combination. As can be observed, the range parameter is more important in obtaining

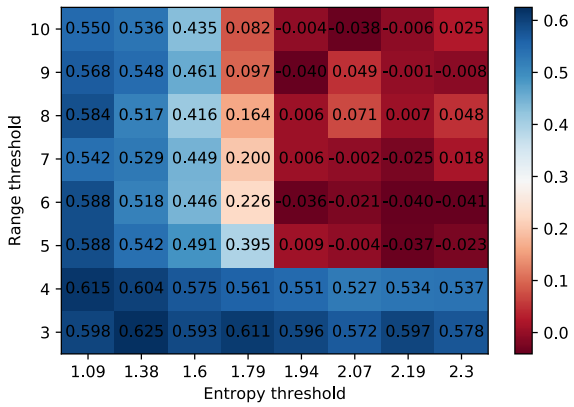


Figure 6.1: Average correlation between modelled and real word valence.

higher correlations, with small ranges resulting in better results. The optimal performance is obtained for a range threshold of 3, and the entropy threshold of 1.38. Note that we consider the distribution described by 10 bins of equal size. Hence the maximum entropy is approximately 2.3. Figure 6.2 displays the modelled and real valences on test data for the run with the best correlation with these parameter values. The plot shows clearly that the valences obtained by our method align well with human-tagged data, validating our approach.

Moreover, we are interested in evaluating the match between valences of an established dictionary, i.e., ANEW, and ones obtained with our epidemic lexicon-based model. To this end, the distribution of valences in the ANEW dictionary is compared to the extended dictionary. Results are shown in Figure 6.3. As can be seen, negative lemmas result in the smaller fraction of the dictionary in both cases. However, the comparison shows some differences which mainly relate to neutral and positive lemmas. The extended dictionary still contains a more significant fraction of neutral and positive lemmas, compared to the ANEW dictionary. This is, however, not a concern since we already observed the same trend for other dictionaries as well, i.e., for instance in SentiWordNet as described in Chapter 5.

Cross-validation Classification Performance

The second criterion for validation of the extended dictionary is classification performance on the Tagged Twitter Dataset - Section 5.1.1. Because of that, we implement a sentiment classifier based on Support Vector Machines (SVM), that used several features to classify sentiment of tweets into three classes: negative, neutral and positive. The features used include, for each tweet, several statistics over the valence of individual lemmas contained by the tweet. Specifically, selected features comprise both arithmetic and geometric mean, median, standard deviation, and minimum and maximum valence. To these, we added the number of lemmas with a valence over 7 and over 9. This feature can be seen as an indicator of the presence of strongly positive terms. Conversely, we also computed

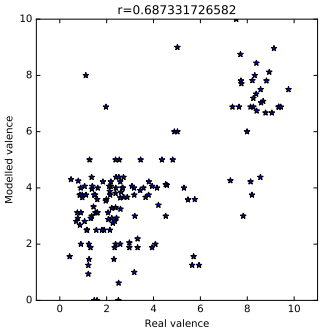


Figure 6.2: Modelled and real word valence for a selected run with best parameters.

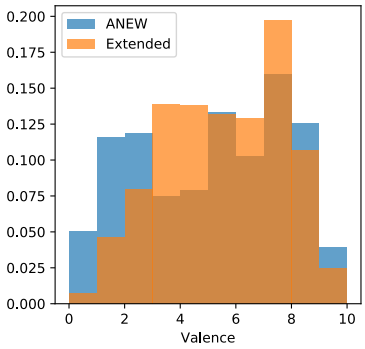


Figure 6.3: Histogram of valences for ANEW and extended dictionary.

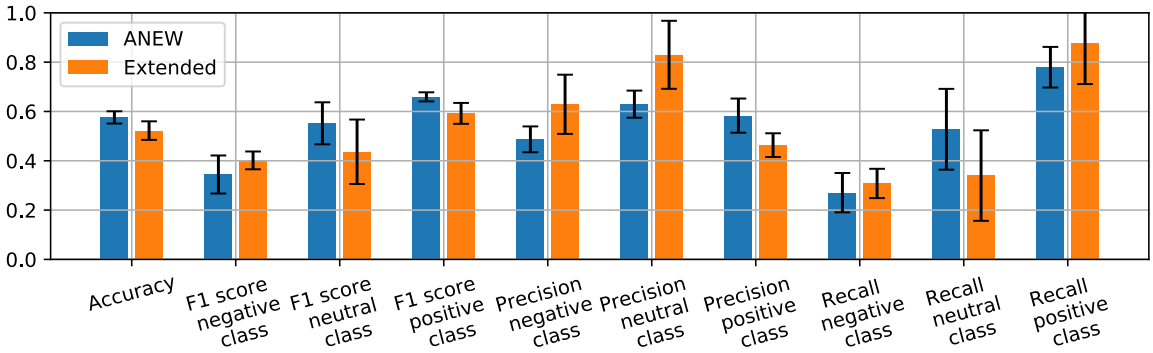


Figure 6.4: Performance of SVM classifier using the original ANEW dictionary only and our extended dictionary.

the number of lemmas with a valence under 3 and under 1, as an indicator of strongly negative terms. Finally, we include the total length of the tweet, and a boolean feature flagging the presence of negation. To obtain effective possible classification, we only considered tweets for which at least three lemmas were found in the dictionary.

The features above can be computed using any dictionary, and SVM performance can vary when changing the dictionary. We compare the performance of our extended dictionary - described in Figures 6.2 and 6.3 - with that of ANEW, which is an established dictionary in the literature. We expect no decrease in performance with our extended dictionary. To validate results, we used a cross-validation approach, where 80% of tagged tweets were used to train the SVM and 20% to test it. The analysis was repeated ten times for each dictionary, with average performance displayed in Figure 6.4. Following the literature, we evaluate performances through the standard four indices described in Section 2.2.3, thus *accuracy, precision, recall,* and *F-score*. Error bars show one standard deviation from the mean. The plot shows that the performance with the extended dictionary is comparable to ANEW, validating our approach again. The F-score increases the negative class and decreases

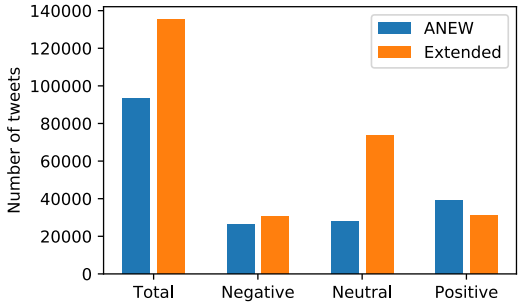


Figure 6.5: Applying the SVM to Untagged Twitter data using the original ANEW dictionary versus our extended dictionary.

Dict.	Positive	Neutral	Negative	Total
`ANEW`	48,409	20,222	24,816	93,447
`Extended`	31,278	73,519	30,544	135,341

Table 6.1: Tweets classification performance with ANEW and our extended dictionary.

on the other two. Precision increases on negative and neutral, while recall increases on negative and positive tweets. Hence, our extended dictionary seems to perform slightly better on negative tweets. However, it overestimates the positives. Given that negative tweets are a small part of the corpus, accuracy decreases slightly. The performance range across repeated runs always overlaps between the two dictionaries compared.

6.1.4 Twitter Sentiment Analysis with Epidemic Model

To further evaluate the goodness of our algorithm, we compare the behaviour of the SVM in classifying the Untagged Twitter Dataset using our epidemic extended sentiment dictionary versus ANEW. Figure 6.5 shows the number of tweets classified by each dictionary, and then the distribution in the negative, neutral and positive classes. The SVM used was trained with all the Tagged Twitter Dataset, while the extended dictionary used was that analysed in Figures 6.2 and 6.3. Applying the SVM using ANEW were are able to label 93,447 tweets. As can be seen from Table 6.1, the number of positive tweets is almost twice that of the neutral and negative tweets. Results obtained applying the SVM classifier with the extended dictionary show an increase in the number of labelled tweets compared to ANEW, by about 45%. We are able to classify 135,341 tweets. As opposed to ANEW, the number of positives - Table 6.1 - and negatives tweets is not that unbalanced. We also see a significant increase in tweets classified as neutral.

Since the dataset has no polarity labels, we cannot validate the classes obtained by each dictionary for individual tweets. However, the distribution of classes obtained with our extended dictionary seems to be closer to what we observed in the Tagged Twitter Dataset, i.e., neutral tweets are a

majority, and negative ones are a minority. ANEW tends to find mostly positive tweets, and this could be because the number of positive words in the dictionary is higher than in other classes. However, this is true also for the extended dictionary. A second reason could be that Twitter users use a youth slang: a large set of unstandardised lemmas which are not included in ANEW lexicon, but which are captured by our extended dictionary.

6.1.5 Discussion and Conclusions

In this section, we have proposed a method for enhancing lexicon-based sentiment analysis by extending the base lexicon of terms. The algorithm produces a sentiment labelled dictionary as output. The obtained dictionary was shown to contain term valences that correlate well with human-labelled lexicons. The performance of the SVM-based sentiment classification was maintained. We believe that our method is particularly suitable for Twitter data, as well as for Big Data in general. This is also because the procedure is very fast. Running times ranges in the order of minutes for over 1.5 million tweets. This characteristic suggests that the method can be applied to vast amounts of data. The results we obtain validated our method.

During our experiments, we have observed an increase in the number of tweets tagged by about 45% compared to ANEW. Furthermore, the extended dictionary is much more extensive than ANEW. This indicates the fact that some terms in the network remain isolated so that sentiment valences do not percolate the entire network, hence many tweets remain untagged. This issue is exacerbated by the fact that we also impose thresholds on the neighbourhood from which a node can be infected with a sentiment. In these conditions, additional data can improve the percolation power of sentiment valences and further increase the performance of our method.

Chapter 7

Superdiversity & Superdiversity Index

Migration has a multitude of important effects both on the migrant population, and on the receiving and source communities, including cultural diversity, economic changes, and social interaction. In human migration studies, migration flows and stocks are typically measured at the national level by official statistics offices, for instance through regular population census. The same goes for other effects such as social and economic integration. These data are critical for the development of policies to optimise the beneficial effects of migration under all criteria. As previously underlined, due to their nature, they can become outdated, or information can be inconsistent when moving from one national statistics office to another. Hence, lately, alternative data are starting to be proposed to measure migration effects in various settings.

The chapter is organised as follows. First, we discuss the current possibility to refer to Big Data to improve the study of human migrations. Then, assumptions on the effectiveness of our index are presented in Section 7.2. After recalling data and preprocessing phase, we describe our Superdiversity Index - Chapter 7.3 and its validation. In detail, we describe the evaluation criteria - see Chapter 7.4.1 -, the null model - see Chapter 7.4.3 -, and the evaluation measures we took into account to demonstrate the effectiveness and robustness of the index, and thus the framework - Chapter 7.4.4. Once defined the assumptions and the index assessment, in Chapters 7.5 and 7.6 we present result obtained about the United Kingdom and Italy, respectively. Starting from issues identified during our analysis, we discuss in Chapter 7.7 the corrective factors able to enhance the algorithm performance. Our conclusions are then proposed in Chapter 7.8.

7.1 Towards Superdiversity 2.0

Every day, each person disseminates digital breadcrumbs over various systems today used for all daily activities. The ever-growing amount of human traces has led to proportional attention towards Big Data. Nevertheless the evolution and the emergence of the various research fields, Big Data Analytics seems to remain considered as a collection of boxes disjoint among themselves. Big Data has been examined through more diverse lines of research, including Sentiment Analysis, e-Science, Linguistics, Healthcare, Geographic knowledge and Geographical Information Systems (GIS), Education, Security, and Privacy. However, these areas often do not communicate between them, and even worse do not collaborate.

As a consequence, in Big Data Analytics still lacks a systematic and multidimensional methodology of analysis allowing observing novel phenomena. Moreover, it is ever more clear that the contents available on social media platforms and microblogs hold great potential. The so-called User Generated Contents (UGC) represent and reflect aspects of human behaviour. These, in conjunction with related metadata, contain various information in terms of geographic, time, sentiment, and language. A conjunct study of multiple fields as the ones just mentioned can allow observing known aspects under different points of view. More interestingly, it can provide novel methods to examine novel patterns, behaviours, and phenomena. The globalisation in conjunction with the evolution of the Internet has created new democratic spaces of interchange which allowed the development of new identities. These identities together with ethnicity, citizenship, residence, origin, and language are all aspects which have brought a "*diversification of diversity*".

As stated by Vertovec [232], the concept of *Superdiversity* aims to acquire an increasingly complex set of relationships between different aspects of human behaviour. Nowadays, the social changes entail superdiversity which in turn urge us to revisit, deconstruct and reinvent many of our established assumptions about language, identity, ethnicity, space, culture, and communication. It is possible then based on Vertovec's concepts to derive a novel contemporary concept of superdiversity. Nowadays, the superdiversity cannot be reduced to the set of changes in migratory flows resulting from globalisation and outline a change in the overall level of migration patterns, as affirmed in the original Vertovec's theory [232, 233]. *Superdiversity 2.0* relies on the perception of socio-cultural communities. This means that superdiversity can aim to identify the phenomena that violate the boundaries of the political, historical, social, cultural and linguistic monocentrism, restricted to a closed spatial framework.

Data Mining applied to microblogs as Twitter creates an invaluable opportunity for changing perspective towards Big Data Analytics. As previously outlined, in this field persists a gap related to the development of a multi-dimensional methodology of user-generated contents concerning sentiment, linguistic, geography, and temporal dimensions. We believe that this gap can be filled through a data-driven framework describing diversity - and superdiversity - in both limited and wide geographical contexts.

7.2 Measuring the *Salad Bowl*

Starting from the concept of Superdiversity as an indicator of significant cultural diversity in the population due to recent migration phenomena, we propose a measure of superdiversity in a population. This *Superdiversity Index* (SI) allows comparing diversity from the point of view of the emotional content of language in different communities. We base our investigation on a base-ground hypothesis:

Hypothesis 1 *Different cultures associate different emotional valences in the same word.*

This means that a culturally diverse community will show a use of the language that is different from a standard expected use. Our SI is built on Twitter data and lexicon-based sentiment analysis. We calculate emotional valences for words used on Twitter by various communities, in the local language, and compare these with emotional valences from a standard tagged lexicon. The distance between the two gives a measure of diversity. We compute our SI at different geographical resolutions, for the United Kingdom (UK) and Italy, and then we compare it with foreign immigration rates. To evaluate the performance of our SI we also compare it with other possible measures of superdiversity extracted from the same Twitter data, such as the use of multiple languages or lexical richness.

7.2.1 Data Description and Resources

The analysis we propose is based on a geolocalised Twitter dataset, on the Data Challenge on Integration of Migrants in Cities (D4I) dataset, and several lexical resources in both English and Italian, i.e., lexicons of words enriched with sentiment valences.

To perform our study, we crossed two datasets, namely the Untagged Extended Twitter Dataset, which is composed of just under 73,175,500 geolocalised tweets, and the *Data Challenge on Integration of Migrants in Cities* (D4I) dataset. Details about the two datasets are provided in Section 5.1.1 and in Section 5.1.2, respectively.

The preprocessing phase we apply to data, aims in obtaining cleaned tweets only coming from the United Kingdom and Italy. Moreover, to be able to perform our analysis we need to match places in our dataset with those in the D4I dataset. To accomplish this task we follow the procedure described in Section 5.2.

At the end of the preprocessing phase, we obtained two cleaned and standardised sub-datasets. The first is composed of all the English tweets posted from the United Kingdom, while the second is composed of all the Italian tweets posted from Italy.

Details regarding the two datasets after preprocessing are shown in Table 7.1.

To apply our sentiment spreading algorithm, we exploit the same resources of lemmas labelled with sentiment valences adopted for developing the algorithm itself. Therefore, we employ:

- the *Affective Norms for English Words* (ANEW) dictionary [37];

Dataset	# tweets	# matched cities
UK	2,088,346	9,603
Italy	274,885	6,050

Table 7.1: Datasets details for the UK and Italy.

- *SentiWordNet* [81];
- the *Full List of Bad Words Banned by Google* of the "What Do You Love" (WDYL) Google project

Details about the three lexical resources are provided in Section 5.3, while the related preprocessing phase is described in Section 5.4.

Due to the multi-language nature of our work, we translate each of the three English lexicons above into the Italian language. As pointed out by Balahur et al. [12], research workers in SA have been reluctant to use machine translation systems due to the low performance they used to have. However, in the last few years, the performance of machine translation systems have greatly improved. For most frequently used languages, open-access services (e.g., Google Translate) offer more and more accurate translations.

To obtain the most accurate translation, we combine and cross-check two different API services, namely Googletrans[1] and Goslate[2]. In addition, we impose some constraints, such a threshold on the translation confidence score. Finally, to be more confident in the translation, we select only lemmas attested in the PAISÀ [138] corpus - see Section 5.3.

7.3 Superdiversity Identification

The SI we propose is based on the emotional content of words for a community. Persons with different cultural backgrounds will necessarily associate different emotional valences to the same word. So, a multi-cultural community will display the use of the local language that is different in its emotional content compared to a standard expected use. We believe that a more diverse community has a larger distance between standard and actual emotional valences. We thus consider a standard lexicon tagged with numeric emotional valences, such as ANEW, and we estimate actual emotional valences for various communities using tweets coming from those communities. We then compute the Pearson correlation r, between the standard and the actual valences of words. We repeat the procedure several times, varying the set of words analysed, and we compute the average

[1] Googletrans is a free and unlimited Google translate API for Python: `https://py-googletrans.readthedocs.io/en/latest/`

[2] Goslate is a free Google Translate API: https://pythonhosted.org/goslate/

correlation $\bar{r}$. The SI is then defined as

$$SI = \frac{1 - \bar{r}}{2} \tag{7.1}$$

A $SI = 0$ indicates no diversity; thus the population uses the language following the standard rules. A $SI = 0.5$ indicates a very significant diversity, where the emotional content of words is not related to the standard one. Finally, a $SI = 1$ - that is unlikely - indicates that the emotional content of words is inverted compared to standard language.

To estimate actual emotional valences for words used by a community on Twitter, we use the algorithm described in Section 6.1. By applying the algorithm, we can construct an extended lexicon tagged with sentiment valences, to perform lexicon-based sentiment analysis on Twitter. Since the extended lexicon is based on the Twitter data to be analysed, it depends highly on the way language is used. We compare the resulting valences of the lexicon with ANEW and compute our SI.

In summary, the algorithm starts by building a network of co-occurrence for all the terms found in a Twitter corpus. Then, a seed dictionary tagged with sentiment valences is used to assign sentiment to some nodes in this network. A spreading process is started, inspired by models of spreading of information and opinions, where nodes that do not have a valence take the aggregated valence of their neighbours if the neighbours agree. The aggregation is done by selecting the mode of the distribution of neighbour valences. Agreement of neighbours is defined by two model parameters. The first is the range of valences of the neighbours, which needs to be below a given threshold R, while the second is the entropy of the distribution of valences of the neighbours, S. These two thresholds are used to control the spreading process and are the only parameters that need to be specified by the user of the algorithm. The final valences of the words are influenced by the initial seed dictionary, but also by the structure of the network where emotional valences spread. This structure is determined by the way Twitter users employ the language. Hence it depends on the cultural mix in the Twitter community. Thus, the final valences show how the language is used in the corresponding community, from the emotional content of words point of view.

To compute our SI, we randomly split the ANEW tagged dictionary into two equally large subsets. One subset is used as seed dictionary by the algorithm, together with SentiWordNet and the Bad Words lexicon. The second subset is used to compute the SI. At the end of the spreading process, the valences in this subset are compared with the valences obtained by the algorithm, to compute r. The process is repeated ten times, with different splits of the ANEW dataset, to obtain the average correlation, $\bar{r}$ and to then apply Equation 7.1 to compute SI.

Our idea is based on the following two assumptions:

Hypothesis 2 *SI, calculated on a Twitter dataset, measures the distance between the emotional content of terms in the language spoken by the Twitter community, as extracted by the algorithm, and the standard language.*

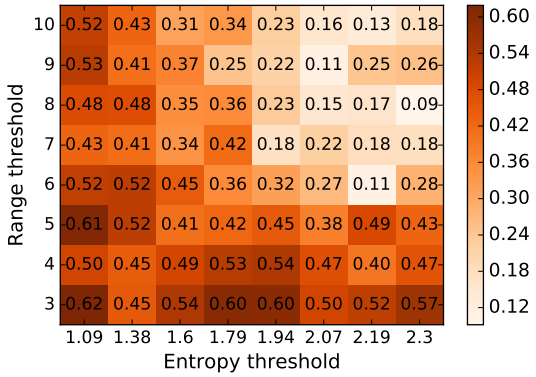


Figure 7.1: Optimisation of model parameters for the UK.

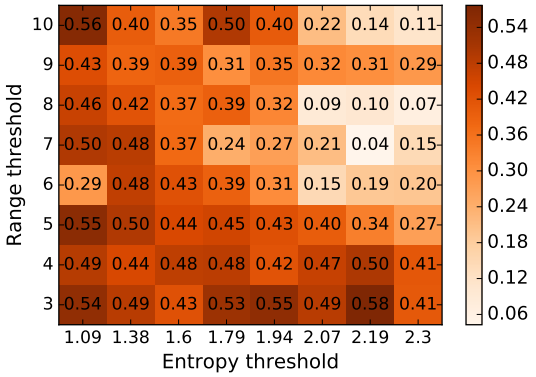


Figure 7.2: Optimisation of model parameters for Italy.

Hypothesis 3 *SI is a measure of diversity in the population from which the tweets are coming.*

SI takes values in the range $[0, 1]$. A value of 0 corresponds to no diversity, i.e., emotional content identical to the expected standard language. A value of 0.5 corresponds to no correlation between the emotional content of words on Twitter and the standard one. A value of 1, which is very unlikely, would correspond to the use of terms with the opposite emotional content compared to standard. The proposed SI inevitably includes a component related to the performance of the algorithm used to compute emotional valences. The algorithm provides an estimate based on a subset of tweets coming from a subset of the total community to be analysed and is based on several assumptions. Thus, this estimate unavoidably includes some error. The correlation r, and in consequence the value of SI, depends both on the different use of the language and on the error of the algorithm. Nevertheless, we believe that the SI we propose can be efficient in quantifying diversity. This is because the error component is stable when changing the cultural mix of the community analysed, or at most it can increase as multiculturalism increases, so it does not negatively affect the relationship between the SI and the diversity, but may even enhance it.

7.3.1 Communities' Superdiversity Index

The selection of the entropy S and the threshold R model parameters was required to compute SI for various communities in Italy and the UK. This was performed through Monte Carlo simulations, using the complete Twitter dataset for Italy and the UK. Specifically, a range of parameters was explored, and for each parameter pair, we applied 10 different instances of the algorithm, on all the Twitter data available. Figure 7.1 shows the average correlation between the valences obtained by the algorithm and the original values in the ANEW dictionary, for each parameter combination, for all the UK data.

We observe that the maximum correlation is obtained for $R = 3$ and $S = 1.09$, as shown

in Figure 7.2. These UK-related parameters are used subsequently to compute the SI at other geographical levels. For Italy, Figure 7.2 shows that best results were obtained for $R = 3$ and $S = 2.19$, which are the values used to obtain all the results related to Italy presented in the following sections.

7.4 SI Evaluation

To demonstrate the efficiency of our model we consider several evaluation criteria, which are described in the first part of this Section. In Section 7.4.1, we expose motivations on which evaluation steps are based. In Section 7.4.3 is presented the Null Model SI, while Section 7.4.4 describes the additional evaluation measures applied to compare obtained performances through our model. For a better comparison, performances obtained through both Null Model SI and the other measures are shown in parallel with model results. Hence, model performances are disclosed in Section 7.5 and in Section 7.6 for the United Kingdom and Italy, respectively. Finally, Sections 7.7 and 7.8 refer to the model's correction factors and conclusions.

7.4.1 Evaluation Criteria

To test whether the proposed SI relates to cultural diversity in a population, we consider as a baseline the foreign immigration rates in the same geographical regions as those where the SI was computed. The immigration rates were extracted from the D4I dataset described in Section 7.2.1. Hence we assume that

Hypothesis 4 *Communities with higher immigration rates also have higher diversity.*

We calculate the Pearson correlation between the SI values and immigration rates, as a measure of the performance of the proposed SI.

7.4.2 Qualitative Evaluation

Besides the Pearson correlation, we perform a qualitative evaluation. Thus, we look at a few examples of assigned valences and the divergence from the new lexicon and the ANEW lexicon. To this purpose, we compare words that are typically polarised, i.e., words typically used only in positive or in negative contexts. Moreover, we look at words that can be used in both positive and negative meanings.

7.4.3 Null Model SI

In the literature, a Null Model is a model generated by using random samples of a specific distribution where elements are allowed to vary stochastically. The null model analysis may be used in specifying

a statistical distribution as well as in randomisation of the observed data when these are designed to predict outcomes of a random process. The process is commonly used to match randomly generated graphs and to prove if and when observed graphs are statistically similar. The Null Model is based on the so-called Null Hypothesis. The latter states that no statistical significance exists in a given set of observations until statistical evidence nullifies it for an alternate hypothesis.

To prove that it is the community that determines the value of the SI and that our correlations are not random, we devise a Null Model SI. To accomplish the test, we reshuffle tweets across geographical regions. Reshuffled tweets are distributed maintaining fixed the number of tweets in each region. We compute the correlation among the Null Model SI and immigration rates and compare them with the original SI correlation.

7.4.4 Evaluation Measures

Besides the Null Model SI, we accomplish an additional evaluation step. During this phase, we compare the performance of the newly proposed SI with that of other possible superdiversity measures extracted from the same data. We consider five additional measures, namely:

1. total number of tweets in the local language.
2. number of tweets *per capita*.
3. the absolute number of languages.
4. the entropy of the distribution of tweets in languages.
5. Token Type Ratio (TTR).

In more detail, the first two measures relate to the frequency of tweeting. One could hypothesise that a more diverse community could tweet more or less. Hence we consider the total number of tweets in the local language, together with the population-normalised version, i.e., number of tweets *per capita*. The second category of measures relates to the different languages spoken by a Twitter community. We would expect that a more diverse community will use more languages. We consider the absolute number of languages, but also the entropy of the distribution of tweets in the various languages. The latter measure takes into account the volume of tweets in each language, besides the number of languages. The fifth possible measure of diversity relates to the lexical richness of the language used by a twitter community. Again, one could expect a richer language from a diverse community. To quantify this, we use a well-known index used in Linguistics, the *Token Type Ratio* (TTR) [219]. The TTR is computed as the ratio between the number of token types that make up the vocabulary and the overall size of the corpus. In our case, it is computed as the ratio between the number of different words and the total number of words in the corpus. The TTR value ranges in $[0, 1]$, where values closer to 1 denote that texts are varied and rich.

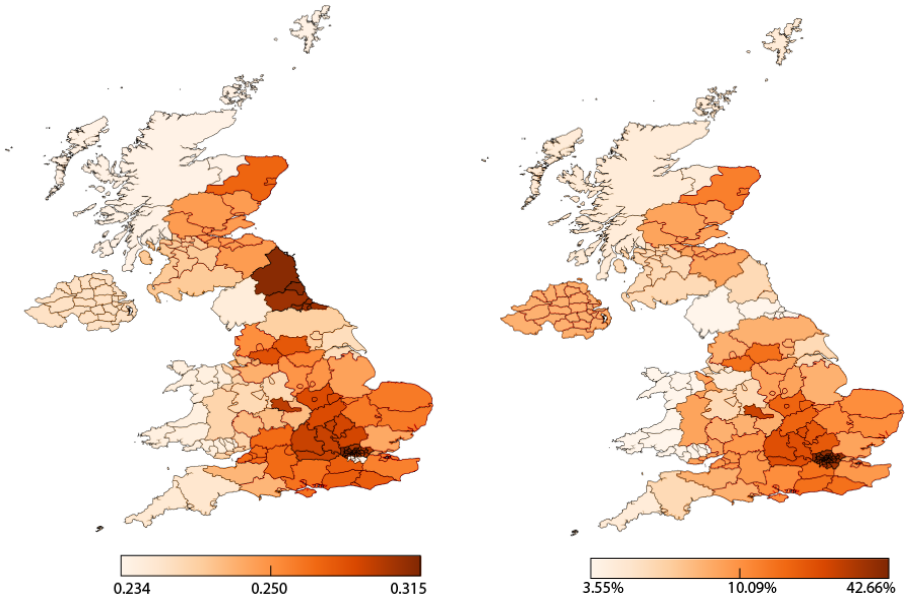


Figure 7.3: Superdiversity index (left) and immigration levels (right) across UK regions at NUTS2 level.

7.5 Superdiversity in the United Kingdom

The proposed SI was computed for the UK at three different geographical resolutions, identified based on the Classification of Territorial Units for Statistics. Hence, this allows us to analyse the NUTS1 level, which corresponds to 12 UK regions while the NUTS2 level to 40 regions. The NUTS3 level contains 174 different regions, out of which we select the 40 with the largest number of tweets.

For computation of the SI, we considered all the tweets in English published in the various regions. Figure 7.3 visually shows the geographical distribution of SI values at level NUTS2, and compares with the distribution of foreign immigration levels, from the D4I dataset. There is a clear similarity between the two maps. To understand better the relation between SI and immigration rates, Figure 7.4 plots the SI values obtained versus the immigration rates, at each NUTS level analysed. We observe that most of the regions align very well on a line, with a very large correlation with the immigration rates.

At all geographical levels, we observe that the regions corresponding to Northeast England and London area appear to have different behaviour, deviating from the main line defined by the other regions. This is also visible on the map, at the regional level. However, when moving from NUTS1 to NUTS2 and NUTS3, we see that, within the two regions, SI grows as the immigration rate grows.

[3]United Kingdom NUTS1: 10 regions
[4]United Kingdom NUTS2: 40 regions
[5]United Kingdom NUTS3: 40 regions

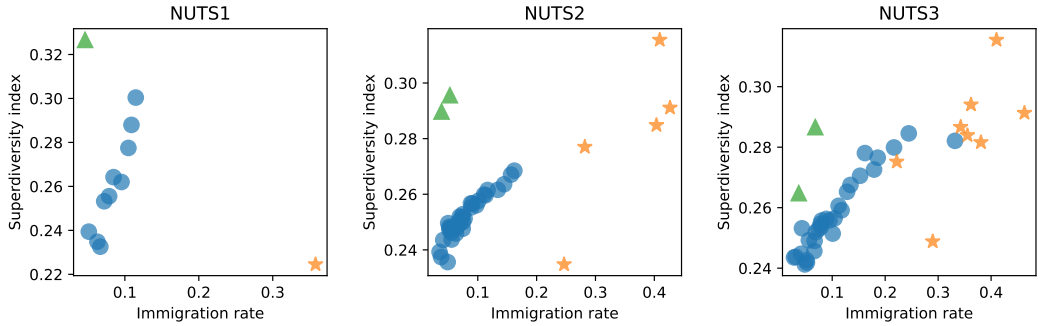


Figure 7.4: SI values versus immigration rates at different geographical levels, for the UK. At the level NUTS3 we selected the top 40 regions based on the number of tweets available in the dataset. The stars correspond to the London area, the triangles represent regions in Northeast England, while the rest of the regions are displayed with circles.

Geo. level	SI	Null model SI	Eng. tweets	Eng. tweets per capita	Languages	Language entropy	TTR
NUTS1[3]	**0.943**	-0.236	0.328	-0.520	0.519	0.481	-0.005
NUTS2[4]	**0.941**	-0.137	0.332	0.007	0.362	0.288	-0.340
NUTS3[5]	**0.928**	-0.221	0.141	0.049	0.322	0.529	0.147

Table 7.2: Correlation between different measures of diversity extracted from Twitter and the immigration rates, at various geographical levels in the UK, excluding London and Northeast England. At the level NUTS3 we selected the top 40 regions based on the number of tweets available in the dataset.

For instance, at level NUTS2, when considering only the five regions from the London area, we see that SI is larger when immigration is larger. The same is true for the regions from Northeast England. We believe this is due to different ranges of SI in different areas. This particular behaviour will be further investigated and discussed. For the rest of this section, we will consider only the remaining regions, i.e., all UK except for Northeast England and the London area. Table 7.2 shows the exact values of the Pearson correlation between immigration rates and the SI, which prove to be remarkably well correlated at all geographical levels.

The null model SI does not correlate at all, as expected, giving evidence that the correlations we obtained are meaningful and related to the source community of the tweets, and not merely due to the number of tweets in each region.

The comparison with other possible measures of diversity is also very favourable to our proposed SI, which is superior to all others, as Table 7.2 shows. No relation between immigration rates and frequency of tweets appears to exist. Some correlation seems to emerge with the number of tweets *per capita*, the number of languages and the language entropy, however much lower than the SI case, and not stable at all geographical levels.

Word	ANEW	UKI	UKC	UKL	UKJ
cheer	8.10	9.09	9.00	8.75	9.05
fun	8.37	9.00	8.48	9.41	8.00
joyful	8.22	9.21	8.00	9.26	7.00
war	2.08	1.10	1.00	1.50	2.50
hostile	2.73	1.96	3.00	1.88	1.97
unhappy	1.57	1.56	0.41	1.02	0.42
news	5.30	5.00	8.00	5.68	2.00
leader	7.63	8.12	2.00	8.43	5.00
social	6.88	7.44	2.00	6.44	1.00

Table 7.3: Examples of valences in ANEW and in new lexicons for selected words. Lexicons displayed relate to UK NUTS1 level. New lexicons refer, from left to right, to the London area (UKI), North East England (UKC), Wales (UKL), and South East England (UKJ).

Finally, we perform a qualitative analysis of assigned emotive valences to check that Hypothesis 4 does indeed make sense. To this purpose, we look at words that typically have a strong polarisation, i.e., words generally considered positive and words generally considered negative. Furthermore, we observe the words that can be associated with both negative and positive contexts. This allows us to observe the divergence from the new lexicon and the ANEW lexicon.

We believe that communities with higher immigration rates also have a higher diversity. However, we also think that some words have a positive or negative connotation that is almost universally accepted. As a naive quantitative analysis, we can consider the valences of these words as a kind of standard parameter. Finally, the words that can be used in both positive and negative contexts allow us to observe if the Superdiversity Index captures the different emotional uses of words depending on the geographical area.

In Table 7.3, we report few examples of words and their assigned valences depending on the geographical dimension. The first column shows words having a positive or negative connotation that is almost universally accepted. The first three words can typically be considered positive, while the second three as negatives. Moreover, we also consider three words that in the new lexicon have different values from those in ANEW. The other columns of the table show the valences assigned in ANEW and in the new lexicon. We report valences obtained for a selected run for four different regions at UK NUTS1 level. We choose two areas in which the SI shows high correlations with immigrants rates - Wales (UKL), and South East England (UKJ). Moreover, we look at the two regions where SI ranges do not match those from the rest of the UK - London area (UKI), North East England (UKC).

As shown in Table 7.3, valences assigned in UKI and UKL lexicons seems coherent with those in ANEW for all words. On the contrary, valences extracted from UKJ and UKC lexicons show

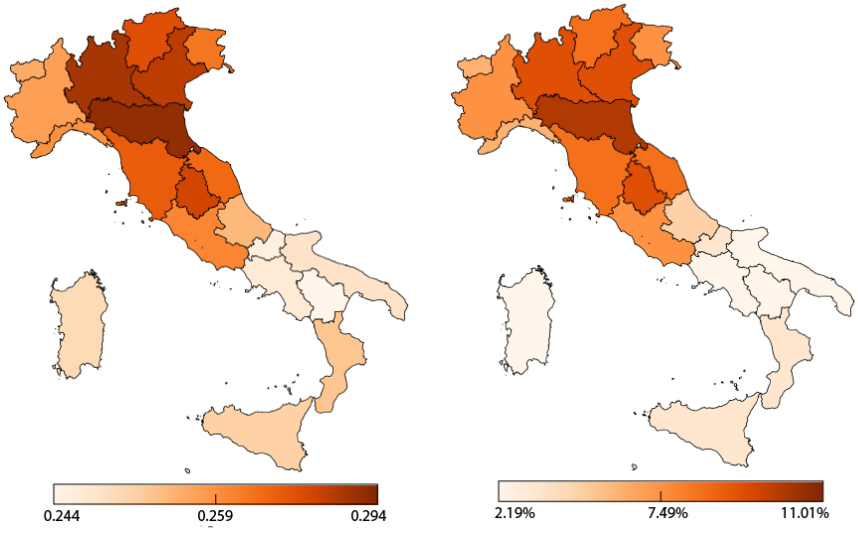


Figure 7.5: Superdiversity index (left) and immigration levels (right) across Italian regions at NUTS2 level.

different behaviour. Positive and negative words show coherent valences concerning those in ANEW. However, valences of words that can be used in both negative and positive contexts do not match with ANEW. Since that positive and negative words show coherent valences, we believe that our SI index makes sense. Moreover, it can catch variations in the emotional use of words depending on the geographical area.

7.6 Superdiversity in Italy

The analysis accomplished over UK-related data was repeated for Italy. Following the method described in Section 7.2.1, the lexical resources are translated into Italian. This allows us to apply the method explained above to all Italian tweets published from different Italian regions. Figure 7.5 shows the geographical distribution of both the SI values and the immigration rates from the D4I dataset, at level NUTS2 - regional. As previously seen, there is a very good similarity between the two maps. In Figure 7.6 we plot the SI values obtained by our method versus the immigration rates. Excellent correlation with immigration can be observed, at all geographical levels. Exact obtained correlations are reported in Table 7.5, with values over 0.85 at all levels.

It is necessary to underline that at level NUTS1, Italy is divided into only five regions, hence here correlations are not really meaningful. We report them for completeness. However, is the behaviour at the following levels that carries significance. At NUTS2 there are 20 regions, while at NUTS3 we select the top 20, similar to the UK case. Table 7.5 also shows results for the null

Word	ANEW	ITH	ITI	ITF
cheer (rallegrare)	8.10	9.00	8.45	9.05
fun (divertimento)	8.37	8.75	9.41	9.38
joyful (gioioso)	8.22	8.33	9.19	8.00
war (guerra)	2.08	1.10	1.23	2.00
hostile (ostile)	2.73	1.99	2.00	1.82
unhappy (infelice)	1.57	1.88	0.89	0.42
news (notizia)	5.30	4.00	5.35	6.00
leader (capo)	7.63	6.00	4.00	6.00
social (sociale)	6.88	7.00	7.44	7.53

Table 7.4: Examples of valences in ANEW and in new lexicons for selected words. Lexicons displayed relate to Italy NUTS1 level. New lexicons refer, from left to right, to Northeast Italy (ITH), Central Italy (ITI), and South Italy (ITF).

model. As are expected, the null model SI does not correlate with immigration rates. As for the other possible diversity measures, none of them seems to give any hint of the immigration rate, at levels NUTS2 and NUTS3. Hence our proposed SI is undoubtedly superior. At level NUTS1 we see some correlation, but again we believe these values to be spurious since we are considering only five geographical areas.

To perform a qualitative analysis of assigned valences, we follow the same approach carried for the UK. Thus, we choose to show the same words identified before as positive and negative. Moreover, we also look at three words that can be used in both negative and positive contexts. Lexicons chosen refer to Northeast Italy (ITH), Central Italy (ITI), and South Italy (ITF). As mentioned before, in the Italian case, the SI index well correlates with immigrant rates in all levels and for all regions. Valences we assign are shown in Table 7.4. Unlike UK case, here all valences seem to be, more or less, coherent with those in ANEW. Concerning the word "leader" ITI-related valence is lower than the one in ANEW. This may be explained by the fact that in Italian, the translation of the word leader can be used to identify the employer. Thus, as a consequence, it could be used as a negative word.

7.7 Corrective Factors

The results we obtain show a strong link between the proposed SI and foreign immigration rates. However, as previously underlined, in the UK case, a small number of regions seemed not to show the same behaviour as the rest of the country. In details, these relate to the London Area and Northeast England. In the first case, although immigration rates in London are much higher than the rest of the country, the SI value extracted from all tweets in London was smaller (NUTS1). However,

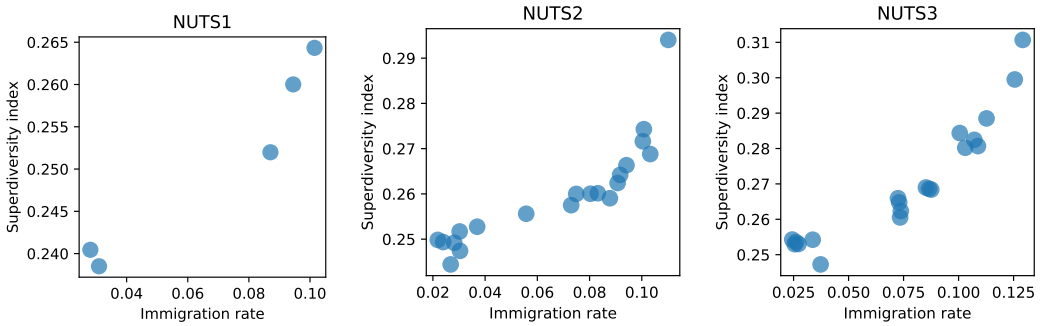


Figure 7.6: SI values versus immigration rates at different geographical levels, for Italy. At the level NUTS3 we selected the top 20 regions based on the number of tweets available in the dataset.

Geo. level	SI	Null model SI	Ita. tweets	Ita. tweets per capita	Languages	Language entropy	TTR
NUTS1[6]	**0.963**	-0.437	0.735	0.696	0.183	-0.585	-0.727
NUTS2[7]	**0.859**	0.143	0.279	0.282	0.304	0.099	-0.243
NUTS3[8]	**0.924**	0.082	0.081	-0.148	0.216	0.021	0.091

Table 7.5: Correlation between different measures of diversity extracted from Twitter and the immigration rates, at various geographical levels in Italy. At county level (NUTS3) we selected the top 20 regions based on the number of tweets available in the dataset.

when dividing tweets per county, within the London area, SI values seem to grow as immigration levels grow (NUTS2 and NUTS3). Hence, it appears that the ranges of the SI obtained are different in London compared to the rest of the country. The same applies to Northeast England. Even if immigration rates are low, SI values were high. Again, those regions seem to form a cluster of their own, where SI ranges do not match those from the rest of the UK.

Considering these particular behaviours, to make the SI range uniform, we believe that the immigration rates are not sufficient. For this reason, we had decided to identify correcting factors or at least determine the various clusters, without using the immigration rate itself. We regard that this should also make SI values comparable across geographical resolutions, since we observe that, for the same immigration rate, SI values can vary from one NUTS level to another.

The first correcting factor we consider is the language entropy. We observed that the London area displays, at NUTS1 level, a much higher language entropy compared to the rest of the country, hence it can be used to re-scale the SI. Figure 7.7 shows language entropy for all NUTS1 regions.

As shown in Figure 7.8, in Italy language entropy is very similar across regions. In fact, in Italy,

[6]Italy NUTS1: 5 regions
[7]Italy NUTS1: 20 regions
[8]Italy NUTS1: 20 regions

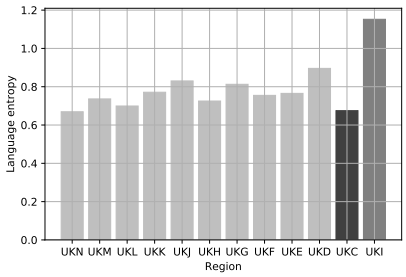


Figure 7.7: Language entropy on tweets originating from the UK, at macro scale (NUTS1 regions). The region UKI corresponds to the London area, while UKC to Northeast England.

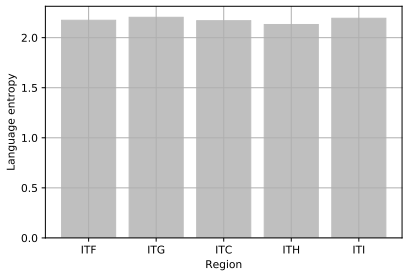


Figure 7.8: Language entropy on tweets originating from Italy, at macro scale (NUTS1 regions).

there seem not to be any range issues, since all regions overlap very well. The uniform and large entropy in Italy could be explained by the fact that Italy is a popular tourist destination. As a consequence, there exist many tweets in different languages. For instance, in central Italy, we have 25,044 tweets in English and 67,903 in Italian, a factor of only 2.7 between the local and a foreign language.

A second important factor could rely on the cultural differences of the local non-immigrant community, such as those represented by local dialects. Higher use of the local dialect could explain the apparent larger SI values in Northeast England. To correct for this, a baseline SI value could be computed from a subset of tweets coming from local users - for example, local newspapers or official accounts. This baseline could be used to correct for range differences to the rest of the country.

Given the high correlation to foreign immigration rates, we believe that our SI represents the first step towards a novel nowcasting model for migration stocks. Once all correcting factors are identified, these could be used, together with the SI, to build a highly accurate model to estimate immigration rates. We expect that a standard machine learning model could prove suitable for this task, that we plan to undertake in future work. Such a model will enable accurate immigration statistics without the need for time - and resource - consuming population censuse. By repeating the analysis regularly, we will be able to maintain updated statistics, valid also in regions where census is not possible or inaccurate due to clandestine immigration.

7.8 Conclusions

In this section, we proposed a novel superdiversity index which quantifies diversity in a population based on changes in the emotional content of words with respect to the standard language. These changes have been estimated in the UK and Italy by using geolocalised Twitter data at different

geographical levels. Results we obtained show a significant correlation with foreign immigration rates in almost all geographical analysed regions. During our analysis, we also compare performances obtained by applying other possible measures of diversity form the same Twitter data. Results show that the proposed Superdiversity Index greatly outperforms all the other measures.

Part III

Superdiversity in Music contexts

Chapter 8

The Music Context

Music has been part of human civilisation for centuries: it has been referred to as the universal language. Certainly, every culture has given birth to its music. The constant collapse of physical barriers caused the overall globalisation of music. During the last decade, the constant growth of online streaming services - such as Spotify, Apple Music, Last.fm - have made available to the public the widest choice of music ever. Emerging bands, as well as famous ones, can obtain global visibility that was unimaginable only a few years ago. Cultural and societal evolution, as well as national and international historical events, reflect their essence in both lyrics and melodies produced by the ones who experienced them. Indeed, music is one of the most valuable expressions of national identities. However, it is rare that a country can be effectively described as a single "cultural" entity: it is natural to expect that each region within it has its peculiarity. In this rapidly evolving multi-faceted scenario, music seems to have lost its geographical-cultural connotation: Are we observing a growing standardisation of music contents? Are there peculiar characteristics able to discriminate the music produced in a given region from the one produced in the country that contains it? Can these observations be applied to the Italian music scene?

Starting from these questions, in the following sections we deeply investigate the current music scene at different granularity levels. This chapter focuses on the music domain with particular attention to the current music scene, that has been already introduced in Chapter 4. A state-of-the-art related to the music has been already discussed in Section 4.1.

In the first section of this chapter - Section 8.1, we describe data, resources and preprocessing we apply to perform our analysis. In the second section of this chapter - Section 8.2, we analyse how the new generation of Italian musicians relates to the musical tradition of their country. Thus, we first investigate characteristics of regional music and then we place emerging artists within the musical current defined by famous artists coming from the same country. Section 8.2 is partitioned as follows. The dataset we employ to perform our analysis is described in Section 8.2.1, as well as its preprocessing. Section 8.2.2 describes how we compute regional and national profiles. Section 8.2.3

Dataset	#Artist	#Tracks	#Genres	#Lyrics
WORLD	833,197 (19,218)	5,525,222	1,380	79,204
ITALY	2,379 (710)	502,582	126	28,582
TUSCANY	513 (58)	24,147	28	91

Table 8.1: Datasets statistics. Within brackets are reported the number of artists for which at least a single song lyric was available.

focuses on dynamics occurring between Tuscany emerging bands and their popularity, and finally, Section 8.3 discusses our observations and conclusions.

The following chapters, namely Chapter 9 and 10, describe the application of a multi-faceted study of the musical context. As previously mentioned, we believe that in Big Data Analytics there exist a gap. This gap is represented by the lack of a conjunct study of various dimensions. To fill the gap, we propose an examination of the musical context which takes into account emotional contents, geography, and linguistic diversity at once.

8.1 Music Data, Lexical Resources and Preprocessing

In this section, we describe the datasets and the preprocessing phase we adopted to develop the studies related to the musical context. Since works presented in Chapters 8.2, 9 and 10 exploits same data and resources, we describe them in the following, while in proper chapters we will limit to recall them.

8.1.1 Musical Dataset

To carry out our analysis, we exploit a musical dataset having three different levels of spatial granularity: world, national and regional. Datasets refer to the *world*s famous musicians - `WORLD` dataset -, *Italian* musicians - `ITALY` dataset, and emerging youth bands in *Tuscany* - `100band` or `TUSCANY` dataset -, respectively. In particular, the `TUSCANY` dataset is referring to emerging artists participating in the "100 Band" contest promoted by "Tuscan Region" and "Controradio" in 2015[1]. The world dataset has more than eight hundred thousand authors, the Italian dataset has more than two thousand top Italian authors and the `TUSCANY` dataset is composed about five hundred emerging Tuscany artists. In Table 8.1 we describe the details of these datasets.

The three datasets are built using the Spotify API[2] and are composed of all the songs present on the platform for the selected artists.

[1]Toscana100band contest: `http://toscana100band.it/`
[2]Spotify API: `https://developer.spotify.com/documentation/web-api/`

For each selected artist, we collect song titles, songs popularity score, album titles, *Spotify ID*, and the list of the genres the artist is associated with. If an artists genre is not set, the array is empty. In the following, we will explain the way we fixed the lack of the musical genre. Regarding the popularity score, each track on Spotify is characterised by a set of "Popularity bars" that when aggregated indicate how popular the track is. In general, the popularity score of a song is based on two parameters: *a*) the total number of plays compared to other tracks; and *b*) how recent those plays are. In addition to the features listed above, for each song, we collect the set of musical features provided by Spotify[3].

Also, each track is described by a set of musical features[4] provided by Spotify, namely *acousticness, danceability, duration, energy, instrumentalness, liveness, loudness, speechiness, tempo*, and *valence*. All the features range in [0, 1] except *duration, loudness*, and *tempo*. In the preprocessing phase, we normalise these latter features to align all the feature scales.

We further integrate the datasets with the lyrics of the songs. In Table 8.1 are also described the details of the lyrics datasets. In particular, the `WORLD-lyric` dataset is collected from the Genius[5], the `ITALY-lyric` dataset is collected using SoundCloud API[6], and, finally, the `TUSCANY-lyric` is built extracting texts from the results of a survey. Using Google Form service[7], we gathered both musical and personal data about artists who participated at the Tuscany 100 Band contest.

Moreover, to analyse the musical context also from the geographical point of view, we integrate the dataset with a geographical field. For this purpose, we use the artists Spotify ID as research key in the Echonest API. Using this API, we obtained the Italian region where each musician comes from. Moreover, to perform our analysis, we integrate these musical datasets with the lyrics of the songs retrieved for each artist. In particular, the `ITALY-lyric` dataset is collected using SoundCloud API. On the contrary, the majority of the Tuscany emerging bands lyrics are not available on public platforms. Due to this reason, the `TUSCANY-lyric` dataset is built by extracting information from the results of a survey. By using the Google Form service[8] we gathered both musical and personal data regarding artists who participated in the Tuscany 100 Band contest of 2015.

8.1.2 Musical Features

As well-known, Spotify has released an API which allows obtaining a JSON metadata about music artists, albums, and tracks, directly from the Spotify Data Catalogue. Into the API, a full `track_object` is described through several metadata. Amongst others, these include album, that is a simplified `album_object` representing the album in which the track appears; artists, as an array

[3]Musical features provided by Spotify have been already described in Section 8.1.2
[4]Spotify Audio Features Object, `https://developer.spotify.com/web-api/get-several-audio-features/`
[5]Genius: `https://genius.com/`
[6]SoundCloud API: `https://developers.soundcloud.com/docs/api/guide`
[7]Google Form service: `https://www.google.com/forms/about/`
[8]Google form service: `https://www.google.com/forms/about/`

of simplified `artist_object` which identifies who performed the track; `available_markets`, as an array of strings containing a list of the countries in which the track can be played, identified by their ISO 3166-1 alpha-2 code. Besides standard information, a `track_object` is also described through a set of ten musical features. Since these features are at the basis of our music-related analysis, in the following, we provide a brief description of each of them.

- `acousticness` is related with how much the track is acoustic
- `danceability` describes how suitable a track is for dancing based on the combination of musical values, i.e., *tempo*, rhythm stability, beat strength, and overall regularity.
- `duration_ms` indicates tracks duration in milliseconds.
- `energy` is a measure that represents a perceptual measure of intensity and activity, which includes dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- `instrumentalness` predicts whether a track contains no vocals.
- `liveness` is related to the presence of an audience in the recording.
- `loudness` is the measure of the overall loudness of a track in decibels (dB).
- `speechiness` is related to the presence of spoken words in a track.
- `tempo` is the *tempo* of a track in beats per minute (BPM).
- `valence` describes the musical positiveness conveyed by a track.

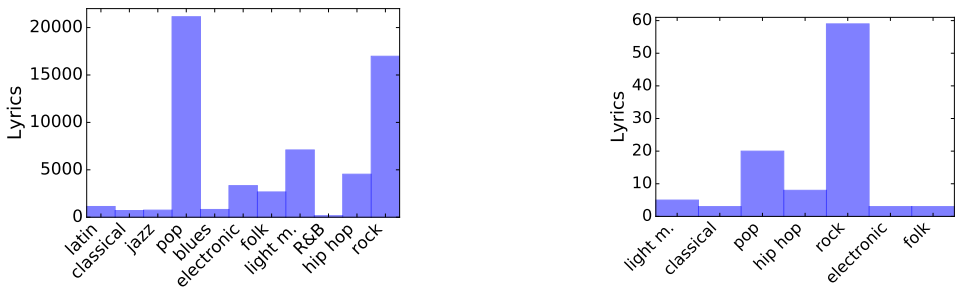
8.1.3 Preprocessing

In this section we discuss the preprocessing phase we perform on the above-mentioned datasets.

The first problem faced is the lack of genres that, once obtained, are normalised and aggregated, as explained later. With this scope, we integrate our data with data from another web resource, namely Wikipedia. For each artist with an empty genre field, we make a call to the Wikipedia Italian version using the name of the artist as the keyword.

Once having obtained all the genres for the musicians who lack them, we proceed by aggregating them into a few major music genres classes. All three datasets show a large number of genres, both minor and major. To reduce noise in music genres, we decide to group them based on a large-grained classification. To this end, by using two lists of popular music genres[9], we assign each songs minor genres to their major class. Once we obtained major-genres classes, we further simplify the classification. Thus, we distribute 234 different genres in 12 classes: *country, blues, religious,*

[9]Lists are gathered from AllMusic - AllMusic, `http://www.allmusic.com/genres` - and from Wikipedia - W. List of popular music genres, `http://en.wikipedia.org/wiki/Psychology`



(a) ITALY (b) TUSCANY

Figure 8.1: Genres distribution among datasets.

hip hop, latin, electronic, folk, jazz, rock, R&B, pop, and *a cappella*. This step allows us to bring together classes' major sub-genres under the principal one. For instance, *hard rock, heavy metal*, and *alternative rock* sub-genres have been labelled as *rock*. Genres distribution in ITALY and TUSCANY datasets is shown in Figure 8.1.

Once we addressed the music-genres related issue, the preprocessing phase is focused on music lyrics. To process music lyrics with unsupervised methods, we first clear the data. Therefore, to obtain normalised texts, we treat lyrics datasets using a rule-based cleaning procedure. Rules we applied are partly related to the specific music domain. The normalisation step covered the following aspects:

- *Stop words*. We detected and eliminated general English and Italian stop words, URLs and domain-specific stop words, such as *strofa, ritornello*, and *rit*.
- *Featuring*. Music is often characterised by several authors featuring, which creates a problem in handling multiple authors. We decided only to maintain the name of the first author.
- *Punctuation*. In this step, punctuation has been removed.

Following this method, we obtain standardised music lyrics that are then treated by a general-purpose pipeline of Natural Language Processing (NLP). After that, music lyrics are lemmatised and tagged with the POS tagger TreeTagger [205]. We also reduce the noise selecting only nouns, verbs and adjectives. In this way, for each text, we obtain only significant words from the sentiment points of view.

8.2 Placing Emerging Artists on the Italian Music Scene

Nowadays, thanks to the maturity of online music-related services, music consumption has become ubiquitous. During the last decade, several works analysed music data collected from online services to study users behaviours and tastes. One peculiar trait that all those works underlined is that the

easy access to enormous music libraries has made possible even for emerging artists to reach audiences that were unimaginable only a few years ago. In this work, we study in a particular context, the Italian music scene, how the new generation of musicians relates to the musical tradition of their country. Taking advantages of data collected from Spotify, we investigate the peculiarity of regional music and try to place emerging artists within the musical movement defined by the already famous colleagues of the same country.

As mentioned before, society evolution reflects its essence in both lyrics and melodies produced by musicians. As a consequence, music is a valuable expression of national identities. Since a country cannot be described as a single "cultural" entity, should we expect that each region within it has its peculiarity? Does this observation apply to the Italian music scene?

Once we profiled the Italian regions by looking at the characteristics of the songs of their most famous artists another question arises: Will emergent artists adhere to the music canon identified for their region? If not which are their strongest influences? Being able to answer such questions will act as a linchpin for shedding lights on how the new generations define themselves, on which are the major changes in music tastes that are currently taking place. Moreover, extending the analysis on broader, national, profiles we can better understand how emergent artists place themselves in the music scene: Do they chose their style to mimic famous artists so to sound more "appealing" to the public, or Do they pursue their passion disregarding the current tastes?

The rest of this section is partitioned as follows. The dataset we employ to perform our analysis is described in Section 8.2.1, as well as its preprocessing. Section 8.2.2 describes how we compute regional and national profiles. Finally, Section 8.2.3 focuses on dynamics occurring between Tuscany emerging bands and their popularity.

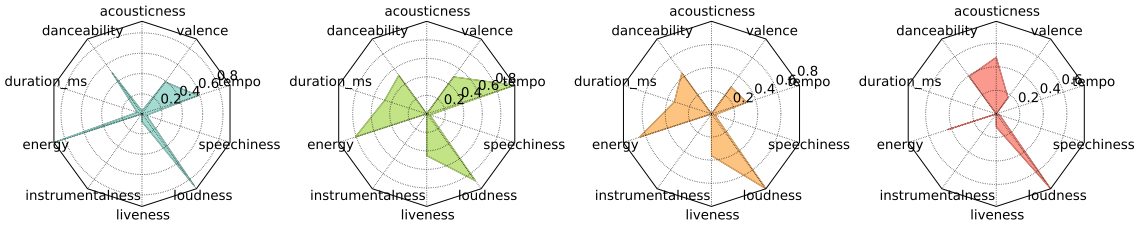
8.2.1 Dataset

To carry out our analysis, we used two different datasets. The first, called `ITALY`, relies on the famous *Italian* musicians, while the second, called `1000band` relies on emerging youth bands in *Tuscany*.

Both datasets are composed of all the songs present on the platform for the selected artists. For each artist, we collect various information, as well as for their songs. Moreover, each track is described by a set of musical features[10] provided by Spotify.

A brief description of Spotify's features we gathered and used is provided in Section 8.1.2. Details regarding the entire dataset and the method adopted to build it are described in Section 8.1.1. Finally, the preprocessing we apply is discussed in Section 8.1.3.

[10]Spotify Audio Features Object, `https://developer.spotify.com/web-api/get-several-audio-features/`



(a) 100band (b) Tuscany (c) Emilia-Romagna (d) Molise

Figure 8.2: Artists Profiles: (a) the `100Band` medoid, (b-d) profiles of the famous artists from Tuscany, Emilia Romagna and Molise.

8.2.2 Regional & National Profiles

As a first step, we grouped songs by artist, and for each of them, we extracted a profile. We describe each artist with his *medoid song*, i.e., his most representative track identified by minimising the sum of the Euclidean distances between the Spotify features among all its discography. Once we profiled both famous and emerging artists, we focused our attention on similarities among them, both at regional and nationwide levels.

Regional profiles. Indeed, music is one of the most valuable expressions of national identities. However, it is rare that a country can be effectively described as a single "cultural" entity: conversely, it is natural to expect that each region expresses its peculiarity. Does this observation apply to the Italian music scene? Can each Italian region be characterised by the music it produces?

To answer such questions, moving from the computed artist profiles, we leveraged the geographical information attached to famous artists in `ITALY` to build region-wide profiles. Thus, we grouped artist profiles by their region of provenance, and then we describe each region with its medoid song. Through radar charts[11], Figure 8.2 shows the profiles obtained for Tuscany, and Emilia Romagna and Molise as examples. Tuscany and Emilia Romagna are both characterised by cheerful and dance songs with a fast pace and instrumental beat. However, Emilia Romagna is represented by more energetic and fast tracks, like the songs of Gem Boy. Molise, instead, is represented by melodic music with a calm and repetitive rhythm, like the songs of Fred Bongusto. Our results suggest that each Italian region has its music peculiarities: what about emerging artists in `100band`?

In Figure 8.2(a) we show the dataset wide profile of emerging Tuscan bands. We can easily notice that such profile differs from the one expressed by famous artists of the same region. Given such distance, to which regional musical tradition Tuscan emerging artists look up to? We assigned each

[11] A radar chart is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point. The relative position and angle of the axes are typically uninformative. In our case, the axes represent the musical features described in Section 8.1.2.

Region	% Artists
Emilia Romagna	0.46
Molise	0.13
Valle d'Aosta	0.11
Piemonte	0.05

Region	% Artists	Avg. Popularity
`Cluster0`	0.61	22/100
`Cluster1`	0.20	15/100
`Cluster2`	0.13	19/100
`Cluster3`	0.05	16/100

Table 8.2: Similarity among Tuscan emerging bands and the profiles of famous Italian artists at regional level *(left)* and with respect to data-driven national profiles *(right)*.

of the bands in `100band` to its most similar regional clusters by applying *K-nearest neighbours* (K-NN). Table 8.2 (left) reveals that most of the emerging bands are assimilated to Emilia Romagna (46%) followed by Molise and Valle dAosta: those are the musical scene that better capture the features expressed by Tuscan emerging artists. The rest of the bands is divided into low percentages among other regions.

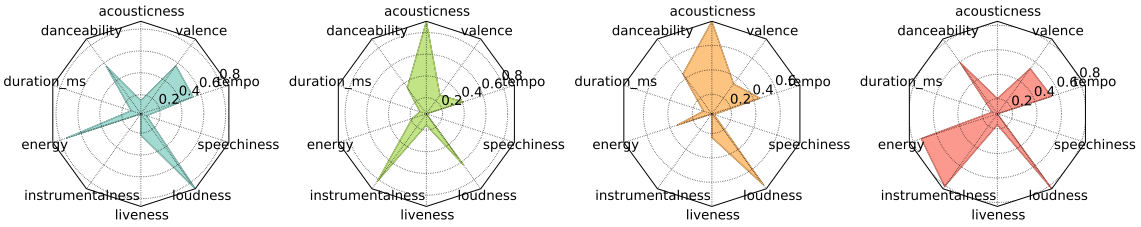
National profiles. To give a more comprehensive classification of the Italian music scene, we employed the K-means clustering algorithm to identify homogeneous clusters among the artists in `ITALY`. After calculating the *Sum of Squared Error* (SSE) distribution for $2 < k < 20$ we selected $k = 4$ because for $k > 4$ the SSE does not decrease significantly anymore. Thus from each cluster, we extracted a medoid resulting in four profiles.

Through radar charts, Figure 8.3 shows the four profiles we obtained. Observing Figure 8.3 we immediately perceive that `cluster0` and `cluster3` are expressions of artists in direct opposition to the ones respectively in `cluster1` and `cluster2`:

- `cluster0` and `cluster3` are represented by happy songs, dance music, and songs with a strong beat. Though the songs of `cluster0` have regular rhythm, `cluster0` has the highest value `speechiness` and includes artists such Datura, Linea77 and rappers (i.e., Emis Killa), while in `cluster3` there are Working Vibes and Persian Jones.

- `cluster1` and `cluster2`, conversely, represent most melodic songs, less rhythmic. Clusters differs only in `instrumentalness`, which is higher in `cluster1`. The `cluster2`, which has lower `instrumentalness` value, includes artists, such as Giovanni Allevi and Stefano Bollani, while `cluster2` includes Lucio Battisti and Carla Bruni.

Outlining the regional analysis, we assigned each of the 100bands to the closest cluster using K-NN fixing as known groups the medoids of the clusters identified by K-means. More than half of the emerging bands (61%) are labelled similar to the cluster0 profile, a cluster containing famous artists of Emilia Romagna and Tuscany. The rest of the bands is mainly allocated in `cluster1` and `cluster2`, with only a small part in `cluster3`, as shown in Table 8.2 (right).



(a) `Cluster0` (45%) (b) `Cluster1` (14%) (c) `Cluster2` (29%) (d) `Cluster3` (12%)

Figure 8.3: Artists Profiles: Italian clusters medoids.

8.2.3 Popularity

How do emerging artists decide their musical style? Do they follow their passion or do they try to mimic what is perceived as the people taste? To provide answers to such questions, we calculated the average popularity of the famous Italian artist for each cluster - as perceived by the Spotify users - and we use this indicator to label the identified groups. As shown in Table 8.2 (right), we observed that there are no strong differences among the clusters means - the same observation holds for their median and standard deviation. However, it needs to be underlined that the *popularity* perception is computed on a service, Spotify, that connects worldwide artists and listeners. In this scenario, the popularity of Italian artists is somehow penalised due to the intrinsic restriction imposed by the non-adoption of English as the primary lyric language. Taking into account such distortion, we can anyhow observe how emerging bands tend to be assimilated to `cluster0` and `cluster3` - almost 74% of `100band` -, clusters grouping together artists having greater average popularity. Somehow, emerging bands seem to mimic characteristics of known artists that receive more attention from the public. *To reach success, first you must learn to play by the rules; then you must forget the rules and play from your heart.*

8.3 Conclusions

With our work, we seek to identify regional, national, and Tuscan emerging bands identities through music analysis. It seems clear that it is not easy to define general national and regional profiles because of the heterogeneity of music. However, our work underlines that some features are discriminating between the various groups identified during the analysis phase.

Our analysis has underlined that the geographical dimension can be used to identify musical identities. Therefore, following a data-driven investigation, we study how geography affects music production and relations between technical characteristics of the music produced at regional, national, and world level.

Chapter 9

The Fractal Dimension of Music

In the previous chapter, we seek to identify regional, national and Tuscan emerging bands identities through a data-driven music analysis. As underlined, to describe the entire national music scene is not a trivial task because of the strong heterogeneity of music. In the sections of this chapter, we propose another data-driven investigation related to the musical context. By leveraging a cross-service multi-level dataset, we study how geography affects music production. Furthermore, we study how artists producing a specific type of music can reach high popularity and how such popularity is often not related to a specific genre. Our data-driven investigation highlights the existence of a "fractal" structure that relates the technical characteristics of the music produced at regional, national and world level. A "fractal" structure is a structure that is characterised with self-similarity, i.. it is composed of fragments whose structural motif is repeated if the scale changes. Thus, we are interested in understanding how technical characteristics of the music change to vary the geographical scale.

Starting from emergent groups of an Italian region - Tuscany -, moving to affirmed Italian artists and finally to a set of world-famous musician, we identify a multi-level set of profiles transverse to the classical concept of genre: we show how such profiles remain stable across the different geographical layers analysed. Finally, we observed how the mood expressed by artists songs as well as their popularity vary with respect to the multilevel traversal profiles they belong to. Once again, we observed the presence of the same structure that clearly emerges both at different geographical levels and over a multi-level set of profiles.

The chapter is organised as follows. Section 9.1 introduces the datasets used in the current study and describes the preprocessing steps performed on them. In Section 9.2 we show and discuss the fractal structure emerging from the datasets and the music profiles that we were able to identify applying an unsupervised learning strategy. Section 9.3 highlights the relationships of the artists popularity and music genres concerning the groups discovered. Following the same approach, in Section 9.4 we analyse the relationships between the songs' lyrics and music profiles enriched with

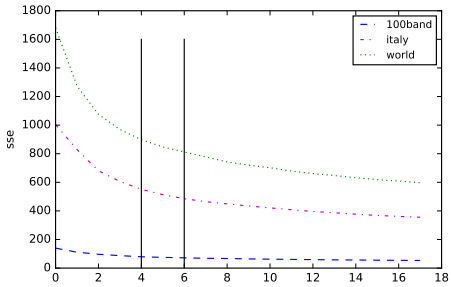


Figure 9.1: Sum of Squared Error (SSE) distributions for $k \in [2, 18]$.

Sentiment Analysis.

9.1 Dataset and Preprocessing

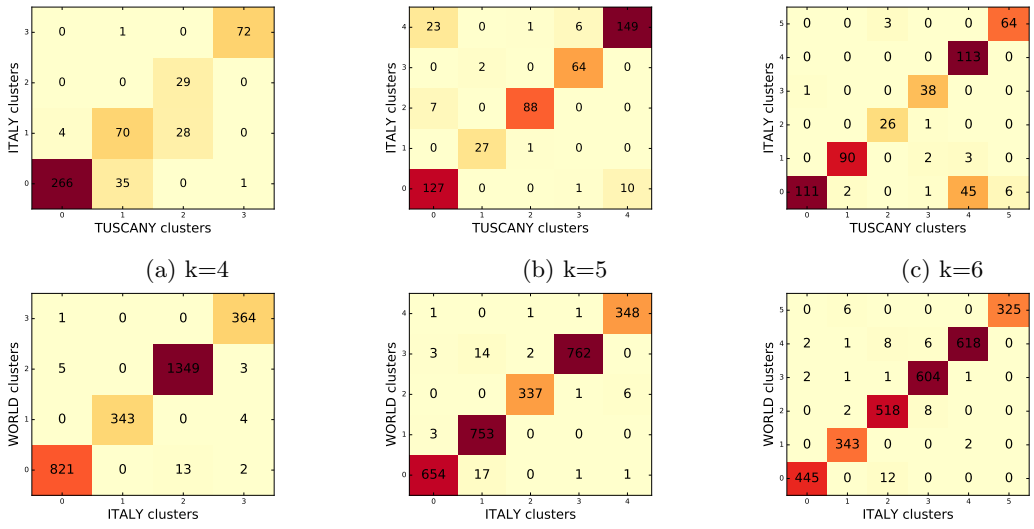
As a proxy of the current musical scene, we exploit a musical dataset having three different levels of spatial granularity: world, national and regional. Datasets, which are described in Section 8.1.1, refer to the *world*s famous musicians, *Italian* musicians, and emerging youth bands in *Tuscany*, respectively. Contrary to the work presented in Chapter 8.2, the `100band` subset is here named `TUSCANY`. The change in naming does not represent any changes in the dataset. It is only related to the comprehensibility and clarity of the phases of the work. Each track in the three datasets is also described by a set of musical features[1] provided by Spotify, which are described in Section 8.1.2.

Moreover, to deepen our analysis and consider only the mood transmitted by the players in terms of lexical content, we integrate these datasets with the lyrics of the songs. To preprocess data, we follow the method presented in Section 8.1.3.

9.2 The Music Scene Fractal Structure

To identify the prototypical type of music produced by each artist in the datasets, we describe every performer through his *medoid* song, i.e., his most representative song identified minimising the sum of the Euclidean distances between the Spotify features among all his discography. Once we built such descriptions of each artist, we move on grouping them on the basis of the music they produce. Since the available datasets allow us to observe the music phenomenon from three different hierarchical levels - thus, regional, national and worldwide -, we perform three different analysis levels. The first level describes the music of regional artists, the second one describes the music of both regional and national artists, and the last one describes the music of all the artists observed worldwide.

[1]Spotify Audio Features Object, `https://developer.spotify.com/web-api/get-several-audio-features/`



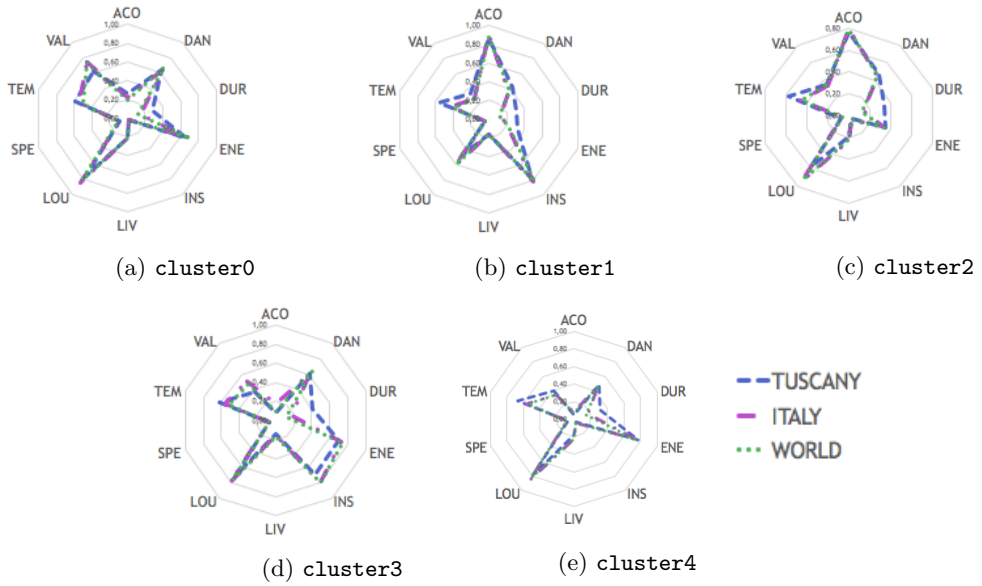
(a) k=4 (b) k=5 (c) k=6

(d) k=4 (e) k=5 (f) k=6

Figure 9.2: Matrices clusters coverage when migrating from `ITALY` to `WORLD` clusters (bottom row), and from `TUSCANY` to `ITALY` clusters (top row). From left to right the coverages for increasing values of k.

Through the analysis of the hierarchical clustering, we aim at understanding if a *fractal* structure emerges among the type of music produced at different geographical levels. We accomplish this task by employing the *K-means* clustering algorithm [218] on the computed artist profiles, as explained in Section 8.2.2. As the first step, to identify a reasonable value of k, i.e., the most appropriate number of clusters, we calculate the *Sum of Squared Error* (SSE) distribution for $k \in [2, 18]$. As shown in Figure 9.1, the SSE distributions for the clustering computed on the three levels follow a common pattern that identifies optimal values of k in the range $[4, 6]$. After that we have identified such range, we extract the clusters for each value of k in it for the three datasets. Finally, from each cluster of each level, we calculate the medoid of the cluster, i.e., the set of features describing its most representative artist.

To understand if our datasets present a fractal structure, we study artists migration among the clusters when moving from the regional to the world level. We repeat this activity for each k in the identified range. Figure 9.2 shows the clusters coverage - for k in range $[4, 6]$ - when migrating from `ITALY` to `WORLD` clusters (bottom row), and from `TUSCANY` to `ITALY` clusters (top row). Indeed, these matrices have a strong diagonal prevalence. The same effect can be observed for all the clusters and all k-values in the range $[4, 6]$. Artists blocks vary in high percentages from a down-geographical cluster to a top-geographical cluster. Matrices have been calculated by matching the pairs of clusters of two different levels with the highest level of coverage, i.e., by maximising the purity using the cluster identifiers as a label. We can observe how, for instance, regional artists of a given cluster are



(a) `cluster0` (b) `cluster1` (c) `cluster2`

(d) `cluster3` (e) `cluster4`

Figure 9.3: Artists Profiles: `TUSCANY, ITALY, WORLD` medoids. From left to right: `cluster0, cluster1, cluster2, cluster3, cluster4`

re-classified into a cohesive block of the national level that, in turn, is re-classified in a single block of the world level.

Since a fractal structure emerges with all the adopted values of k in the following, we detail the clustering obtained only for $k = 5$. Moreover, music genres distribution varies within each of the three datasets, consequently affecting their clustering. Thus, for each level, entire musical sub-genres fall into specific clusters. To better analyse and understand the characteristics of these clusters, we compare the clusters for each dataset by taking advantage of the radar chart in Figure 9.3 computed for $k = 5$ - we obtain comparable results for $k = 4$ and $k = 6$. They describe the medoids of the five clusters identified for each dataset. Axes of radar charts represents the musical features already presented in Section 8.1.2. Radar charts underline, one more time, the presence of a fractal structure capturing very similar profiles across the observed hierarchy levels. This demonstrates that, for each different level of observation, musically homogeneous artists are well clustered with artists of the superior step. First of all, we notice how the fractal structure is perfectly highlighted also by the radar charts. The spikes of the three different datasets follow the same shape almost for each cluster. At first glance of Figure 9.3, we can observe that some features are more discriminant than others. Features like *speechiness, liveness, loudness*, and *tempo* present similar values in each cluster. Despite this, the other features are determinants for cluster discrimination. Despite some little discrepancy among datasets, we can group clusters by their similarities.

We perceive that `cluster0, cluster3`, and `cluster4` are expressions of artists in direct opposition to the ones respectively in `cluster1` and `cluster2`:

- `cluster1` and `cluster2` represent melodiously, bit danceable without a strong beat, and negative songs. Clusters diverge only for instrumental scores: `cluster1` present low values, on contrary `cluster2` show high values.

- `cluster0`, `cluster3`, and `cluster4` represent non-melodic, strongly rhythmic and fairly dancing tracks. `cluster0` and `cluster4` show very low instrumental values, while `cluster3` show high scores. Furthermore, clusters differ for valence values: `cluster0` presents the highest values. On the contrary, `cluster4` presents the lowest values.

Since a fractal structure able to relate different geographical levels also emerges on the other dimensions, in the following Sections we analyse the *popularity, followers, genres* and *sentiment* of the clusters and we will detail only the results observed at the worldwide level.

9.3 Genres, Popularity and Followers

In this section, we analyse the collected information regarding artist popularity and followers as well as song genres. Starting from a dataset-wide discussion, we detail how such dimensions can be used to characterise the identified clusters. Figure 9.4 (a) shows the overall distributions of both artists and followers concerning the 12 genres identified after the cleaning stage. Indeed, the most represented genres are rock, electronic and pop, meta-classes. As was foreseeable, these genres also attract a considerable number of followers. However, it is interesting noticing that, despite rock ranks a first among the most played genres, it presents a smaller number of followers than pop music.

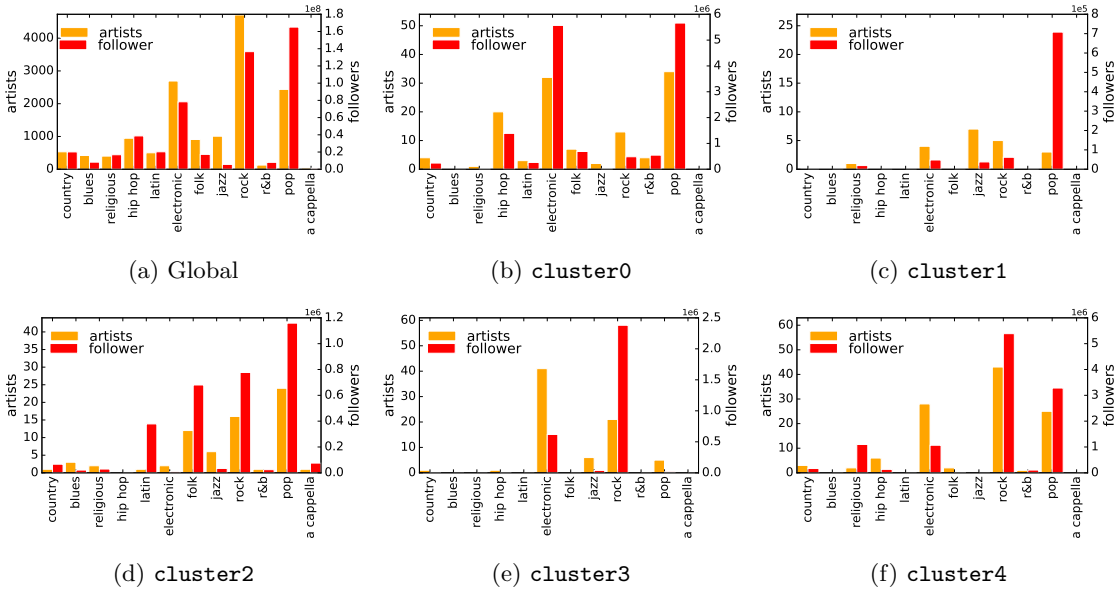
Figure 9.5 (a) shows a relation between artists popularity and followers. Starting from such plot three sub-classes of artists can be identified:

- *Low popularity.* A large number of artists present low popularity, between 0 and 40, and a small number of followers. A large number of artists are followed by few people. Such artists could be emerging artists or are likely to play niche music.

- *Medium-high popularity.* Artists have medium popularity, between 40 and 70, are followed by a consistent number of users.

- *High popularity.* Artists have very high popularity that has a relatively low number of followers.

While the former two classes were somehow expected, the latter one breaks the common intuition that expects very popular artists to be the ones attracting more followers. Once observed the general behaviours of popularity, followers and genre components, we study how they relate to the clustering we obtained in Section 9.2. As a first step, to provide a semantic annotation of the identified clusters, we describe them exploiting their main genres as well as their profiles. As we can see from Figure 9.4, clusters are strongly heterogeneous, since they represent different music genres. Indeed, the clusters



(a) Global (b) `cluster0` (c) `cluster1`

(d) `cluster2` (e) `cluster3` (f) `cluster4`

Figure 9.4: Artists and Followers distribution among clusters.

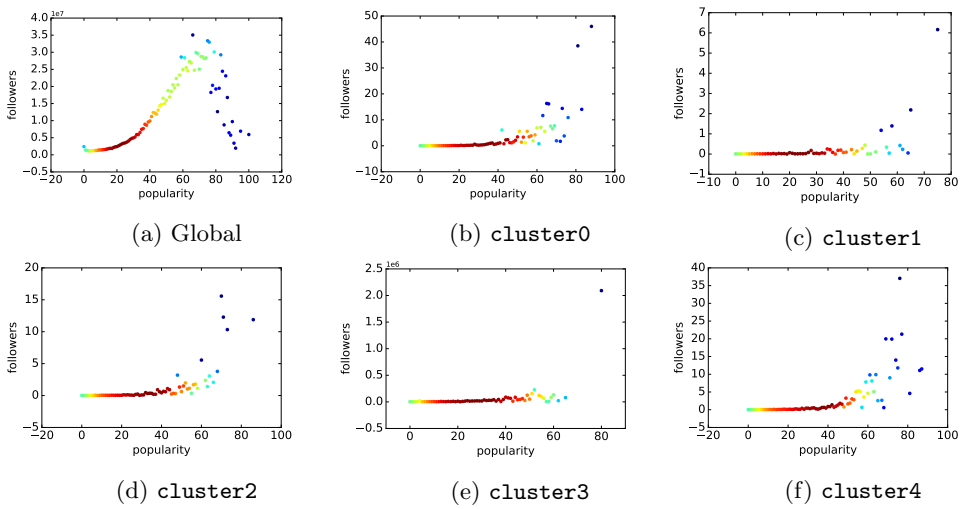
obtained are markedly different among each other, and each cluster distinctly identifies a subset of genres with specific levels of popularity and number of followers. In the following, we describe clusters as well as their main characteristics.

- `cluster0`, Figure 9.4 (b), can be identified as the *pop/hip hop* cluster. It is represented by artists such as Laura Pausini and Vanilla Sky. In this cluster fall a few genres. However, musicians that belong to this profile are followed by the highest amount of users with respect to all clusters. Songs are very suitable for dancing, repetitive, cheerful and singing. In this cluster also fall all few rap artists, but probably, they have little influence on medoids values. However, `speechiness` values are still the highest of the dataset.

- `cluster1`, Figure 9.4 (c), is the *jazz* cluster. The majority of the artists belonging to this cluster, like Stefano Bollani and Doctor Dixie Jazz Band, have few followers. The beat pace and high acoustics make tracks unsuitable for dancing and influence the valence values that are the lowest of all other clusters.

- `cluster2`, Figure 9.4 (d), is the *folk* music cluster. These genres are the most represented by the artist. In this cluster falls artists like Norah Jones and Lucio Battisti. Tracks are perceived as calm, unsuitable for dancing, primarily vocal and sad. As a consequence, valence scores are fairly negative.

- `cluster3`, Figure 9.4 (e), is the *electronic* cluster and it is represented by artists such as Calvin Harris and Go!Zilla. The most representative genre is electronic music. However, users that



(a) Global (b) `cluster0` (c) `cluster1`

(d) `cluster2` (e) `cluster3` (f) `cluster4`

Figure 9.5: Artists and Followers distribution among clusters.

follow this genre are a few compared by the rock ones. This cluster represents non-acoustic, strongly danceable with a strong beat, like dance, house, and minimal music.

- `cluster4`, Figure 9.4 (f), is the *pop/rock* cluster and in this cluster falls artists like U2 and Green Day. The number of rock and pop followers is moderately high, while the amount of electronic music is quite low. Songs are strongly rhythmic, noisy and energetic but slightly danceable. Often, tracks are perceived as angry, so valences are tendentially negatives.

As the final step, we study for each cluster the relations between its artists popularity and their number of followers. Figure 9.5 shows the relation between artists popularity and followers only for all five clusters. All the clusters are characterised by the same trend. Most of their artists have medium-low popularity, with scores between 20 and 40, and are followed by a low amount of users. Moreover, as popularity also grows followers increase. However, musicians having both high popularity, over 70, and a high number of followers are few and almost uniformly spread across all the clusters. Moreover, it is interesting to observe that in each cluster there are very few artists with a very high popularity score, over 80, that are followed by a high number of followers: supposedly, these are famous international artists.

9.4 Analysing Sentiments in Music Lyrics

As a final analysis, we are interested in observing the correlation between the songs and the feelings they transmit, framing our analysis within our clusters. We proceed by adopting a lexicon-based approach, exploiting ANEW [37] as seed-lexicon resource. ANEW provides a set of emotional ratings for 1,034 words in terms of *valence, arousal,* and *dominance*. To determinate artists polarity, we

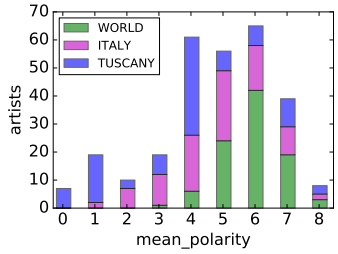


Figure 9.6: Artists' polarity score distribution among polarity class.

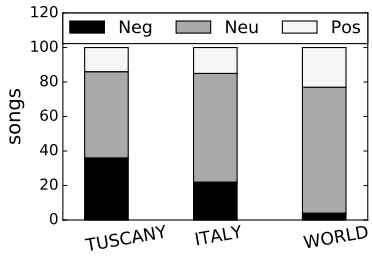


Figure 9.7: Artists' polarity score distribution among datasets.

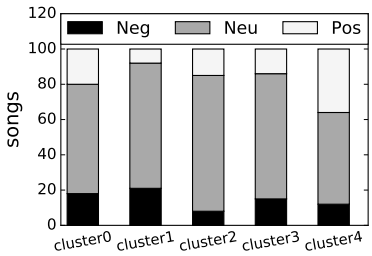


Figure 9.8: Artists' polarity score distribution among clusters.

select the valence values provided by both male and female subjects. After calculating the weighted average of the words valences v_{text} as polarity score, we grouped songs by artists, to obtain a polarity value for each of them. For each artist, we compute the weighted average among their tracks polarity. To enhance the differences among the levels of the various scores we apply a logistic function[230]:

$$f(x) = \frac{L}{1 + e^{k(xx_0)}} \tag{9.1}$$

where L is the curve maximum value (we set it equal to 1), $k = 10$ is the steepness of the curve, $x_0 = 0.5$ is the x-value of the sigmoids midpoint, and $x = v_{text}$. We apply the already mentioned procedure to each lyrics dataset. Then we grouped emotionally-tagged tracks for each artist and computed the weighted average among the i-artists discography. To analyse texts belonging to `ITALY-lyric` and `TUSCANY-lyric`, we translated the ANEW lexicon in Italian by using the Python library Goslate[2]. As a result, we obtain a comprehensive list of artists each one of them having a polarity score in the range $[0, 10]$. Studying the polarity distributions, we note that most artists have a polarity score in the range $[5, 7]$, as showed in Figure 9.6. It is necessary to keep in mind that a *neutral* score could indicate that *a*) the artists tracks transmit no strong emotions or, *b*) they present conflicting emotions. To give a better comprehension of the results, we split artists based on the polarity score into three class: *a*) *positive*, scores higher than 6; *b*) *negative*, scores lower than 4, and *c*) *neutral*, scores between 4 and 6. Figure 9.7 shows the distribution of the artists polarity among the three datasets using ANEW. Finally, we analyse how the clusters are affected by the polarity scores. Indeed, we apply the same procedure described above for each of the five clusters. The other clusters are described by a Gaussian curve similar to the one expressed by the complete dataset. All such distributions are multi-modal and define three peaks. The remaining artists reach more extreme polarity scores, with a long tail to negative values. Going further into the analysis, we split the artists based on their scores into the three polarity classes identified above. Figure 9.8 shows the

[2]Goslate: `http://pythonhosted.org/goslate/`

distribution of the artists polarity among the three clusters using ANEW. As we can observe, the most represented category remains the *neutral* one, while the less represented is the *negative* one.

9.5 Conclusions

By relying on a composite dataset built upon heterogeneous online resources, we have proposed a data-driven investigation of the music scene. We compared song technical features, lyrics, and artist popularity across three hierarchical geographic layers (world, national, and regional). Our investigation reveals the existence of a stable clustering structure describing cross-genre music profiles. Such clusters describe a fractal structure in which, disregarding the geographical granularity observed, all the artists observed can be profiled and categorised in a reduced and well-defined set of clusters.

Results we obtained led us to investigate the music context deeply. In particular, we are interested in investigating lexical features also and in observing the relationship between melodic and lexical features. Furthermore, due to the lack of works, we decided to focus our data-driven investigation on the Italian music-specific domain. In the following chapter, characteristics of Italian music have been analysed exploiting the geographical and emotional dimensions, through melodic and lexical features, to evaluate the Italian music superdiversity.

Chapter 10

The Italian Music Superdiversity

Starting from the observations described in the previous two chapters, we decide to develop a data-driven investigation of the Italian music-specific domain. The built-in-house dataset exploits heterogeneous online resources. Peculiarities of Italian music have been analysed leveraging the geographical and emotional dimensions, through melodic and lexical features, to evaluate the Italian music superdiversity. Our analysis aims to examine the relationship between melodic and lexical features. To build the cross-service datasets, we aggregate information collected from different online resources, such as Spotify, Sound-Cloud[1], Echonest[2], and Wikipedia. As mentioned, through the data, we inspect a particular context, namely the Italian music scene. The main factors able to identify homogeneous groups of individuals are spatial and linguistic. These dimensions vary in a *diachronic* sense, and their changes can be observed over time. Regarding the language, observations of *superdiversification* have led to consider the notion of language as the set of *trans-idiomatic* practices, in order to describe the communicative practices in particular places and situations. In particular, we focused on some questions. Are we observing a growing standardisation of the specific Italian music contents? Are there lexical and melodic-specific features able to discriminate the music from a geographical, lexical and emotional point of view? As is well-known, emotions can also be induced by music. If emotions may be caused by music, what is the weight of melodies and lyrics in this process? To answer these questions, we develop two different kinds of musical profiles through a melodic and a lexical approach. In particular, regarding this latter point of view, we exploit the sentiment spreading epidemic model proposed in Chapter 6. By choosing to analyse this model on Italian music lyrics, we aim to evaluate Italian music superdiversity since the resulting dictionary is strongly population-dependent and can provide important insight into how language is used by various populations. Furthermore, we aim to evaluate the model performance on long texts instead

[1]Sound-Cloud site: `https://soundcloud.com/`
[2]The Echonest service is now inactive, with the URL resulting in a 404 response.

Lexicon	#Lemmas	#Italian lemmas	#Balanced lemmas
ANEW	1,034	1,034	1,034
SentiWordNet	117,659	45,882	3,932
Bad words	550	500	500

Table 10.1: Lexicons statistics

of short texts.

The rest of the chapter is partitioned as follows. In Section 10.1 we describe the building of our cross-service composite dataset, the lexical resources we employ and the preprocessing steps performed to make manageable the data gathered from different online resources. At the end of the section, we introduce a brief discussion on the problems encountered during the data mining and the preprocessing phases. In Section 10.2 we show the Italian regional profiles identified applying an unsupervised learning strategy and a method for enhancing lexicon-based sentiment analysis by extending a base lexicon of terms. In Section 10.3 we evaluate the obtained profiles by applying the two different procedures, and we discuss the relationships and peculiarities between the lexical and melodic profiles. In addition to this, we show a sentiment analysis experiment.

10.1 Data, Resources and Preprocessing

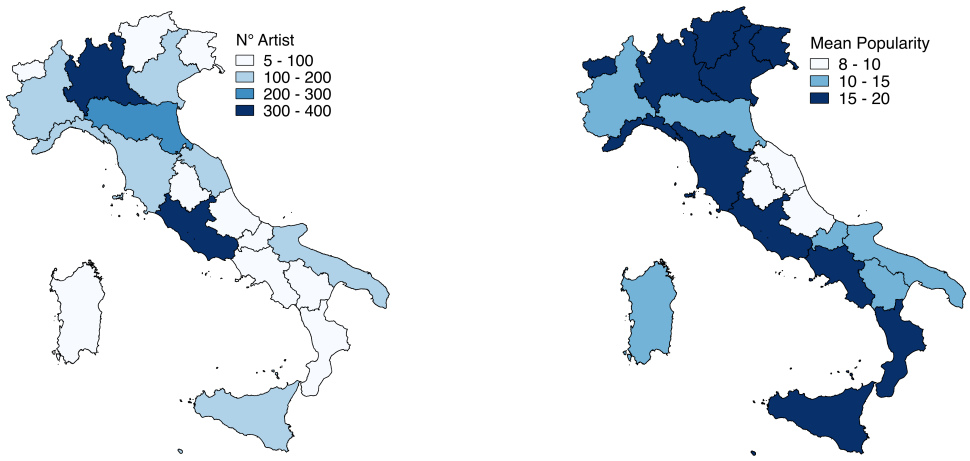
To obtain a wide overview of the current Italian music scene, we exploit a musical dataset showing two different levels of spatial granularity; thus *national* and *regional*. These two datasets refer to *Italian* musicians - both famous and less famous -, and *emerging* youth bands in Tuscany, respectively. Dataset information is provided in Section 8.1.1.

Moreover, since we are interested in musical data also from a geographical point of view, we integrate this information with a geographical field. Finally, we describe each track through a set of musical features[3] provided by Spotify - see Section 8.1.2.

To employ the sentiment spreading algorithm and validate the obtained results, we exploit the same resources used for developing the algorithm itself, thus

- The *Affective Norms for English Words* (ANEW) [37].
- *SentiWordNet* [81];
- *Full List of Bad Words Banned by Google.*
- *PAISÀ.*

[3]Spotify Audio Features Object, `https://developer.spotify.com/web-api/get-several-audio-features/`



(a) Artists distribution over Italian regions.

(b) Region average popularity score.

Figure 10.1: `ITALY` datasets statistics.

A comprehensive description of ANEW, SentiWordNet, the Full List of Bad Words Banned by Google and PAISÀ has been already provided in Section 5.3.

Since we are interested in appling the sentiment spreading algorithm to Italian music lyrics, we translate each word in the lexical resources. We follow the approach presented in Section 7.2.1. Details about lexicons obtained after the translation are reported in Table 10.1.

During the preprocessing phase, we address some problems, such as the lack of genres, the genre sparsity and, finally, the lyrics noise. Details regarding each performed step are provided in Section 8.1.3.

Once we performed the entire preprocessing phase, we obtain cleaned lyrics composed by only significant words from the sentiment points of view.

To allow experiment reproducibility, the final dataset is made publicly available on *Zenodo* at `https://bit.ly/2MUUwEx`.

10.1.1 Problems Faced

Mixing different resources referred to a unique specific domain is not such an easy task, and this is harder when starting with user-generated contents. In addition to the difficulty in obtaining the data, processing this information presents a problem. On one side, the non-standardised data variety does not allow to use standard rules. From the other side, a standard pipeline is not effective with domain-specific data. As explained before, data requires a dual-phase cleanup: an "ad hoc" supervised step to identify particular lexical forms attested, and then, an automatic standard cleaning phase. Another problem encountered is the difficulty of correct identification of the Italian authors in the international music scene. To curtail this problem, we explore only the Italian Wikipedia version of the platform. However, in musical APIs, many Italian musicians are unknown or have residual

positions. Moreover, due to the commercial orientation of the APIs, less known artists are at the bottom of their lists and this can lead to songs misallocation. A preferential attachment phenomenon may be noted: the more famous an author is, the more likely a searched information is credited to him and vice-versa.

10.2 Italian Regional Profiles

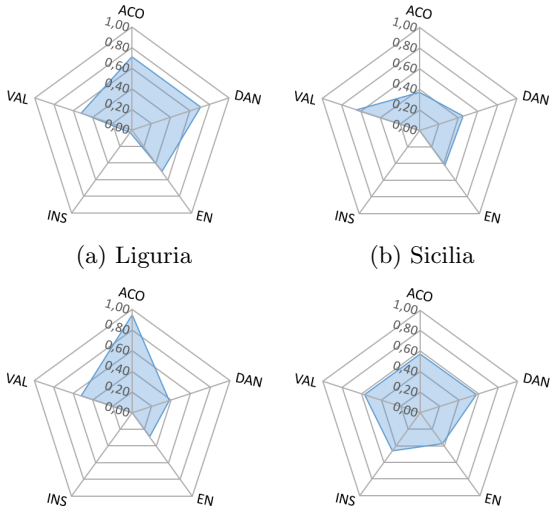
In this section, we describe the novel application of the algorithm presented in Chapter 6 to Italian musical lyrics, and our approach targeted to compute and evaluate Italian music regional profiles.

10.2.1 Regional Profiles: The Sound Point of View

The music is one of the most ancient cultural and national expression. It can also be said that is one of the most valuable expressions of national identities, since, through its performer, the music tells the history, events, and transformations of cultures and nations. However, nowadays it is particularly rare that a country can be described through a singular "cultural" entity, like a singular music genre. Conversely, as already underlined in previous chapters a national music scene is characterised by different aspects. Indeed, regarding the Italian music scene, each region can be characterised by the music it produces.

To obtain Italian regional profiles, we first exploit geographical information to partition the `ITALY` dataset into different subsets. Using the artists' region of provenance, for each Italian region, we build a regional Italian lyrics dataset. Figure 10.1 shows regional dataset statistics. In particular, Figure 10.1a shows the artists distribution among Italian regions while Figure 10.1b shows the regions average popularity. This latter score is computed through two steps. At the first step, for each artist, we calculate the average popularity over their full discography by using the popularity scores provided by Spotify. Then, we aggregated artists based on their regional provenience and, finally, we calculate each average regional popularity. As it can be noted by comparing Figures 10.1a and 10.1b, artists distribution and popularity are not correlated. Indeed, the number of artists belonging to a region does not affect the region popularity score. Regions having a relatively low number of artists, such as Calabria and Campania show a high popularity score.

To compute regional profiles, as a first step, songs in the `ITALY` dataset are grouped by artist and, for each of them, we extract a profile, as the medoid song computed over the entire artist' discography. A medoid song is not a real song attested in the artist discography, but a "sample song" which combines the main artist's musical characteristics. Indeed, we extract each medoid song as the most artist's representative song identified by minimising the sum of the Euclidean distances between the artist's tracks' musical features gathered from Spotify. Given the results obtained in Chapter 8.2 and 9, the medoids computation does not take into account all the ten musical features obtained



(a) Liguria (b) Sicilia

(c) Basilicata (d) Marche

Figure 10.2: The four Italian "super-profiles" represented by (a) Liguria, (b) Sicilia, (c) Basilicata, and (d) Marche.

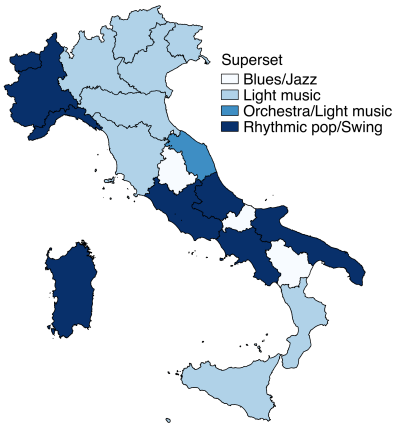


Figure 10.3: Melodic profiles.

from Spotify. Indeed, previous results display that features like *speechiness, liveness, loudness,* and *tempo* present similar values in each type of data aggregation and comparison, while other features are conversely high discriminant. Following a relevant music parameters selection phase similar to the one presented in [100], when computing the artists' medoid song we select only some features: *acousticness*, *energy*, *danceability*, and *instrumentalness*.

Once we profiled all the Italian artists, we focused our attention on the regional level by aggregating artists' medoids based on their geographical provenience. Following the approach previously explained, we extract a profile for each region, as the most representative artist among all those who come from the selected region. Our results suggest on one side that each Italian region has its music peculiarities, and on the other side that regions share several common characteristics. This is more interesting if we consider that these joint features are visible independently from the regions' geographical locations.

If we focused our attention on similarities among them, we can group regional melodic profiles based on their analogies. This brings us to identify four "super-profiles" (Figure 10.3) composed as follow:

a) Abruzzo, Valle D'Aosta, Campania, Lazio, Liguria, Piemonte, Puglia, and Sardegna.

b) Emilia Romagna, Calabria, Lombardia, Sicilia, Trentino Alto Adige, Veneto, Toscana, and Friuli Venezia Giulia.

c) Basilicata, Molise, and Umbria.

d) Marche.

We underline that these clusters are the result of the obtained melodic profiles aggregation. Therefore, we cannot find in these clusters genres shown before, as those of Figure 8.1.

Figure 10.2 shows the profiles obtained (one for each super-profile identified). Each axis in radar graphs represents one of the musical features gathered from Spotify and described in Section 8.1.2. Super-profiles are computed based on region instead of genres; therefore, they contain a vast and heterogeneous range of kind of music. However, by observing the regional profiles characteristics, each super-profile can be described through sufficiently specific music genres. For more in-depth comprehension, we describe each super-profile and some of their representative artists.

a) *Rhythmic pop/Swing*: is described by upbeat pop and the most rhythmic easy listening music, and by the swing and ska. Tracks have a good beat strength and claim the listener to move, like songs played by Max Gazze (Lazio), Fred Buscaglione (Piemonte) and 99 Posse (Campania).

b) *Light music*: is the super-set characterised by the most relaxing part of pop, by the light music and by tracks played by singer-songwriters. Songs have a slow rhythm and are suitable only for slow dance. Artists that fall into this category can be recognised in Sergio Cammariere (Calabria), Carmen Consoli (Sicilia) and Laura Pausini (Emilia Romagna).

c) *Blues/Jazz*: despite the low number of Italian region that falls into this super-set, the profile is well characterised. Indeed, it is represented by the most "commercial" portion of the jazz and blues music. In particular, in this set, we found the few Italian *croones*[4], like Fred Bongusto (Basilicata).

d) *Orchestra/Light music/Light pop*: this is a separate super-set composed by only a region. Indeed, this region show mean values for each musical feature. Probably artists coming from this region are well balanced from the genres point of view. As the most famous representative artist who comes from Marche we can found only Jimmy Fontana.

To understand if emerging bands are characterised as well as the not emerging artists of the same region, we compute a melodic profile also for them. Following the method applied to not emerging Italian artists, we group the `TUSCANY` dataset by artist and, for each of them, we extract a profile. Once profiled all emerging bands, we compute the regional profile as the most representative artist among all the emerging bands. The obtained profile is shown in Figure 10.4a in comparison with the profile computed for the Toscana region - Figure 10.4b. As can be seen, the two profiles regarding

[4] "Crooner" is an American term given to male singers of jazz standards, accompanied by either a full orchestra, a big band or a piano. The most famous America crooner is Frank Sinatra, even though he does not consider himself a crooner.

(a) TUSCANY (b) Toscana

Figure 10.4: Comparison between Tuscany emerging youth bands and the Toscana profile computer for not emerging artists.

Toscana are perfectly aligned. Our results show that emerging bands and famous artists present similar musical characteristics. This led us to affirm that youth bands tend to mimic characteristics of known artists, probably to receive more attention from the public instead of pursuing their musical way and risk to fail.

10.2.2 Regional Profiles: The Lexical Point of View

Once we identified and inspected prototypical types of regional music based on artists' musical features, we move on the lexical and emotional content of the songs. Indeed, we aim to investigate if using two completely different approaches it is possible to identify consistent and correlated regional profiles. To compile regional lexical profiles, we choose to apply the algorithm presented in Chapter 6.1 to the `ITALY-lyric` dataset.

Following the sentiment spreading algorithm method, we build a network of lemmas for each regional dataset, where each lemma corresponds to one node. Hence, the network is an unweighted co-occurrence graph based on the target lyrics to be classified. Once each regional network of lemmas is obtained, valences are added to each node in the network.

To be able to evaluate obtained dictionaries, we use cross-validation on the translated ANEW dictionary. In particular, the ANEW dictionary is randomly split into two halves. One half is used as a seed dictionary during the spreading epidemic process, merged with SentiWordNet and Bad words translated dictionaries. The second half is used as a test dataset, to compare the valence obtained through the algorithm to the original valence in the ANEW dictionary. The Pearson correlation is then used to quantify the similarity and to obtain an indication of whether the process produces valid sentiment valences. As explained before, the algorithm requires two parameters - range and entropy -, since agents can influence their neighbours as a consensual group rather than isolated; hence a heterogeneous group will not influence its neighbours. To obtain optimal values for entropy and range thresholds, ten runs are repeated for each couple of thresholds values. In this case, even though we

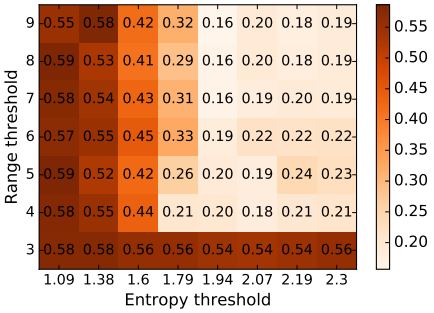


Figure 10.5: Avg correlation between modelled and real word valence (`ITALY-lyric` dataset).

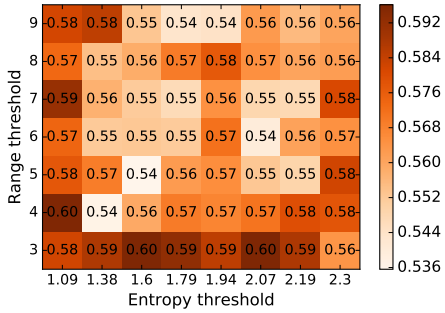


Figure 10.6: Avg correlation between modelled and real word valence (`TUSCANY-lyric` dataset).

will compute singular regional profiles, we decide to extract optimal threshold values by computing the runs on the entire `ITALY-lyric` dataset. This choice aims to better compare the obtained results. Therefore, before applying the method to Italian regions, we extract the optimal range and entropy values from the entire `ITALY-lyric` dataset. We applied the same procedure to compare the `TUSCANY-lyric` dataset with `ITALY-lyric` dataset, and once again we extract the optimal range and entropy values for the `TUSCANY-lyric` dataset. For transparency and not for comparison, Figure 10.5 and Figure 10.6, show the average correlation obtained for each threshold combination for both the `ITALY-lyric` and `TUSCANY-lyric` datasets respectively. As already observed by the algorithm, the range parameter is more important in obtaining higher correlations. Even then, small ranges resulting in better results. The optimal performance of the `ITALY-lyric` dataset is obtained for a range threshold of 5, and the entropy threshold of 1.09. Indeed, we consider the distribution described by ten bins of equal size. Hence the maximum entropy obtainable is approximately 2.3. For the `TUSCANY-lyric` dataset, the optimal performance is obtained for a range threshold of 3, and the entropy threshold of 1.6. Note that is this case, range and entropy threshold values differ less. This result is due to the small number of lyrics in the dataset.

10.3 Regional Profiles Evaluation

To evaluate the procedure, we perform two different analyses. As the first criterion, we focus on the valences obtained after the sentiment spreading process using cross-validation on the translated ANEW dictionary. So, once extracted optimal threshold values from the entire dataset, we use them as the algorithm's parameters. Therefore, for each Italian region, we compute ten different runs. For each run, the seed dictionary, composed by the 50% of the translated ANEW together with the Italian translation of SentiWordNet and Bad words lexicon, is randomly recomputed. Following this

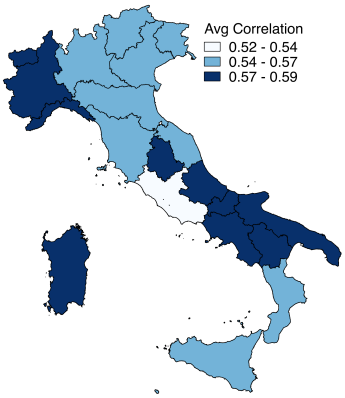
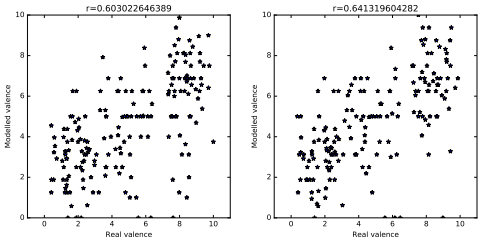
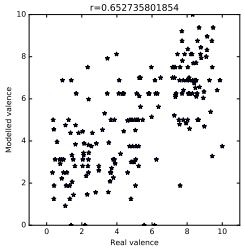


Figure 10.7: Lexical profiles based on regional average correlation over ten runs.



(a) Lazio (0.60) (b) Basilicata (0.64)



(c) `TUSCANY-lyric` (0.65)

Figure 10.8: Modelled and real word valence for a selected run with best parameters.

approach, for each region, we obtained a lexicon of words labelled with a polarity score[5].

To provide more comprehensive results, Figure 10.7 shows for each region the average correlation over ten runs, so the regional lexical profiles. Focusing our attention on closeness among their correlation values, we can group regional lexical profiles based on their analogies. This brings us to identify three "super-profiles" - Figure 10.3 - composed as follow:

a) Valle D'Aosta, Piemonte,Liguria, Sardegna, Campania, Liguria, Puglia, Basilicata, Molise, Umbria, and Sardegna.

b) Emilia Romagna, Calabria, Lombardia, Sicilia, Trentino Alto Adige, Veneto, Toscana, and Friuli Venezia Giulia and Marche.

c) Lazio.

Moreover, Figure 10.8 displays for two sample regions (Lazio and Basilicata, Figure 10.8a and 10.8b respectively) the modelled and real valences on test data for selected runs with best parameters. Within brackets are shown best correlation values for the selected run. In particular, the Lazio is the Italian region with the lowest average correlation value (0,53), while the Basilicata is one of the

[5]For example, the dictionary we obtained for Lazio for a selected run with best parameters (the same showed in Figure 10.8a) is composed of 11,243 Italian lemmas, each labelled with a polarity score in the range $[0, 10]$.

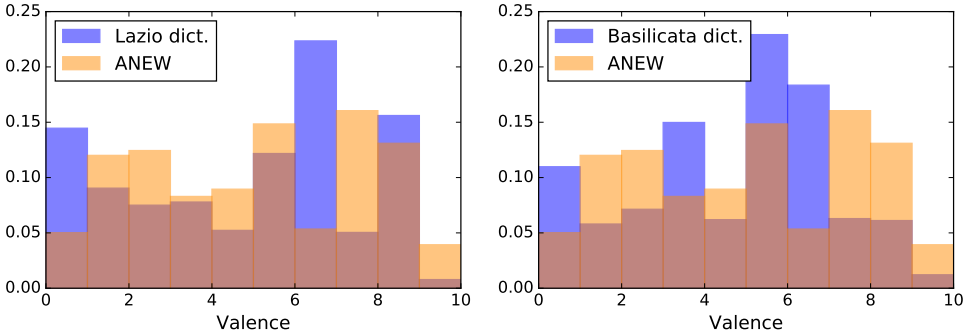


Figure 10.9: Histograms of valences for ANEW and two obtained dictionaries: Lazio (left), Basilicata (right).

regions with the highest average correlation value (0,59). The two examples plots show that also in the "worst" case - Lazio 10.8a - the valences obtained by applying the spreading epidemic sentiment algorithm to Italian music lyrics align well with human-tagged data.

Finally, Figure 10.8c shows the modelled and real valances on the `TUSCANY-lyric` dataset for a selected run with best parameters. Here too, the obtained valances align well with human tagged data. Unfortunately, the huge difference of dataset sizes between the `TUSCANY-lyric` dataset and Toscana's lyrics will not let us compare correlation results.

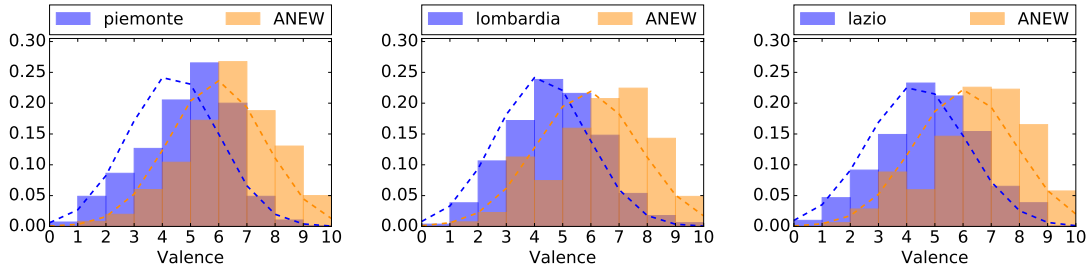
Obtained results, besides validating the original method itself, validate its usage for a different language and a different kind of initial data. As already underlined, until now, the algorithm presented in Chapter 6.1 was tested and validated only on Twitter tweets exclusively in English. Our results bear witness to the algorithm efficiency also for the Italian language and longer texts, like an entire song text. For further comparison, we also display - Figure 10.9 - the distribution of valences in the ANEW dictionary, compared with two obtained regional dictionaries related to Lazio and Basilicata. The results confirm that the spreading algorithm can identify different relations between lemmas by their usage in texts. Concerning the region, valences distributions differ widely. The only likeness is that in all regional lexicon there is a higher number of lemmas tagged with a score in the range $[0, 1]$, and, conversely, the less number of lemmas tagged in the range $[9, 10]$ than ANEW.

Once we obtained both melodic and lexical profiles, displayed in Figure 10.3 and Figure 10.7 respectively, we focus on their similarities. Starting from melodic ones, merging regional medoids, we identify four different "super-profiles" characterisable based on their genres. We obtain two major and two minor super-profiles: *Rhythmic pop/Swing* and *Light music*; *Blues/Jazz* and *Orchestra/Light music*, respectively. The two majors regional blocks are composed of 8 regions each. The *Rhythmic pop/Swing* profile includes the regions at north-west (Valle d'Aosta, Piemonte, and Liguria), the Sardegna island, and regions placed between Central Italy and south Italy, excluded Calabria, Basilicata, and Molise. The *Light music* profile is composed of a major cohesive block

including all the North and North-East regions, until the Toscana, plus two southern regions, Calabria and Sicilia. Finally, observing the two minor super-profiles, the *Blues/Jazz* profile is composed of three regions of Central Italy, while the *Orchestra/Light* music profile includes only a region of Central-East Italy. Moving to lexical profiles - Figure 10.7 - we found three "super-profiles" instead of four. However, the distribution of regions in blocks is highly related to the one previously observed. Indeed, we identify two major super-profiles, plus a minor super-profile composed of an alone region. The two major profiles enclose regions having the highest and middle average correlations, which include ten and nine regions respectively. As can be seen, regions having the highest average correlations are the same regions that fall into the *Rythmic pop/Swing*, together with regions of the *Blues/Jazz* profile. From the other side, regions having the middle average correlations are the same that fall in the Light music profile together with Marche. Alongside this allocation, Lazio shows the lowest average correlation, going out of the relative cohesive block. It is probably caused by the high number of very heterogeneous artists or by a different usage of the language by its artists.

In light of these considerations, we observe that using different features over the geographical dimension leads to two similar, comparable and coherent results. In practice, through the language as a set of trans-idiomatic practices, and specific musical features, we are able to highlight discriminant characteristics that violate the regional political boundaries, reconfiguring them following the current musical communicative practices. In details, by computing lexical profiles we obtain a coarse-grained representation of the superdiversity. Motivation can be found in the lyrics' style. Indeed, it is common that in a single can be lyrics found both positive and negative parts of the text. For example, in a romantic lyrics, the main part of the text could speak about the "positive" aspects of love, while the chorus could be focused on the love's painful part. This lyrics' characteristic could lead to flattening and to standardisation of valence scores. Indeed, since the lemmas' valence scores spread among a network of words that co-occurs in texts, it is frequent that strong positive lemmas include in their networks strong negative ones, and vice-versa. On the other side, computing melodic profiles, we obtain fine-grained details.

Unfortunately, we cannot align our lexicons evaluation with one presented in Chapter 6. The leading causes are *(a)* the lack of an already existing Italian musical tagged dataset and *(b)* the high costs required to tag a larger set of Italian songs manually. Due to this, on one side we cannot train a sentiment classifier based on Support Vector Machines (SVM), on the other side, also applying another sentiment analysis method, we cannot compare our results with others already evaluated. Finally, to build a tagged dataset to use as a training dataset we do not take into account the method applied in [69]. Indeed, this latter approach exploits ANEW as seed dictionary, so we believe that an SVM performance comparison with ANEW would be misleading.



(a) Piemonte (b) Lombardia (c) Lazio

Figure 10.10: Histograms of lyrics' valences in selected regional lyrics subsets using ANEW and the related regional lexicon.

10.3.1 Music Sentiment Analysis

Due to the lack of an already tagged and evaluated Italian musical dataset, to further evaluate the goodness of obtained lexicons, we compare the behaviour of ANEW and each regional lexicon in classifying our untagged regional lyrics datasets. We choose to follow the method proposed in [69] because it is already attested to estimate the overall valence score for musical lyrics.

For each regional subset in the `ITALY-lyric` dataset, we apply the proposed approach by using both ANEW and the respective regional lexicon obtained applying the sentiment spreading epidemic model - see Section 10.2.2.

$$v_{text} = \frac{\sum_{i=1}^{n} vi \cdot fi}{\sum_{i=1}^{n} fi} \tag{10.1}$$

where vi is the ANEW average valence for the lemma i.

Figure 10.10 shows the lyrics' valences in selected regional lyrics subsets (Piemonte, Lazio, and Lombardia) using ANEW and the related regional lexicon. As can be noted, valences obtained with the related regional lexicon are aligned. In particular, histograms show a common behaviour. ANEW tends to assign scores higher of 1 or 2 points compared to the regionals' lexicons. We observe that regional lexicons tend to assign to lyrics values in the range $[4, 6]$, while ANEW in the range $[6, 8]$. These behaviours are aligned with those observed in Section 6.1.4, where ANEW tend to evaluate tweets as positive, while the obtained dictionary balance better negatives and positives classes, but tends on assign mean valences.

10.4 Conclusions

In this chapter, we have proposed a data-driven investigation of the Italian music scene. We have analysed the characteristics of Italian music by focusing on the geographical and emotional dimensions. Such dimensions have been investigated leveraging melodic and lexical profiles. Results we obtained show that melodic and lexical features lead to coherent profiles able to highlight Italian

music's peculiarities. To the best of our knowledge, we have presented the first work on the Italian music scene focused on geographical and emotional dimensions. Moreover, our investigation represents the first attempt to apply the Superdiversity theory to a specific context, such as the musical one. Finally, the dataset we built by exploiting heterogeneous online resources is the first free available annotated Italian music dataset.

Chapter 11

Conclusions

In this thesis, we have presented the first applied investigation on Superdiversity. We have proposed to shift from the Superdiversity theory of Vertovec to a real measurement of it by using the first Superdiversity Index. We defined the Superdiversity Index as a measure of the distance between the sentiment valence of words used by a community and the standard valences from a manually tagged dictionary. Thus, it can quantify diversity in a population based on the changes in the emotional content of words, compared to the standard language. We estimated this change in communities from the UK and Italy, using geolocalised Twitter data in English and Italian, at various geographical resolutions. In almost all of the geographical regions analysed we observed a remarkable correlation with foreign immigration rates, extracted from the dataset available from the Joint Research Center of the European Community, through the D4I data challenge. The proposed Superdiversity Index greatly outperforms other possible measures of diversity from the same Twitter data.

The index is based on a similarity measure between standard and community use of a language. To obtain this measure, we have proposed a method for enhancing lexicon-based Sentiment Analysis by extending a base lexicon of terms. The output of the algorithm consists of an emotively-labelled dictionary containing terms' valences that correlate well with human-labelled lexicons. Moreover, we have shown that the performance of the SVM-based sentiment classification was maintained and the number of tweets labelled by the new dictionary grew by about 45% in comparison with a well-established dictionary in the literature. Results have shown that the algorithm is particularly suitable for Twitter data where the short length of the text to be analysed makes classification impossible with a small dictionary since most tweets do not contain any of the terms for which the sentiment valence is known. However, the algorithm can be also applied to long texts, as demonstrated on music lyrics. Additionally, the procedure we propose to extend the dictionary is very fast since running times are of the order of minutes for over 1.5 million tweets. This characteristic suggests that the method can be applied to vast amounts of data, i.e. Big Data.

The extended dictionary we obtain is strongly domain dependent. Even if we used a random

selection of tweets, one could also select user sub-populations or various languages. An advantage of our method is that it is easily expandable to other languages since it is enough to translate the seed dictionary to obtain a much larger annotated lexicon. The correlation between modelled and real valences becomes a measure to describe the way language is used in a sub-population, allowing for comparison for various purposes.

While useful for standard Sentiment Analysis, as shown here, our method is also suitable for novel and more advanced techniques similar to supervised aggregated SA [50]. These techniques concentrate not on the classification of individual tweets, but on quantifying the aggregated distribution of sentiment in a collection of tweets. Our extended dictionary could prove to be an important resource to develop new such methods. These would be useful not only to predict election or debate results, which is why they were proposed in the first place but also to quantify aggregated sentiment in various populations. This could help understand superdiversity better.

This work paves the way for a novel nowcasting model of immigration rates, that can be applied with higher time and space resolution compared to official statistics. In future works, we plan to investigate the use of machine learning to achieve this task. We expect our Superdiversity Index to be a major feature in this model, with other features used to correct for range differences, including language entropy, local dialects, and population density.

Together with the application of the Superdiversity theory to real data, we also proposed several data-driven investigations of the music scene by relying on a dataset built upon heterogeneous online resources. Peculiarities of music have been analysed by leveraging the geographical and emotional dimensions. We analysed song technical features, lyrics, popularity, followers and genres across three hierarchical geographic layers - world, national and regional. In our opinion, it is clear that it is not easy to define general national and regional profiles of music exhaustively to describe the entire national scene because of the strong heterogeneity of music.

However, our analysis identifies the existence of a very stable clustering structure able to describe cross-genre music profiles. We highlighted how such clusters describe a fractal structure. Disregarding the geographical granularity observed, all the artists observed can be profiled and categorised in a reduced and well-defined set of clusters. Moreover, we analysed artists popularity and fan base observing how their distributions describe a similar trend in all the identified clusters. Finally, looking to the artists song lyrics we were able to observe the emotional valence of the identified meta-profiles.

Further analysis has focused on the Italian music scene to evaluate Italian Music Superdiversity. Our results show that our melodic and lexical features lead to coherent profiles. Our sentiment spreading algorithm has allowed us to highlight both the lexical and emotive Italian music's characteristics. It can be argued that being two parts of the same phenomenon, lexical and melodic features can be combined in favour of a better understanding of the analysed Italian scene. Moreover, since there are no attested publicly annotated Italian music datasets, the free availability of

our data represents an important contribution to start to fill this gap.

The results of our experiments leave open options for future developments in different directions. We believe that it could be interesting to perform a finer grained analysis on lyrics to investigate other discriminating features, i.e., linguistic that can help to describe songs sociologically and emotionally. At the same time, we also plan to perform comparisons at the international level. This analysis can be conducted both in Europe and globally to investigate the importance of music in the identification of cultures and traditions of the populations, in other words, in the worldwide musical Superdiversity. Finally, due to the sparsity of contributions regarding the Italian music panorama, we would like to consolidate the cross-domain dataset to enrich the current one.

Bibliography

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.

[2] Shaishav Agrawal et al. Using syntactic and contextual information for sentiment polarity analysis. In *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, pages 620–623. ACM, 2009.

[3] Rein Ahas, Siiri Silm, and Margus Tiru. Measuring transnational migration with roaming datasets. In Peter Kiefer, Haosheng Huang, Nico Van de Weghe, and Martin Raubal, editors, *Adjunct Proceedings of the 14th International Conference on Location Based Services*, pages 105 – 108. ETH Zurich, 2018-01-15. 14th International Conference on Location Based Services (LBS 2018); Conference Location: Zurich, Switzerland; Conference Date: January 15-17, 2018.

[4] Silvio Amir, Ramón Astudillo, Wang Ling, Bruno Martins, Mario J Silva, and Isabel Trancoso. Inesc-id: A regression model for large scale twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 613–618, 2015.

[5] Xiaoran An, Auroop R Ganguly, Yi Fang, Steven B Scyphers, Ann M Hunter, and Jennifer G Dy. Tracking climate change opinions from twitter data. In *Workshop on Data Science for Social Good*, 2014.

[6] E Andrea. *Automatic generation of lexical resources for opinion mining: Models, algorithms and applica-tions*. PhD thesis, Pisa: University dipisa. Italy, 2008.

[7] Ravi Arunachalam and Sandipan Sarkar. The new eye of government: Citizen sentiment analysis in social media. In *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 23–28, 2013.

[8] Amir Asiaee T, Mariano Tepper, Arindam Banerjee, and Guillermo Sapiro. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1602–1606. ACM, 2012.

[9] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

[10] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

[11] Akshat Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, and Vasudeva Varma. Mining sentiments from tweets. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 11–18, 2012.

[12] Alexandra Balahur and Marco Turchi. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52–60. Association for Computational Linguistics, 2012.

[13] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

[14] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 36–44. Association for Computational Linguistics, 2010.

[15] Andrea Baronchelli, Vittorio Loreto, and Francesca Tria. Language dynamics, 2012.

[16] Laurie Bauer. Inferring variation and change from public corpora. *The handbook of language variation and change*, pages 97–114, 2002.

[17] Nuria Bel, Cornelis HA Koster, and Marta Villegas. Cross-lingual text categorization. In *International Conference on Theory and Practice of Digital Libraries*, pages 126–139. Springer, 2003.

[18] Gema Bello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, 2016.

[19] Farah Benamara, Baptiste Chardon, Yannick Mathieu, and Vladimir Popescu. Towards context-based subjectivity analysis. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1180–1188, 2011.

[20] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[21] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti. *PLoS medicine*, 8(8):e1001083, 2011.

[22] Adam Bermingham and Alan F Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836. ACM, 2010.

[23] Karla Z Bertrand, Maya Bialik, Kawandeep Virdee, Andreas Gros, and Yaneer Bar-Yam. Sentiment in new york city: A high resolution spatial and temporal view. *arXiv preprint arXiv:1308.5010*, 2013.

[24] Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In *2004 AAAI spring symposium on exploring attitude and affect in text*, volume 2224, 2004.

[25] Mark A Beyer and Douglas Laney. The importance of big data: a definition. *Stamford, CT: Gartner*, pages 2014–2018, 2012.

[26] Kerstin Bischoff, Claudiu S Firan, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier, and Mohamed Sordo. Music mood and theme classification-a hybrid approach. In *ISMIR*, pages 657–662, 2009.

[27] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era*, volume 14, pages 339–348, 2008.

[28] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics, 2006.

[29] Jan Blommaert, Karel Arnaut, Ben Rampton, and Massimiliano Spotti. Language and superdiversity. 2016.

[30] Jan Blommaert and Ben Rampton. Language and superdiversity. 2012.

[31] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10, 2015.

[32] Joshua E Blumenstock. Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development*, 18(2):107–125, 2012.

[33] Marcus Böhme, André Gröger, and Tobias Stöhr. Searching for a better life: Now-casting international migration with online search keywords.

[34] Marco Bonzanini, Miguel Martinez-Alvarez, and Thomas Roelleke. Opinion summarisation through sentence extraction: An investigation with movie reviews. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1121–1122. ACM, 2012.

[35] Efthimios Bothos, Dimitris Apostolou, and Gregoris Mentzas. Using social media to predict future events with agent-based markets. *IEEE Intelligent Systems*, 25(6):50–58, 2010.

[36] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[37] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[38] Felipe Bravo-Marquez. *Acquiring and Exploiting Lexical Knowledge for Twitter Sentiment Analysis*. PhD thesis, University of Waikato, 2017.

[39] Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the conference on empirical methods in natural language processing*, pages 562–570. Association for Computational Linguistics, 2011.

[40] Julian Brooke, Milan Tofiloski, and Maite Taboada. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the international conference RANLP-2009*, pages 50–54, 2009.

[41] Rogers Brubaker. Ethnicity without groups. *European Journal of Sociology/Archives Européennes de Sociologie*, 43(2):163–189, 2002.

[42] José R Bucheli, Matías Fontenla, and Benjamin James Waddell. Return migration and violence. *World Development*, 116:113–124, 2019.

[43] Erik Cambria and Amir Hussain. Sentic computing. *Cognitive Computation*, 7(2):183–185, Apr 2015.

[44] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer, 2012.

[45] Erik Cambria, Dheeraj Rajagopal, Daniel Olsher, and Dipankar Das. Big social data analysis. *Big data computing*, 13:401–414, 2013.

[46] Erion Çano and Maurizio Morisio. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pages 118–124. ACM, 2017.

[47] Erion Çano and Maurizio Morisio. Music mood dataset creation based on last. fm tags. In *2017 International Conference on Artificial Intelligence and Applications, Vienna, Austria*, 2017.

[48] Claudio Castellano, Miguel A Muñoz, and Romualdo Pastor-Satorras. Nonlinear q-voter model. *Physical Review E*, 80(4):041129, 2009.

[49] Oscar Celma. Music recommendation. In *Music recommendation and discovery*, pages 43–85. Springer, 2010.

[50] Andrea Ceron, Luigi Curini, and S Iacus. Using social media to fore-cast electoral results: A review of state-of-the-art. *Ital J Appl Stat*, 25:237–259, 2015.

[51] Andrea Ceron, Luigi Curini, and Stefano Maria Iacus. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*, volume 9. Springer Science & Business Media, 2014.

[52] Jack K Chambers and Peter Trudgill. *Dialectology*. Cambridge University Press, 1998.

[53] Pimwadee Chaovalit and Lina Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 112c–112c. IEEE, 2005.

[54] Iti Chaturvedi, Erik Cambria, Roy E Welsch, and Francisco Herrera. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77, 2018.

[55] Lisa Chauvet and Marion Mercier. Do return migrants transfer political norms to their origin country? evidence from mali. *Journal of Comparative Economics*, 42(3):630–651, 2014.

[56] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2005.

[57] Md Faisal Mahbub Chowdhury, Marco Guerini, Sara Tonelli, and Alberto Lavelli. Fbk: Sentiment analysis in twitter with tweetsted. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 466–470, 2013.

[58] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[59] Mauro Coletto, Andrea Esuli, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016*, pages 1270–1277, 2016.

[60] Mauro Coletto, Andrea Esuli, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. Perception of social phenomena through the multidimensional analysis of online social networks. *Online Social Networks and Media*, 1:14 – 32, 2017.

[61] Mauro Coletto, Andrea Esuli, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. Perception of social phenomena through the multidimensional analysis of online social networks. *Online Social Networks and Media*, 1:14–32, 2017.

[62] Alfredo Cuzzocrea, Il-Yeol Song, and Karen C Davis. Analytics over large-scale multidimensional data: the big data revolution! In *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, pages 101–104. ACM, 2011.

[63] Kushal Dave, Steve Lawrence, and David 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*, pages 519–528. ACM, 2003.

[64] Felice DellOrletta. Ensemble system for part-of-speech tagging. *Proceedings of EVALITA*, 9:1–8, 2009.

[65] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[66] Xiaowen Ding, Bing Liu, and Philip Yu. A holistic lexicon-based approach to opinion mining. pages 231–240, 01 2008.

[67] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM, 2008.

[68] Frédéric Docquier and Hillel Rapoport. Globalization, brain drain, and development. *Journal of Economic Literature*, 50(3):681–730, 2012.

[69] Peter Sheridan Dodds and Christopher M Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of happiness studies*, 11(4):441–456, 2010.

[70] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752, 2011.

[71] Laney Douglas. 3d data management: Controlling data volume, velocity and variety. *Gartner. Retrieved*, 6(2001):6, 2001.

[72] XHJS Downie, Cyril Laurier, and MBAF Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *Proc. 9th Int. Conf. Music Inf. Retrieval*, pages 462–467, 2008.

[73] Gabriel Doyle. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106, 2014.

[74] Antoine Dubois, Emilio Zagheni, Kiran Garimella, and Ingmar Weber. Studying Migrant Assimilation Through Facebook Interests. jan 2018.

[75] Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21, 1997.

[76] Edd Dumbill. What is big data. *An introduction to the big data landscape.[online] http://strata. oreilly. com/2012/01/what-is-big-data. html*, 2012.

[77] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[78] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114, 2014.

[79] Andrea Esuli and Alejandro Moreo Fernández. Distributional correspondence indexing for cross-language text categorization. In *European Conference on Information Retrieval*, pages 104–109. Springer, 2015.

[80] Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM, 2005.

[81] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[82] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, 17:1–26, 2007.

[83] Eurostat. Migrants in Europe: A statistical portrait of the first and second generation. *Luxembourg: Publications Office of the European Union.*, 2011.

[84] Song Feng, Ritwik Bose, and Yejin Choi. Learning general connotation of words using graph-based algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1092–1103. Association for Computational Linguistics, 2011.

[85] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse: Mining customer opinions from free text. In *international symposium on intelligent data analysis*, pages 121–132. Springer, 2005.

[86] Manoochehr Ghiassi, James Skinner, and David Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.

[87] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28, 2016.

[88] Alec Go, Lei Huang, and Richa Bhayani. Twitter sentiment analysis. *Entropy*, 17:252, 2009.

[89] Bruno Gonçalves and David Sánchez. Crowdsourcing dialect characterization through twitter. *PloS one*, 9(11):e112074, 2014.

[90] Marco Guerini, Lorenzo Gatti, and Marco Turchi. Sentiment analysis: How to derive prior polarities from sentiwordnet. *arXiv preprint arXiv:1309.5843*, 2013.

[91] Marco Guerini, Carlo Strapparava, and Oliviero Stock. Valentino: A tool for valence shifting of natural language texts. In *LREC*. Citeseer, 2008.

[92] Riccardo Guidotti and Lorenzo Gabrielli. Recognizing residents and tourists with retail data using shopping profiles. In *International Conference on Smart Objects and Technologies for Social Good*, pages 353–363. Springer, 2017.

[93] Riccardo Guidotti, Lorenzo Gabrielli, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. Discovering temporal regularities in retail customers shopping behavior. *EPJ Data Science*, 7(1):6, 2018.

[94] Riccardo Guidotti, Anna Monreale, Mirco Nanni, Fosca Giannotti, and Dino Pedreschi. Clustering individual transactional data for masses of users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 195–204. ACM, 2017.

[95] Vishal Gupta, Gurpreet S Lehal, et al. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1):60–76, 2009.

[96] Robin Hanson. Foul play in information markets. 2006.

[97] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of big data on cloud computing: Review and open research issues. *Information Systems*, 47:98–115, 2015.

[98] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[99] Vasileios Hatzivassiloglou and Janyce M Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305. Association for Computational Linguistics, 2000.

[100] Patrick Helmholz, Dominik Siemon, and Susanne Robra-Bissantz. Summer hot, winter not!–seasonal influences on context-based music recommendations.

[101] Herdagdelen, Amaç, Bogdan State, Lada Adamic, and Winter Mason. The social ties of immigrant communities in the United States. In *WebSci*, 2016.

[102] David A Hollinger. *Postethnic America: beyond multiculturalism*. Hachette UK, 2006.

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

[104] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 537–546. ACM, 2013.

[105] Xiao Hu and J Stephen Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *ISMIR*, pages 67–72, 2007.

[106] Xiao Hu and J Stephen Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *ISMIR*, pages 619–624, 2010.

[107] Xiao Hu, J Stephen Downie, and Andreas F Ehmann. Lyric text mining in music mood classification. *American music*, 183(5,049):2–209, 2009.

[108] Darrell Huff. *How to lie with statistics*. WW Norton & Company, 1993.

[109] Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338, 2018.

[110] Hossam S Ibrahim, Sherif M Abdou, and Mervat Gheith. Sentiment analysis for modern standard arabic and colloquial. *arXiv preprint arXiv:1505.03105*, 2015.

[111] Philip J. Stone, Robert F. Bales, J Zvi Namenwirth, and Daniel Ogilvie. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7:484 – 498, 10 2007.

[112] Jaap Kamps, Maarten Marx, Robert J Mokken, Maarten De Rijke, et al. Using wordnet to measure semantic orientations of adjectives. In *LREC*, volume 4, pages 1115–1118. Citeseer, 2004.

[113] Farhan Hassan Khan, Saba Bashir, and Usman Qamar. Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57:245–257, 2014.

[114] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Mahmoud Ali, Waleed Kamaleldin, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal*, 2014, 2014.

[115] Vinh Ngoc Khuc, Chaitanya Shivade, Rajiv Ramnath, and Jay Ramanathan. Towards building large-scale distributed systems for twitter sentiment analysis. In *Proceedings of the 27th annual ACM symposium on applied computing*, pages 459–464. ACM, 2012.

[116] Riivo Kikas, Marlon Dumas, and Ando Saabas. Explaining international migration in the skype network: The role of social network features. In *Proceedings of the 1st ACM Workshop on Social Media World Sensors*, pages 17–22. ACM, 2015.

[117] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[118] Svetlana Kiritchenko and Saif M Mohammad. The effect of negators, modals, and degree adverbs on sentiment composition. *arXiv preprint arXiv:1712.01794*, 2017.

[119] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[120] Moshe Koppel and Jonathan Schler. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109, 2006.

[121] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*, 2011.

[122] Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Freshman or fresher? quantifying the geographic variation of language in online social media. In *ICWSM*, pages 615–618, 2016.

[123] Akshi Kumar and Teeja Mary Sebastian. Sentiment analysis on twitter. *International Journal of Computer Science Issues (IJCSI)*, 9(4):372, 2012.

[124] William Labov, Sharon Ash, and Charles Boberg. *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter, 2005.

[125] Frank Laczko. Improving Data on International Migration and Development: Towards a Global Action Plan? " Improving Data on International Migration -towards Agenda 2030 and the Global Compact on Migration ", 2015.

[126] Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, and José J Ramasco. Immigrant community integration in world cities. *PloS one*, 13(3):e0191612, 2018.

[127] Paul Lamere, Elias Pampalk, C Schmitz, JP Bello, E Chew, and D Turnbull. Social tags and music information retrieval. In *ISMIR*, page 24, 2008.

[128] Cyril Laurier, Mohamed Sordo, Joan Serra, and Perfecto Herrera. Music mood representations from social tags. In *ISMIR*, pages 381–386, 2009.

[129] Jin Ha Lee and Xiao Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 129–138. ACM, 2012.

[130] Adrienne Lehrer. Semantic fields and lexical structure. 1974.

[131] Gang Li, Rob Law, Huy Quan Vu, Jia Rong, and Xinyuan Roy Zhao. Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism management*, 46:311–321, 2015.

[132] Gang Li and Fei Liu. A clustering-based approach on sentiment analysis. In *Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on*, pages 331–337. IEEE, 2010.

[133] Tao Li and Mitsunori Ogihara. Music artist style identification by semi-supervised learning from both lyrics and content. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 364–367. ACM, 2004.

[134] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.

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

[136] Bing Liu et al. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666, 2010.

[137] Chong Long, Jie Zhang, and Xiaoyan Zhut. A review selection approach for accurate feature rating estimation. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 766–774. Association for Computational Linguistics, 2010.

[138] Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. The paisa'corpus of italian web texts. In *9th Web as Corpus Workshop (WaC-9)@ EACL 2014*, pages 36–43. EACL (European chapter of the Association for Computational Linguistics), 2014.

[139] Amr Magdy, Thanaa M Ghanem, Mashaal Musleh, and Mohamed F Mokbel. Exploiting geo-tagged tweets to understand localized language diversity. In *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data*, page 2. ACM, 2014.

[140] Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Pedro Paiva. Classification and regression of music lyrics: Emotionally-significant features. 8th International Conference on Knowledge Discovery and Information Retrieval, 2016.

[141] Lev Manovich. Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2:460–475, 2011.

[142] Marco Martiniello. How to combine integration and diversities: The challenge of an eu multicultural citizenship. 2004.

[143] Chetan Mate. Product aspect ranking using sentiment analysis: A survey. *International Research Journal of Engineering and Technology*, 3(01):126–127, 2015.

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

[145] Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284. ACM, 2009.

[146] Alice Mesnard et al. Temporary migration and self-employment: evidence from tunisia. *Brussels Economic Review*, 47(1):119–138, 2004.

[147] Johnnatan Messias, Fabricio Benevenuto, Ingmar Weber, and Emilio Zagheni. From migration corridors to clusters: The value of Google+ data for migration studies. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 421–428. IEEE, aug 2016.

[148] Alessandra Micalizzi. *Come un altro mondo. Pratiche di socializzazione dell'esperienza della perdita dentro e fuori della rete*, volume 7. Ledizioni, 2012.

[149] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 976–983, 2007.

[150] Rada Mihalcea and Carlo Strapparava. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599. Association for Computational Linguistics, 2012.

[151] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[152] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11(5th):25, 2011.

[153] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5):e64417, 2013.

[154] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981, 2013.

[155] Saif Mohammad. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[156] Saif M Mohammad. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics, 2012.

[157] Saif M Mohammad. Challenges in sentiment analysis. In *A practical guide to sentiment analysis*, pages 61–83. Springer, 2017.

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

[159] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.

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

[161] Izabela Moise, Edward Gaere, Ruben Merz, Stefan Koch, and Evangelos Pournaras. Tracking language mobility in the twitter landscape. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 663–670. IEEE, 2016.

[162] Roser Morante and Caroline Sporleder. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260, 2012.

[163] Kaitlyn Mulcrone. Detecting emotion in text. *University of Minnesota–Morris CS Senior Semminar Paper*, 2012.

[164] Diala Naboulsi, Razvan Stanica, and Marco Fiore. Classifying call profiles in large-scale mobile traffic datasets. In *INFOCOM, 2014 Proceedings IEEE*, pages 1806–1814. IEEE, 2014.

[165] Ramanathan Narayanan, Bing Liu, and Alok Choudhary. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 180–189. Association for Computational Linguistics, 2009.

[166] Sascha Narr, Michael Hulfenhaus, and Sahin Albayrak. Language-independent twitter sentiment analysis. *Knowledge discovery and machine learning (KDML), LWA*, pages 12–14, 2012.

[167] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Sentiful: Generating a reliable lexicon for sentiment analysis. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE, 2009.

[168] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593, 2016.

[169] Hieu Nguyen and Kiran Garimella. Understanding International Migration using Tensor Factorization. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, pages 829–830, New York, New York, USA, 2017. ACM Press.

[170] Reynier Ortega, Adrian Fonseca, and Andres Montoyo. Ssa-uo: unsupervised twitter sentiment analysis. In *Second joint conference on lexical and computational semantics (* SEM)*, volume 2, pages 501–507, 2013.

[171] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The measurement of meaning*. Number 47. University of Illinois press, 1957.

[172] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390, 2013.

[173] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[174] Bo Pang and Lillian 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*, page 271. Association for Computational Linguistics, 2004.

[175] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

[176] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[177] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[178] Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. Emoticon style: Interpreting differences in emoticons across cultures. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[179] Priyanka Patil and Pratibha Yalagi. Sentiment analysis levels and techniques: A survey. *space*, 1:6, 2016.

[180] S Perna, A Maisto, P Vitale, and R Guarasci. Il linguaggio del rap. possibilità di unanalisi multidisciplinare. *XXVI Convegno internazionale ass. i. term. Terminologia e organizzazione della conoscenza nella conservazione della memoria digitale*, 34:209–217, 2016.

[181] Andrew Perrin. Social media usage. *Pew research center*, pages 52–68, 2015.

[182] John C Platt, Kristina Toutanova, and Wen-tau Yih. Translingual document representations from discriminative projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 251–261. Association for Computational Linguistics, 2010.

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

[184] Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. Do all birds tweet the same?: characterizing twitter around the world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1025–1030. ACM, 2011.

[185] Raluca-Elena Podicu, Diana Gratie, and Octavian Voicu. Inferring song moods from lyrics.

[186] Laura Pollacci, Riccardo Guidotti, and Giulio Rossetti. Are we playing like music-stars? placing emerging artists on the italian music scene. In *9th international workshop on machine learning and music*, 2016.

[187] Laura Pollacci, Riccardo Guidotti, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. The fractal dimension of music: geography, popularity and sentiment analysis. In *International conference on smart objects and technologies for social good*, pages 183–194. Springer, 2017.

[188] Laura Pollacci, Riccardo Guidotti, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. The italian music superdiversity. *Multimedia Tools and Applications*, pages 1–23, 2018.

[189] Laura Pollacci, Alina Sîrbu, Fosca Giannotti, Dino Pedreschi, Claudio Lucchese, and Cristina I. Muntean. Sentiment spreading: an epidemic model for lexicon-based sentiment analysis on twitter. In *Conference of the Italian Association for Artificial Intelligence*, pages 114–127. Springer, 2017.

[190] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[191] Peter Prettenhofer and Benno Stein. Cross-lingual adaptation using structural correspondence learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(1):13, 2011.

[192] Allan Puur and Luule Sakkeus. International migration and the demographic challenges of Europe. *Estonian Human Development Report 2016/2017, "Estonia at the Age of Migration"*., 2017.

[193] Zhenxin Qin. *A framework and practical implementation for sentiment analysis and aspect exploration*. PhD thesis, University of Manchester, 2017.

[194] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[195] David Rawlings and Vera Ciancarelli. Music preference and the five-factor model of the neo personality inventory. *Psychology of Music*, 25(2):120–132, 1997.

[196] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, pages 43–48. Association for Computational Linguistics, 2005.

[197] Hilke Reckman, Cheyanne Baird, Jean Crawford, Richard Crowell, Linnea Micciulla, Saratendu Sethi, and Fruzsina Veress. teragram: Rule-based detection of sentiment phrases using sas sentiment analysis. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 513–519, 2013.

[198] Peter J Rentfrow and Samuel D Gosling. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236, 2003.

[199] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017.

[200] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[201] Hassan Saif, Yulan He, and Harith Alani. Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org), 2012.

[202] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *International semantic web conference*, pages 508–524. Springer, 2012.

[203] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*, 52(1):5–19, 2016.

[204] Markus Schedl, Nicola Orio, Cynthia Liem, and Geoffroy Peeters. A professionally annotated and enriched multimodal data set on popular music. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 78–83. ACM, 2013.

[205] Helmut Schmid. Part-of-speech tagging with neural networks. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 172–176. Association for Computational Linguistics, 1994.

[206] Gayane Shalunts, Gerhard Backfried, and Nicolas Commeignes. The impact of machine translation on sentiment analysis. *DATA ANALYTICS 2016*, page 63, 2016.

[207] David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, pages 3–10. ACM, 2009.

[208] Vikas Sindhwani and Prem Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *2008 Eighth IEEE International Conference on Data Mining*, pages 1025–1030. IEEE, 2008.

[209] Alina Sîrbu, Vittorio Loreto, Vito DP Servedio, and Francesca Tria. Opinion dynamics: models, extensions and external effects. In *Participatory sensing, opinions and collective awareness*, pages 363–401. Springer, 2017.

[210] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[211] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics, 2011.

[212] Spredfast. Displaying status entities.

[213] Stephen V Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89, 1997.

[214] Josef Steinberger, Tomáš Brychcín, and Michal Konkol. Aspect-level sentiment analysis in czech. In *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 24–30, 2014.

[215] Xiaodan Zhu Svetlana Kiritchenko and Saif M. Mohammad. Sentiment analysis of short informal texts. 50:723–762.

[216] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[217] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of twitter networks. *Social networks*, 34(1):73–81, 2012.

[218] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 2013.

[219] Mildred C Templin. Certain language skills in children; their development and interrelationships. 1957.

[220] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[221] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[222] Elizabeth Thomas-Hope. Return migration to jamaica and its development potential. *International migration*, 37(1):183–207, 1999.

[223] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, volume 8, pages 325–330, 2008.

[224] Angela Charng-Rurng Tsai, Chi-En Wu, Richard Tzong-Han Tsai, and Jane Yung-jen Hsu. Building a concept-level sentiment dictionary based on commonsense knowledge. *IEEE Intelligent Systems*, 28(2):22–30, 2013.

[225] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[226] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

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

[228] Silvia Vázquez, Óscar Muñoz-García, Inés Campanella, Marc Poch, Beatriz Fisas, Nuria Bel, and Gloria Andreu. A classification of user-generated content into consumer decision journey stages. *Neural Networks*, 58:68–81, 2014.

[229] Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan 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*, pages 777–785. Association for Computational Linguistics, 2010.

[230] PF Verhulst. Note sur la loi daccroissement de la population. *Bulletins de lAcadémie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, 13:226–227, 1846.

[231] Steven Vertovec. Conceiving and researching transnationalism. *Ethnic and racial studies*, 22(2):447–462, 1999.

[232] Steven Vertovec. Super-diversity and its implications. *Ethnic and racial studies*, 30(6):1024–1054, 2007.

[233] Steven Vertovec. Towards post-multiculturalism? changing communities, conditions and contexts of diversity. *International social science journal*, 61(199):83–95, 2010.

[234] Luis Von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: a game for collecting commonsense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78. ACM, 2006.

[235] Jackline Wahba. Selection, selection, selection: the impact of return migration. *Journal of Population Economics*, 28(3):535–563, 2015.

[236] Jackline Wahba and Yves Zenou. Out of sight, out of mind: Migration, entrepreneurship and social capital. *Regional Science and Urban Economics*, 42(5):890–903, 2012.

[237] Xiaojun Wan. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the conference on empirical methods in natural language processing*, pages 553–561. Association for Computational Linguistics, 2008.

[238] Xiaojun Wan. Co-training for cross-lingual 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 1-volume 1*, pages 235–243. Association for Computational Linguistics, 2009.

[239] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626. ACM, 2011.

[240] Ingmar Weber, Emilio Zagheni, et al. Studying inter-national mobility through ip geolocation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 265–274. ACM, 2013.

[241] Wouter Weerkamp, Simon Carter, and Manos Tsagkias. How people use twitter in different languages. 2011.

[242] Albert Weichselbraun, Stefan Gindl, and Arno Scharl. Enriching semantic knowledge bases for opinion mining in big data applications. *Knowledge-based systems*, 69:78–85, 2014.

[243] Michael J White. Segregation and diversity measures in population distribution. *Population index*, pages 198–221, 1986.

[244] Janyce M Wiebe, Rebecca F Bruce, and Thomas 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*, pages 246–253. Association for Computational Linguistics, 1999.

[245] Gail Wilson and Elizabeth Stacey. Online interaction impacts on learning: Teaching the teachers to teach online. In *Interact, Integrate, Impact: proceedings of the 20th annual conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE)*, pages 541–551. ASCILITE, 2003.

[246] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, 2005.

[247] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

[248] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? finding strong and weak opinion clauses. In *aaai*, volume 4, pages 761–769, 2004.

[249] Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of economic perspectives*, 18(2):107–126, 2004.

[250] Chi-En Wu and Richard Tzong-Han Tsai. Using relation selection to improve value propagation in a conceptnet-based sentiment dictionary. *Knowledge-Based Systems*, 69:100–107, 2014.

[251] Chunxu Wu, Lingfeng Shen, and Xuan Wang. A new method of using contextual information to infer the semantic orientations of context dependent opinions. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 4, pages 274–278. IEEE, 2009.

[252] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.

[253] Min Xiao and Yuhong Guo. Semi-supervised matrix completion for cross-lingual text classification. In *AAAI*, pages 1607–1614, 2014.

[254] Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati, and Jennifer A. Holland. Using Twitter data for demographic research. *Demographic Research*, 37:1477–1514, nov 2017.

[255] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.

[256] Sheng Yu and Subhash Kak. A survey of prediction using social media. *arXiv preprint arXiv:1203.1647*, 2012.

[257] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. Inferring international and internal migration patterns from Twitter data. In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, pages 439–444, New York, New York, USA, 2014. ACM Press.

[258] Emilio Zagheni and Ingmar Weber. You are where you e-mail: using e-mail data to estimate international migration rates. In *Proceedings of the 4th annual ACM web science conference*, pages 348–351. ACM, 2012.

[259] Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review*, 43(4):721–734, dec 2017.

[260] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89, 2011.

[261] Yaohui Zhao. Causes and consequences of return migration: recent evidence from china. *Journal of Comparative Economics*, 30(2):376–394, 2002.

[262] Zhixin Zhou, Xiuzhen Zhang, and Mark Sanderson. Sentiment analysis on twitter through topic-based lexicon expansion. In *Australasian Database Conference*, pages 98–109. Springer, 2014.

[263] Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 304–313, 2014.